Российская академия наук
Вычислительный центр РАН ФИЦ ИУ РАН

Интеллектуализация обработки информации

14-я Международная конференция

Москва, 2022

УДК 004.85+004.89+004.93+519.2+519.25+519.7
ББК 22.1:32.973.26-018.2
И 73

Интеллектуализация обработки информации: Тезисы докладов 14-й Международной конференции, г. Москва 2022 г. — М.: Российская академия наук, 2022. — 472 с.

ISBN 978-5-907366-77-0

В сборнике представлены тезисы докладов 14-й Международной конференции «Интеллектуализация обработки информации», проводимой Российской академией наук, Вычислительным центром Федерального исследовательского центра «Информатика и управление» РАН.
Конференция проводится регулярно, начиная с 1989 г., и является представительным научным форумом в области интеллектуального анализа данных, машинного обучения, распознавания образов, анализа изображений, обработки сигналов, дискретного анализа.
Сайт конференции `http://machinelearning.ru/wiki?title=IIP`.

ISBN 978-5-907366-77-0

UDK 004.85+004.89+004.93+519.2+519.25+519.7
BBK 22.1:32.973.26-018.2

Intelligent Data Processing: Theory and Applications: Book of abstract of the 14th International Conference, Moscow, 2022. — Moscow: Russian Academy of Sciences, 2022. — 472 p.

ISBN 978-5-907366-77-0

The volume contains the abstracts of the 14th International Conference "Intelligent Data Processing: Theory and Applications". The conference is organized by the Russian Academy of Sciences, Federal Research Center "Computer Science and Control" of RAS. The conference has being held biennially since 1989. It is one of the most recognizable scientific forums on data mining, machine learning, pattern recognition, image analysis, signal processing, and discrete analysis.

The conference website `http://machinelearning.ru/wiki?title=IIP`.

ISBN 978-5-907366-77-0

Оргкомитет

Председатель: Соколов Игорь Анатольевич, *акад. РАН*
Ученый секретарь: Чехович Юрий Викторович, *к.ф.-м.н.*

Борисова Татьяна Игоревна
Грабовой Андрей Валериевич, *к.ф.-м.н.*
Громов Андрей Николаевич
Инякин Андрей Сергеевич, *к.ф.-м.н.*
Лемтюжникова Дарья Владимировна, *к.ф.-м.н.*
Петров Игорь Борисович, *чл.-корр. РАН*
Рейер Иван Александрович, *к.т.н.*
Шананин Александр Алексеевич, *акад. РАН*

Программный комитет

Сопредседатели: Стрижов Вадим Викторович, *д.ф.-м.н.*
Воронцов Константин Вячеславович, *д.ф.-м.н.*
Члены комитета: Ватолин Дмитрий Сергеевич, *к.ф.-м.н.*
Гимади Эдуард Хайрутдинович, *д.ф.-м.н.*
Горнов Александр Юрьевич, *д.т.н.*
Громова Ольга Алексеевна, *д.м.н.*
Двоенко Сергей Данилович, *д.ф.-м.н.*
Дяконов Александр Геннадьевич, *д.ф.-м.н.*
Конушин Антон Сергеевич *к.ф.-м.н*
Краснопрошин Виктор Владимирович, *д.т.н.*
Лазарев Александр Алексеевич, *д.ф.-м.н.*
Матвеев Иван Алексеевич, *д.т.н.*
Местецкий Леонид Моисеевич, *д.т.н.*
Пытьев Юрий Петрович, *д.ф.-м.н.*
Рязанов Владимир Васильевич, *д.ф.-м.н.*
Семенов Алексей Львович, *акад. РАН*
Сойфер Виктор Александрович, *акад. РАН*
Хачай Михаил Юрьевич, *чл.-корр. РАН*
Чехович Юлия Викторовна
Чуличков Алексей Иванович, *д.ф.-м.н.*

Organizing Committee

Chair: Igor Sokolov, *acad. of RAS*

Secretary: Yury Chekhovich, *C.Sc.*

Tatiana Borisova
Andrey Grabovoy, *C.Sc.*
Andrey Gromov
Andrey Inyakin, *C.Sc.*
Dariya Lemtushnikova, *C.Sc.*
Igor Petrov, *corr. member of RAS*
Ivan Reyer, *C.Sc.*
Alexander Shananin, *acad. of RAS*

Program Committee

Chair: Vadim Strijov, *D.Sc.*,
Konstantin Vorontsov, *D.Sc.*

Committee members: Dmitriy Vatolin, *C.Sc.*
Edward Gimadi, *D.Sc.*
Alexander Gornov, *D.Sc.*
Olga Gromova, *D.Sc.*
Sergey Dvoenko, *D.Sc.*
Alexander Dyakonov, *D.Sc.*
Anton Konushin, *C.Sc.*
Viktor Krasnoproshin *D.Sc.*
Alexander Lazarev *D.Sc.*
Ivan Matveev *D.Sc.*
Leonid Mestetskiy, *D.Sc.*
Yury Pytiev, *D.Sc.*
Vladimir Ryazanov, *D.Sc.*
Alexey Semenov, *acad. of RAS*
Viktor Soyfer, *acad. of RAS*
Michael Khachay, *corr. member of RAS*
Yulia Chekhovich
Alexey Chulichkov, *D.Sc.*

Рецензенты

Адуенко А. А.
Анциперов В. Е.
Бахтеев О. Ю.
Бунакова В. Р.
Визильтер Ю. В.
Володин С. Е.
Воронцов К. В.
Гасников А. В.
Генрихов И. Е.
Гнеушев А. Н.
Голиков А. И.
Гончаров А. В.
Гороховский К. Ю.
Грабовой А. В.
Двоенко С. Д.
Дьяконов А. Г.
Жариков И. Н.
Животовский Н. К.
Загоруйко Н. Г.
Зайцев А. А.
Ивахненко А. А.
Игнатов А. Д.
Игнатов Д. И.
Игнатьев В. Ю.
Инякин А. С.
Исаченко Р. Г.
Карасиков М. Е.
Катруца А. М.
Копылов А. В.
Кочетов Ю. А.
Красоткина О. В.
Крымова Е. А.
Кудинов М. С.
Кузнецов М. П.
Кузьмин А. А.
Кулунчаков А. С.
Кушнир О. А.
Ланге М. М.
Ломов Н. А.
Лукашевич Н. В.
Майсурадзе А. И.
Максимов Ю. В.
Матвеев И. А.
Местецкий Л. М.
Михеева А. В.
Мнухин В. Б.
Мотренко А. П.
Мурашов Д. М.
Неделько В. М.
Нейчев Р. Г.
Новик В. П.
Одиноких Г. А.
Панов А. И.
Панов М. Е.
Потапенко А. А.
Пушняков А. С.
Рейер И. А.
Рудой Г. И.
Рябенко Е. А.
Сафонов И. В.
Сенько О. В.
Середин О. С.
Сотнезов Р. М.
Стенина М. М.
Стрижов В. В.
Сулимова В. В.
Талипов К. И.
Таханов Р. С.
Торшин И. Ю.
Трёкин А. Н.
Турдаков Д. Ю.
Федоряка Д. С.
Фрей А. И.
Хачай М. Ю.
Черепанов Е. В.
Чуличков А. И.
Янина А. О.

Reviewers

Aduenko A.
Antsiperov V.
Bakhteev O.
Bunakova V.
Cherepanov E.
Chulichkov A.
Dvoenko S.
D'yakonov A.
Fedoryaka D.
Frei A.
Gasnikov A.
Genrikhov I.
Gneushev A.
Golikov A.
Goncharov A.
Gorokhovskiy K.
Grabovoy A.
Ignat'ev V.
Ignatov A.
Ignatov D.
Inyakin A.
Isachenko R.
Ivakhnenko A.
Karasikov M.
Katrutsa A.
Khachay M.
Kochetov Yu.
Kopylov A.
Krasotkina O.
Krymova E.
Kudinov M.
Kulunchakov A.
Kushnir O.
Kuz'min A.
Kuznetsov M.
Lange M.
Lomov N.
Lukashevich N.
Maksimov Yu.
Matveev I.
Maysuradze A.
Mestetskiy L.
Mikheeva A.
Mnukhin V.
Motrenko A.
Murashov D.
Nedel'ko V.
Nejchev R.
Novik V.
Odinokikh G.
Panov A.
Panov M.
Potapenko A.
Pushnyakov A.
Reyer I.
Rudoy G.
Ryabenko E.
Safonov I.
Sen'ko O.
Seredin O.
Sotnezov R.
Stenina M.
Strizhov V.
Sulimova V.
Takhanov R.
Talipov K.
Torshin I.
Trekin A.
Turdakov D.
Vizil'ter Yu.
Volodin S.
Vorontsov K.
Yanina A.
Zagorujko N.
Zajtsev A.
Zharikov I.
Zhivotovskiy N.

Краткое оглавление

Brief contents

От кючевых слов Ю.И. Журавлева: “алгоритмы с оценками” и “почти всегда” до асимптотически точных алгоритмов

Гимади Эдуард Хайрутдинович[1,2]⋆ gimadi@math.nsc.ru

[1]Новосибирск, Институт математики им. С.Л.Соболева

[2]Новосибирск, Новосибирский государственный университет

Ю.И. Журавлев был моим первым завлабом в Институте математики Сибирского Отделения (в лаборатории теории вычислений). Его ключевые слова “алгоритмы с оценками” и “почти всегда” предопределили всю мою дальнейшую научную деятельность в области дискретной оптимизации в исследовании операций.

Для задач дискретной оптимизации основным фактором, определяющим реализуемость алгоритмов их решения, является размерность (длина записи входа) задачи, в зависимости от которой имеют место те или иные оценки качества работы алгоритма. Наиболее важнымими оценками качества алгоритма в зависимости от размерности задачи считают его временную сложность и точность получаемого решения, а для задач на случайных входах также важными считаются оценки степени надежности получения решения.

Оценкой **относительной погрешности** (relative error) алгоритма A решения задачи на детерминированных входах $\mathcal{I}_n$ размера n называют такую величину $\varepsilon_A(n)$, что на любом входе $I \in \mathcal{I}_n$ верно $\frac{|W_A(I)-OPT(I)|}{OPT(I)} \leqslant \varepsilon_A(n)$, где $OPT(I)$ и $W_A(I)$ — оптимальное и найденное в результате работы алгоритма A значения целевой функции задачи на входе I.

Для задач на случайных входах качество алгоритма характеризуется также вероятностью несрабатывания (failure probability).

Алгоритм A на множестве входов $\mathcal{I}_n$ имеет оценки **относительной погрешности** $\varepsilon_A(n)$ и **вероятностью несрабатывания** $\delta_A(n)$ в классе задач размера n, если на входах $I \in \mathcal{I}_n$ верно следующее неравенство $\mathbb{P}\left\{\frac{|W_A(I)-OPT(I)|}{OPT(I)} > \varepsilon_A(n)\right\} \leqslant \delta_A(n)$, где $\delta_A(n)$ означает долю случаев, когда алгоритм A не гарантирует получение решения с анонсированной погрешностью.

Алгоритм тем лучше, чем меньше $\varepsilon_A(n)$ и $\delta_A(n)$. Алгоритм с оценками относительной погрешности $\varepsilon_A(n)$ и вероятности несрабатывания $\delta_A(n)$ называем **асимтотически точным** (АТ), если обе оценки стремятся к нулю при $n \to \infty$. 50-70 г.г. прошлого века ассоциировались с понятием **“проклятия размерности”** ("curse of dimensionality" , Ричард Беллман, 1961 г.). Эта проблема, связанная с экспоненциальным возрастанием времени решения задачи при увеличении длины записи входных данных. В противовес понятию "проклятия размерности" в рамках АТ подхода к решению трудных задач дискретной оптимизации размерность задачи является нашим другом и союзником.

К настоящему моменту определилось немало примеров реализации АТ подхода к решению таких большеразмерных задач дискретной оптимизации в исследованиии операций, как задачи маршрутизации (в том числе задача крммивояжера (ЗК)), многоиндексные задачи о назначениях, задачи кластеризации, экстремальные задачи на графах и сетях и т.п. Обычно эти задачи труднорешаемы [1], что обуславливает актуальность разработки эффективных алгоритмов решения таких задач с гарантированными оценками качества их работы.

Первый пример алгоритма с почти всегда гарантированной оценкой точности был дан А.А. Боровковым 60 лет тоиу назад в работе [2] для ЗК$_{min}$.

Теорема [2]. *В случае равномерного распределения точек в единичном квадрате алгоритм АБ решает ЗК$_{min}$ за время $\mathcal{O}(n\log n)$ почти всегда ($\delta_{АБ}(n)\to 0$) с оценкой относительной погрешности $\varepsilon_{АБ}(n)=0.48$.*

Приведем первые в мире примеры АТ подхода с использованием $\mathcal{O}(n^2)$-алгоритма ИБГ ("Иди в ближпйший непройденный город") для ЗК$_{min}$ на полном графе с весами ребер — независимыми случайными величинами (н.сл.в.) с функцией распределения (ф.р.) дискретного (д.ф.р.) и непрерывного (н.ф.р.) типа.

Определим класс $\mathcal{C}_n,\ n>3$, входов ЗК$_{min}$, задаваемый $n\times n$-матрицей (c_{ij}) расстояний между городами, элементы которой — н.сл.в. с общей ф.р.

Теорема [3]. *Алгоритм ИБГвыдает АТ решение ЗК$_{min}$ в классе $\mathcal{C}_n$ с общей д.ф.р. $p_k=\mathbb{P}\{c_{ij}=k\}, k=\overline{1,K_n}$, при условии $\sum_{k=1}^{K_n}(p_1+\ldots+p_k)^{-1}=o(n)$.*

Теорема [4]. *Пусть задан класс $\mathcal{C}_n$ входов ЗК$_{min}$ с $n\times n$-матрицей (c_{ij}), элементы которой — н.сл.в. с общей н.ф.р. на отрезке $(a_n,b_n), a_n>0$, и пусть $\mathcal{P}_\xi(x)$ — ф.р. н.сл.в. $\xi=\frac{c_{ij}-a_n}{b_n-a_n}\leqslant 1,\ 1\leqslant i,j\leqslant n$. Тогда алгоритм ИБГ является алгоритмом с оценками: $\varepsilon_{\tilde{A}}(n)=\mathcal{O}\big(\frac{b_n/a_n}{n/\max(n\gamma_n;\mathcal{J}_n)}\big)$ и $\delta_{\tilde{A}}(n)=\mathcal{O}\big(\frac{1}{n\gamma_n+\mathcal{J}_n}\big)$, где обозначено $\mathcal{J}_n=\int_{\gamma_n}^1\frac{dx}{\mathcal{P}_\xi(x)}$ и γ_n — корень уравнения $\mathcal{P}_\xi(x)=\frac{1}{n}$. При этом на выходе алгоритма ИБГ получается АТ решение ЗК$_{min}$ при выполнении следующих двух условий: $\frac{b_n}{a_n}=o\Big(\frac{n}{\max(n\gamma_n,\mathcal{J}_n)}\Big)$ и $\mathcal{J}_n\to\infty$ при $n\to\infty$.*

Следствие. *В случае всех ф.р. вида $\mathcal{P}_\xi(x)\leqslant x$ (мажорирующих равномерное распределение UNI(a_n,b_n)) алгоритм ИБГ имеет оценки $\varepsilon_{\tilde{A}}(n)=\mathcal{O}\Big(\frac{b_n/a_n}{n/\ln n}\Big)$ и $\delta_{\tilde{A}}(n)=\mathcal{O}\big(\frac{1}{\ln n}\big)$, а АТ решение достигается при более компактном условии: $\frac{b_n}{a_n}=o\big(\frac{n}{\ln n}\big)$.*

Первые результаты по обоснованию асимптотической точности были получены с использованием неравенства Чебышева. Позже более продуктивной оказалась

Теорема Петрова *Пусть $S=\sum_{j=1}^n X_j$ — сумма н.сл.в. и существуют положительные константы $h_1,\ldots,h_n$ и T такие, что верны неравенства $\mathbb{E}e^{tX_j}\leqslant e^{\frac{1}{2}h_jt^2}$ для всяких $j=\overline{1,n}$ и $0\leqslant t\leqslant T$.*

Тогда $\mathbb{P}\{S>x\}\leqslant\begin{cases} exp\{-x^2/2H\}\ при\ 0\leqslant x<HT,\\ exp\{-Tx/2\}\ при\ x\geqslant HT\end{cases}$, где $H=\sum_{j=1}^n h_j$.

С использованием теоремы Петрова имеют место следующие оценки качества алгоритма ИБГ для ЗК$_{min}$: $\varepsilon_A(n) = O\Big(\frac{\beta_n/a_n}{n/\ln n}\Big)$; $\delta_A(n) = \frac{1}{n}$ с условием асимптотической точности $\frac{\beta_n}{a_n} = o\Big(\frac{n}{\ln n}\Big)$, где параметр $\beta_n = b_n,\ \alpha_n,\ \sigma_n$, соответственно, для непрерывных ф.р.: равномерного UNI(a_n, b_n) и усеченно-смещенных: экспоненциального $EXP(a_n, \alpha_n)$ и нормального NOR(a_n, σ_n).

К числу примеров асимптотического подхода к труднорешаемым задачам на детерминированных входах, в первую очередь следует отнести ЗК$_{max}$ (TSP$_{max}$) в многомерных эвклидовых пространствах и ее обобщение (m-PSP$_{max}$) на случай поиска нескольких реберно непересекающихся маршрутов коммивояжера максимального суммарного веса [5].

Еще примеры труднорешаемых задач с реализациями подхода.

- Многоиндексная аксиальная задача о назначениях.
- Трехиндексная планарная m-слойная задача о назначениях.
- Задача о нескольких коммивояжерах (m-PSP) на случ. входах с одинаковыми и с различными весовыми функциями маршрутов коммивояжера.
- Покрытие полного графа m несмежными циклами заданных размеров с экстремальным суммарным весом ребер (m-Cycles Cover Problem —m-CCP).
- Задача отыскания связного остовного подграфа с максим. весом ребер в полном неор-ом графе с заданными степенями вершин.
- Задача отыскания в графе одного и нескольких остовных деревьев с ограниченным (снизу, сверху, либо фиксированным) диаметром.
- Задачи маршрутизации транспортных средств (VRP) с одним и несколькими депо, с возвратом и необязательным возвратом в депо и др.
- Упаковки в контейнеры и в полосу с ограниченным общим ресурсом.
- Поиск подмножества векторов заданного размера с наибольшей суммой.

[1] *Garey M. R., Johnson D. S.*: Computers and Intractability, Freeman, San Francisco, 1979, 340 p.

[2] *Боровков А. А.* К вероятностной постановке двух экономических задач // ДАН СССР, 1962, 146(5). С. 983–986.

[3] *Перепелица В. А., Гимади Э. Х.* К задаче нахождения минимального гамильтонова контура на графе со взвешенными дугами // Дискр. анализ. Новосибирск, 1969. Вып. 15. С. 57–65.

[4] *Гимади Э. Х., Перепелица В. А.* Асимптотически точный подход к решению задачи коммивояжера // Управляемые системы. Сб. науч. тр. Новосибирск: Ин-т математики СО АН СССР. 1974. Вып. 12. С. 35–45.

[5] *Гимади Э. Х., Хачай М. Ю.* Экстремальные задачи на множествах перестановок // Екатеринбург: Изд-во УМЦ УПИ, 2016. 219 с.

From the key words of Yu.I. Zhuravleva: "algorithms with estimates" and "almost always" to asymptotically exact algorithms

Gimadi Eduard[1,2]⋆ gimadi@math.nsc.ru

[1]Novosibirsk, Institute of Mathematics. S.L.Soboleva

[2]Novosibirsk, Novosibirsk State Universitys

Yu.I. Zhuravlev was my first head of the laboratory at the Institute of Mathematics of the Siberian Branch (in the laboratory of the theory of computation). His keywords "estimated algorithms" and "almost always" predetermined all my further scientific activity in the field of discrete optimization in operations research.

For discrete optimization problems, the main factor determining the feasibility of algorithms for solving them is the dimension (the length of the input record) of the problem, depending on which one or another assessment of the quality of the algorithm takes place. The most important estimates of the quality of an algorithm, depending on the dimension of the problem, are considered to be its time complexity and the accuracy of the solution obtained, and for problems on random inputs, estimates of the degree of reliability of obtaining the solution are also considered important.

The estimate **relative error** (relative error) of the algorithm A for solving a problem on deterministic inputs $\mathcal{I}_n$ of size n is a value $\varepsilon_A(n)$ such that at any input $I \in \mathcal{I}_n$ true $\frac{|W_A(I)-OPT(I)|}{OPT(I)} \leqslant \varepsilon_A(n)$, where $OPT(I)$ and $W_A(I)$ – - the optimal and the value of the objective function of the problem at the input I found as a result of the A algorithm.

For tasks on random inputs, the quality of the algorithm is also characterized by the failure probability.

Algorithm A on the set of inputs $\mathcal{I}_n$ has estimates **relative error** $\varepsilon_A(n)$ and **failure probability** $\delta_A(n)$ in the class of problems of size n if the following inequality is true on the inputs $I \in \mathcal{I}_n$ $\mathbb{P}\left\{\frac{|W_A(I)-OPT(I)|}{OPT(I)} > \varepsilon_A(n)\right\} \leqslant \delta_A(n)$, where $\delta_A(n)$ means the proportion of cases where the A algorithm does not guarantee a solution with the announced error.

The algorithm is better, the smaller $\varepsilon_A(n)$ and $\delta_A(n)$. An algorithm with estimates of the relative error $\varepsilon_A(n)$ and the probability of failure $\delta_A(n)$ is called **asymptotically exact** (AT) if both estimates tend to zero as $n \to \infty$. 50-70 years of the last century were associated with the concept of the **"curse of dimensionality"** ("curse of dimensionality" Richard Bellman, 1961). This problem is associated with an exponential increase in the time to solve the problem with an increase in the length of the input data record. In contrast to the notion of the "curse of dimensionality" in the framework of the AT approach to solving difficult problems of discrete optimization, the dimensionality of the problem is our friend and ally.

To date, many examples of the implementation of the AT approach have been identified for solving such large-scale discrete optimization problems in operations research as routing problems (including the CRM problem), multi-index assignment problems, clustering problems, extremal problems on graphs and networks, etc. .P. Usually, these problems are difficult to solve [1], which makes it important to develop efficient algorithms for solving such problems with guaranteed performance estimates.

[1] *Garey M. R., Johnson D. S.*: Computers and Intractability, Freeman, San Francisco, 1979, 340 p.

[2] *Borovkov A. A.* On the probabilistic formulation of two economic problems // DAN SSSR, 1962, 146(5). S. 983–986.

[3] *Perepelitsa V. A., Gimadi E. Kh.* On the problem of finding a minimal Hamiltonian contour on a graph with weighted arcs // Diskr. analysis. Novosibirsk, 1969. Issue. 15. S. 57–65.

[4] *Gimadi E. Kh., Perepelitsa V. A.* An asymptotically exact approach to solving the traveling salesman problem // Controlled Systems. Sat. scientific tr. Novosibirsk: Institute of Mathematics SO AN USSR. 1974. Issue. 12. S. 35–45.

[5] *Gimadi E. Kh., Khachai M. Yu.* Extremal problems on sets permutations // Ekaterinburg: Publishing House of UMTs UPI, 2016. 219 p.

Линейная бинарная классификация при интервальной неопределенности данных

Ерохин Владимир Иванович[1]⋆ erohin_v_i@mail.ru
Кадочников Андрей Павлович[1] kado162@mail.ru
Сотников Сергей Владимирович[1] svsotnikov66@gmail.com
[1]Санкт-Петербург, Военно-космическая академия имени А.Ф. Можайского

Пусть в пространстве $\mathbb{R}^n$ заданы два конечных множества

$$\mathbf{P} = \{\mathfrak{p}_i\}_{i=1}^{m_1} \text{ и } \mathbf{Q} = \{\mathfrak{q}_i\}_{i=1}^{m_2}.$$

Указанным множествам соответствуют матрицы

$$P = (p_{ij}) \in \mathbb{R}^{m_1 \times n} \text{ и } Q = (q_{ij}) \in \mathbb{R}^{m_2 \times n}.$$

Элементы матриц P и Q заданы с интервальной неопределенностью. Вместо точных значений элементов p_{ij} и q_{ij} известны их нижние и верхние границы: $\underline{P} = (\underline{p}_{ij}) \in \mathbb{R}^{m_1 \times n}$, $\overline{P} = (\overline{p}_{ij}) \in \mathbb{R}^{m_1 \times n}$, $\underline{Q} = (\underline{q}_{ij}) \in \mathbb{R}^{m_2 \times n}$, $\overline{Q} = (\overline{q}_{ij}) \in \mathbb{R}^{m_2 \times n}$. Приведенные ниже неравенства выполнены поэлементно:

$$\underline{P} \leqslant P \leqslant \overline{P},\ \underline{Q} \leqslant Q \leqslant \overline{Q}. \tag{1}$$

Хорошо известно (см., например, [1]), что при отсутствии интервальной неопределенности задачу *нестрогого* отделения множеств $\mathbf{P}$, $\mathbf{Q}$ гиперплоскостью, не проходящей через начало координат, можно формализовать как задачу решения системы линейных неравенств вида

$$Pw \leqslant 1, Qw \geqslant 1 \Leftrightarrow \begin{bmatrix} P \\ -Q \end{bmatrix} w \leqslant \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \tag{2}$$

где $w \in \mathbb{R}^n$ – вектор неизвестных коэффициентов левой части уравнения гиперплоскости (правая часть уравнения нормирована к 1). Очевидно, что для решения задач машинного обучения используются более тонкие методы, чем система (2) (см., например, [3]), поэтому будем её рассматривать в качестве удобного модельного примера и отправной точки для дальнейших рассуждений. Пусть $\mathbf{W}$ – допустимое множество системы (2). Заметим, что при $\mathbf{W} \neq \varnothing$ в общем случае система (2) имеет бесконечное множество решений, и все они, в смысле задачи нестрогого отделения множеств $\mathbf{P}$ и $\mathbf{Q}$ гиперплоскостью, являются *эквивалентными*.

Если $\mathbf{W} = \varnothing$, при определенных условиях (в контексте задач машинного обучения) может быть полезно *псевдорешение* системы (2), формализованное, например, как решение задачи выпуклой негладкой безусловной минимизации

$$\left\| \left[\begin{bmatrix} P \\ -Q \end{bmatrix} w - \begin{bmatrix} 1 \\ -1 \end{bmatrix} \right]_+ \right\| \to \min_w, \tag{3}$$

где $||\cdot||$ – некоторая векторная норма, $[\,\cdot\,]_+$ – операция положительной срезки, применяемая к векторному аргументу поэлементно.

Некоторым парадоксом является то факт, что вне контекста машинного обучения задача (3) является более привлекательной, чем задача решения системы (2), поскольку при выборе строго выпуклой нормы (например, евклидовой), она имеет единственное решение. Рассмотрим следующие задачи:

Задача 1. Найти вектор $w \in \mathbb{R}^n$, являющийся решением системы линейных неравенств (2) для *любых* матриц P, Q, удовлетворяющих системе линейных неравенств (1).

Задача 2. Найти вектор $w \in \mathbb{R}^n$, являющийся решением задачи (3) для *любых* матриц P, Q, удовлетворяющих системе линейных неравенств (1).

Заметим, что множества матриц $\mathbf{P}$, $\mathbf{Q}$, удовлетворяющих системе неравенств (1), являются бесконечными. При этом $\mathbf{P}$ имеет $m_1 \times 2^n$ крайних точек, $\mathbf{Q}$ – $m_2 \times$ $\times\, 2^n$ крайних точек, что дает основание опасаться NP-трудности задач 1 и 2. Указанные опасения, к счастью, не оправдываются, поскольку в интервальном анализе известна следующая

Теорема 1. *[2] Система линейных неравенств $Ax \leqslant b$ разрешима для всех A, b таких что $\underline{A} \leqslant A \leqslant \overline{A}$, $\underline{b} \leqslant b \leqslant \overline{b}$ тогда и только тогда, когда разрешима система $\overline{A}x^+ - \underline{A}x^- \leqslant \underline{b}$, $x^+, x^- \geqslant 0$. При этом $x = x^+ - x^-$.*

В силу теоремы 1 легко указать методы решения задач 1 и 2.

Утверждение 1. *Задача 1 имеет решение тогда и только тогда, когда имеет решение система линейных неравенств*

$$\begin{bmatrix} \overline{P} \\ -\underline{Q} \end{bmatrix} w^+ - \begin{bmatrix} \underline{P} \\ -\overline{Q} \end{bmatrix} w^- \leqslant \begin{bmatrix} 1 \\ -1 \end{bmatrix}, w^+, w^- \geqslant 0. \tag{4}$$

При этом $w = w^+ - w^-$.

Утверждение 2. *Если $||\cdot||$ – абсолютная векторная норма, то задача 2 эквивалентна задаче*

$$\|\Psi(w)\| = \left\| \begin{bmatrix} P_c w + P_r\,|w| - 1 \\ -Q_c w + Q_r\,|w| + 1 \end{bmatrix}_+ \right\| \to \min_w, \tag{5}$$

где $P_c = (\underline{P} + \overline{P})/2$, $P_r = (\overline{P} - \underline{P})/2$, $Q_c = (\underline{Q} + \overline{Q})/2$, $Q_r = (\overline{Q} - \underline{Q})/2$, $|\cdot|$ – операция взятия абсолютной величины, применяемая к векторному аргументу поэлементно.

Заметим, что (5), также как и (3), является задачей выпуклой негладкой безусловной минимизации, имеющей при использовании строго выпуклой нормы

единственное решение. Рассмотрим задачи

$$\left\| \left[\begin{bmatrix} P \\ -Q \end{bmatrix} w - \begin{bmatrix} 1 \\ -1 \end{bmatrix} + \gamma \cdot 1 \right]_+ \right\| \to \min_w (= \delta_\gamma), \tag{6}$$

$$\left\| \left[\begin{bmatrix} P_c \\ -Q_c \end{bmatrix} w + \begin{bmatrix} P_r \\ Q_r \end{bmatrix} |w| - \begin{bmatrix} 1 \\ -1 \end{bmatrix} + \gamma \cdot 1 \right]_+ \right\| \to \min_w (= \delta_\gamma), \tag{7}$$

где $\gamma > 0$ – скалярный параметр.

Задачи (6) и (7) актуальны в случае линейно отделимых множеств (соответственно точных или обладающих интервальной неопределенностью). Очевидно, что $\delta_\gamma = 0$ при $\gamma = 0$ и задачи (6) и (7) имеют бесконечное множество решений. В то же время можно показать, что начиная с некоторого значения $\gamma > 0$ будет выполняться условие $\delta_\gamma > 0$ и упомянутые выше задачи при использовании строго выпуклой нормы будут иметь единственное решение.

На рисунке 1 представлены иллюстрации результатов решения модельных задач 7 и 5 (с интервальной неопределенностью данных) при $n = 2$ и использовании евклидовой нормы.

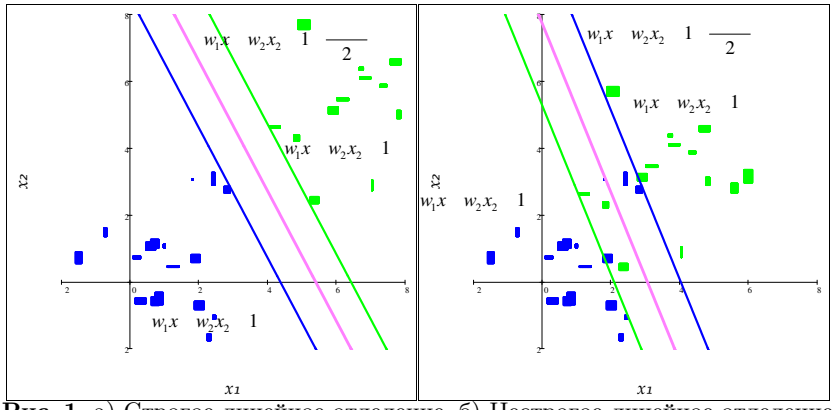


Рис. 1. а) Строгое линейное отделение. б) Нестрогое линейное отделение

Определение параметров α, β не описано в силу ограниченного объема тезисов доклада.

[1] *Вапник В. Н., Червоненкис А. Я.* Теория распознавания образов (статистические проблемы обучения). М.: Наука, 1974.

[2] *Фидлер М., Недома Й., Рамник Я. и др.,* Задачи линейной оптимизации с неточными данными. М.-Ижевск: НИЦ «Регулярная и хаотическая динамика», 2008.

[3] *Deisenroth M. P., Faisal A. A., Ong C. S.* Mathematics for machine learning. Cambridge University Press, 2020.

[4] *Ланкастер П.* Теория матриц. М.: Наука, 1982.

Linear binary classification under interval uncertainty of data

Erokhin Vladimir[1]⋆ erohin_v_i@mail.ru
Kadochnikov Andrey[1] kado162@mail.ru
Sotnikov Sergey[1] svsotnikov66@gmail.com
[1]St Petersburg, A.F. Mozhaisky Military-Space Academy

Let two finite sets be given in $\mathbb{R}^n$

$$\mathbf{P} = \{\mathfrak{p}_i\}_{i=1}^{m_1} \text{ and } \mathbf{Q} = \{\mathfrak{q}_i\}_{i=1}^{m_2}.$$

The specified sets correspond to the matrices

$$P = (p_{ij}) \in \mathbb{R}^{m_1 \times n} \text{ and } Q = (q_{ij}) \in \mathbb{R}^{m_2 \times n}.$$

The elements of the matrices P and Q are given with interval uncertainty. Instead of the exact values of the elements p_{ij} and q_{ij} their lower and upper boundaries are known: $\underline{P} = (\underline{p}_{ij}) \in \mathbb{R}^{m_1 \times n}$, $\overline{P} = (\overline{p}_{ij}) \in \mathbb{R}^{m_1 \times n}$, $\underline{Q} = (\underline{q}_{ij}) \in \mathbb{R}^{m_2 \times n}$, $\overline{Q} = (\overline{q}_{ij}) \in \mathbb{R}^{m_2 \times n}$. The following inequalities are satisfied element by element:

$$\underline{P} \leqslant P \leqslant \overline{P},\ \underline{Q} \leqslant Q \leqslant \overline{Q}. \tag{1}$$

It is well known (see, for example, [1]) that in the absence of interval uncertainty the problem of *unstrict* separation of sets $\mathbf{P}$, $\mathbf{Q}$ by a hyperplane not passing through the origin can be formalized as a problem of solving a system of linear inequalities of the form

$$Pw \leqslant 1, Qw \geqslant 1 \Leftrightarrow \begin{bmatrix} P \\ -Q \end{bmatrix} w \leqslant \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \tag{2}$$

where $w \in \mathbb{R}^n$ is the vector of unknown coefficients of the left side of the hyperplane equation (the right side of the equation is normalized to 1). Obviously, more subtle methods are used to solve machine learning problems than the (2) system (see, for example, [3]), so we will consider it as a convenient model example and a starting point for further reasoning. Let $\mathbf{W}$ be a feasible set systems (2). Note that for $\mathbf{W} \neq \varnothing$ in the general case, the system (2) has an infinite set of solutions, and all of them, in the sense of the problem of unstrict separation of sets $\mathbf{P}$ and $\mathbf{Q}$ hyperplane, are *equivalent*.

If $\mathbf{W} = \varnothing$, under certain conditions (in the context of machine learning problems), a *pseudo-solution* of the system (2) could be useful, formalized, for example, as a solution to the convex non-smooth unconstrained minimization problem

$$\left\| \left[\begin{bmatrix} P \\ -Q \end{bmatrix} w - \begin{bmatrix} 1 \\ -1 \end{bmatrix} \right]_+ \right\| \to \min_w, \tag{3}$$

where $||\cdot||$ is some vector norm, $[\,\cdot\,]_+$ is a positive cut operation applied to a vector argument element by element.

Quite a paradox that outside the context of machine learning, the problem (3) is more attractive than the problem of solving the system (2), because when choosing a strictly convex norm (for example, Euclidean), it has a unique solution. Consider the following problems:

Problem 1. Find the vector $w \in \mathbb{R}^n$, which is the solution of the system of linear inequalities (2) for *any* matrices P, Q satisfying the system of linear inequalities (1).

Problem 2. Find the vector $w \in \mathbb{R}^n$, which is the solution of the problem (3) for *any* matrices P, Q satisfying the system of linear inequalities (1).

Note that the sets of matrices $\mathbf{P}$, $\mathbf{Q}$ satisfying the system of inequalities (1) are infinite. At the same time $\mathbf{P}$ has $m_1 \times 2^n$ extreme points, $\mathbf{Q}$ – $m_2 \times 2^n$ extreme points, which gives reason to fear NP-difficulties of problems 1 and 2. Fortunately, these fears were not justified, since the following is known in interval analysis:

Theorem 1. [2] *The system of linear inequalities $Ax \leqslant b$ is solvable for all A, b such that $\underline{A} \leqslant A \leqslant \overline{A}$, $\underline{b} \leqslant b \leqslant \overline{b}$ if and only if the system is solvable $\overline{A}x^+ - \underline{A}x^- \leqslant \underline{b}$, $x^+, x^- \geqslant 0$. In this case, $x = x^+ - x^-$.*

By virtue of the theorem 1 it is easy to specify methods for solving problems 1 and 2.

Statement 1. *Problem 1 has a solution if and only if the following system of linear inequalities has a solution*

$$\begin{bmatrix} \overline{P} \\ -\underline{Q} \end{bmatrix} w^+ - \begin{bmatrix} \underline{P} \\ -\overline{Q} \end{bmatrix} w^- \leqslant \begin{bmatrix} 1 \\ -1 \end{bmatrix}, w^+, w^- \geqslant 0. \tag{4}$$

In this case, $w = w^+ - w^-$.

Statement 2. *If $||\cdot||$ is an absolute vector norm, then Problem 2 is equivalent to the problem*

$$\|\Psi(w)\| = \left\| \begin{bmatrix} P_c w + P_r |w| - 1 \\ -Q_c w + Q_r |w| + 1 \end{bmatrix}_+ \right\| \to \min_w, \tag{5}$$

where $P_c = (\underline{P} + \overline{P})/2$, $P_r = (\overline{P} - \underline{P})/2$, $Q_c = (\underline{Q} + \overline{Q})/2$, $Q_r = (\overline{Q} - \underline{Q})/2$, $|\cdot|$ is an absolute value operation applied to the vector argument element by element. Note that (5), as well as (3), is a convex nonsmooth unconstrained minimization problem that has a unique solution when using a strictly convex norm. Consider the problems

$$\left\| \left[\begin{bmatrix} P \\ -Q \end{bmatrix} w - \begin{bmatrix} 1 \\ -1 \end{bmatrix} + \gamma \cdot 1 \right]_+ \right\| \to \min_w (= \delta_\gamma), \tag{6}$$

$$\left\| \left[\begin{bmatrix} P_c \\ -Q_c \end{bmatrix} w + \begin{bmatrix} P_r \\ Q_r \end{bmatrix} |w| - \begin{bmatrix} 1 \\ -1 \end{bmatrix} + \gamma \cdot 1 \right]_+ \right\| \to \min_w (= \delta_\gamma), \tag{7}$$

where $\gamma > 0$ is a scalar parameter.

The problems (6) and (7) are relevant in the case of linearly separable sets (respectively exact or having interval uncertainty). It is obvious that $\delta_\gamma = 0$ for $\gamma = 0$ and the problems (6) and (7) have an infinite set of solutions. At the same time, it can be shown that starting from a certain value $\gamma > 0$, the condition $\delta_\gamma > 0$ will be fulfilled and the problems mentioned above will have a unique solution when using a strictly convex norm.

Figure 1 illustrate the results of solving model problems 7 and 5 (with interval uncertainty of data) at $n = 2$ and using the Euclidean norm.

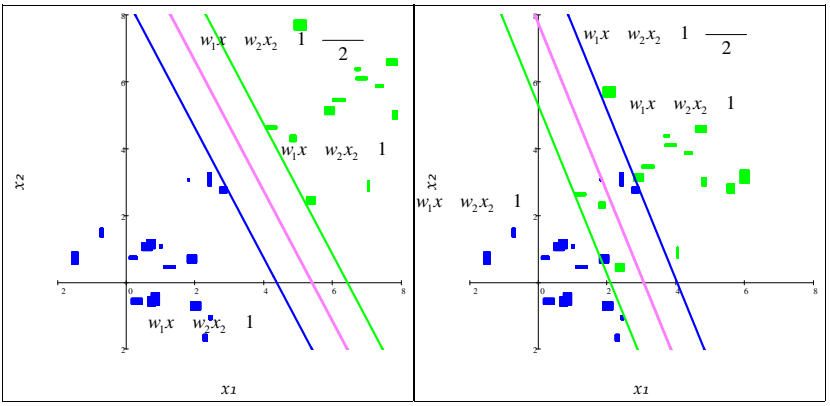


Fig. 1. a) Strict linear separation. b) Unstrict linear separation

The way of finding the parameters α, β is not described due to the limited scope of the paper.

[1] *Vapnik V. N., Chervonenkis A. Ya.* Theory of Pattern Recognition, Nauka, Moscow, 1974 (in Russian).

[2] *Fiedler M., Nedoma J., Ramnik J. et al,* Linear Optimization Problems with Inexact Data, Springer Science+Business Media, Inc., New York, 2006.

[3] *Deisenroth M. P., Faisal A. A., Ong C. S.* Mathematics for Machine Learning. Cambridge University Press, 2020.

[4] *Lankaster P.* Theory of Matrices, Academic Press, New York – London, 1969.

Восстановление пропусков парных сравнений

Двоенко Сергей Данилович[1]⋆ sergedv@yandex.ru
Копылов Андрей Валериевич[1] av.kopylov@yandex.ru
[1]Тула, Тульский государственный университет

Рассматривается известная задача восстановления пропущенных значений в экспериментальных данных. С теоретической точки зрения данная задача оказывается весьма нетривиальной. Проблема восстановления пропущенных значений в данных интенсивно развивается с 70-х годов прошлого века. Для ее решения были предложены различные статистические и регрессионные модели на основе оценок параметров соответствующих вероятностных распределений и оценок степени их искажений из-за потерянных измерений.

В тоже время на практике, восстановление пропущенных значений часто основано на применении известной неформальной гипотезы компактности. Данная гипотеза является важной парадигмой интеллектуального анализа данных. В соответствии с ней предполагается, что исследуемое явление может находиться в конечном числе состояний, где результаты измерений его характеристик косвенно представляют эти состояния. На основе такого подхода были предложены различные локальные параметрические и непараметрические подходы, алгоритмы блочной аппроксимации и т.д.

Следует отметить, что упомянутые подходы нацелены на восстановление пропущенных значений непосредственно измеренных характеристик, представленных в традиционной матрице данных 'объекты-признаки'. Если проблема восстановления пропусков на этом этапе анализа решена, то дальнейшие преобразования данных выполняются без затруднений, когда вычисляются расстояния или близости между элементами множества (объектами), представляющими состояния исследуемого явления.

В данной работе рассмотрена другая ситуация, когда экспериментальные данные представлены только в виде парных сравнений между элементами множества матрицами расстояний или близостей, и некоторые из парных сравнений утрачены. Если оказывается, что повторить измерения исходных характеристик невозможно, а исходная матрица данных недоступна, то невозможно повторно вычислить соответствующее расстояние или близость.

Такие условия возникают, например, при работе измерительных комплексов типа Kinect, когда из-за сбоев возникают пропуски измерений некоторых характеристик, затрудняющих последующие сравнения различных состояний объекта [1]. В другом случае необходимость восстановления пропусков в данных возникает в задаче улучшения видимости на изображениях, полученных в условиях тумана, при наличии локализованных источников освещения, которые не позволяют получить оценку карты рассеивания в соответствующих им областях [2].

Следовательно, возникает проблема восстановления утраченных парных сравнений, для решения которой необходимо разработать соответствующие новые методы.

В работе предложен подход [3] на основе восстановления допустимых значений скалярных произведений между элементами множества, погруженных в евклидово пространство при условии сохранения положительной определенности соответствующей матрицы скалярных произведений. В данном подходе близости понимаются как скалярные произведения между элементами множества, когда все они расположены только в одном квадранте метрического пространства. Восстановление расстояний основано на переходе от скалярных произведений к расстояниям на основе теоремы косинусов.

Предложенный метод непосредственного восстановления пропущенных парных сравнений является примером применения разработанной практической технологии коррекции метрических нарушений в данных, представленных парными сравнениями [4,5].

Работа поддержана грантами РФФИ No. 20-07-00055, 20-07-00441.

[1] *Seredin O. S., Kopylov A. V., Huang S. -C., Rodionov D. S.* A Skeleton Features-Based Fall Detection Using Microsoft Kinect v2 with One Class-Classifier Outlier Removal // ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences., 4212, 189–195, (2019).

[2] *Filin A., Gracheva I., Kopylov A., Seredin O.* Fast Channel-Dependent Transmission Map Estimation for Haze Removal with Localized Light Sources // Artificial Intelligence in Data and Big Data Processing. ICABDE 2021, LNDECT, Springer, Cham., 124, 461–471 (2021).

[3] *Dvoenko S. D.* Recovering Missing Values of Paired Comparisons // Pattern Recognit. Image Anal., 32(3), 522–527 (2022).

[4] *Dvoenko S. D., Pshenichny D. O.* On Metric Correction and Conditionality of Raw Featureless Data in Machine Learning // Pattern Recognit. Image Anal., 28(4), 595–604 (2018).

[5] *Dvoenko S., Pshenichny D.* Metric Correction of Similarities Based on Orthgonal Decomposition // 15th Int. Conf. PRIP'2021, Minsk, UIIP NASB, pp. 19–21 (2021).

Restoring the missing paired comparisons

Dvoenko Sergey[1]⋆ sergedv@yandex.ru
Kopylov Andrey[1] av.kopylov@yandex.ru
[1]Tula, Tula State University

The well-known problem of recovering missing values in experimental data is considered. From a theoretical point of view, this problem appears to be rather nontrivial. The problem of recovering missing values in data has been intensively developed since the 70s of the last century. To solve it, various statistical and regression models were proposed based on the parameter evaluation of the corresponding probability distributions and estimations of the degree of their distortion because of lost measurements.

At the same time, in practice, the recovery of missing values is often based on the so-called compactness informal hypothesis. This hypothesis is an important paradigm of intelligent data analysis. In accordance with it, it is assumed that the phenomenon under study can be in a finite number of states, where the results of measurements of its characteristics indirectly represent these states. Based on this idea, various local parametric and non-parametric approaches, block approximation algorithms, etc. have been proposed.

It should be noted that the mentioned approaches are aimed at recovering the missing values of the directly measured characteristics presented in the traditional 'objects-features' data matrix. If the problem of recovering missing values at this stage of the analysis is solved, then further data processing is performed without difficulties. In this case, all the distances or similarities between the elements of the set (objects) representing the states of the phenomenon under study are calculated.

In this paper, another situation is considered, when the experimental data are presented only by pairwise comparisons between the elements of the set in the form of matrices of distances or similarities, and some pairwise comparisons have been lost. If it is not possible to repeat the measurements of the original features, or the original data matrix is not available, then it is not possible to recalculate the corresponding distances or similarities.

Such conditions can arise, for example, in measuring complexes of the Kinect type. Due to failures, the missing values of some characteristics make difficult the subsequent comparisons of the various object states [1]. In another case, it is necessary to restore the missing values in the problem of improving visibility in images obtained in foggy conditions. In the presence of localized light sources, they do not allow obtaining an estimation of the scatter map in the respective areas [2].

Therefore, the problem of restoring the lost pairwise comparisons needs to be solved, and it is necessary to develop the appropriate new methods.

The paper proposes an approach [3] based on restoring the correct values of scalar products between elements of a set immersed in Euclidean space, subjected to the condition that the corresponding matrix of scalar products is positive definite.

In this approach, similarities are treated as scalar products between the elements of the set, when all of them are located in only one quadrant of the metric space. The recovery of distances is based on the transformation from scalar products to distances based on the law of cosines.

The proposed method is an example of solving the practical problem based on the developed technology of the correction of metric violations in data represented by paired comparisons [4,5].

This research is funded by RFBR, grants 20-07-00055, 20-07-00441.

[1] *Seredin O. S., Kopylov A. V., Huang S. -C., Rodionov D. S.* A Skeleton Features-Based Fall Detection Using Microsoft Kinect v2 with One Class-Classifier Outlier Removal // ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences., 4212, 189-195, (2019).

[2] *Filin A., Gracheva I., Kopylov A., Seredin O.* Fast Channel-Dependent Transmission Map Estimation for Haze Removal with Localized Light Sources // Artificial Intelligence in Data and Big Data Processing. ICABDE 2021, LNDECT, Springer, Cham., 124, 461-471 (2021).

[3] *Dvoenko S. D.* Recovering Missing Values of Paired Comparisons // Pattern Recognit. Image Anal., 32(3), 522-527 (2022).

[4] *Dvoenko S. D., Pshenichny D. O.* On Metric Correction and Conditionality of Raw Featureless Data in Machine Learning // Pattern Recognit. Image Anal., 28(4), 595-604 (2018).

[5] *Dvoenko S., Pshenichny D.* Metric Correction of Similarities Based on Orthgonal Decomposition // 15th Int. Conf. PRIP'2021, Minsk, UIIP NASB, pp. 19-21 (2021).

Разнородный кластерный ансамбль: вероятностная модель, степень разнообразия и оценка качества

Бериков Владимир Борисович[1] berikov@math.nsc.ru
[1]Новосибирск, Институт математики им. С.Л. Соболева СО РАН

В кластерном анализе требуется получить разбиение некоторого множества объектов на относительно небольшое число однородных подмножеств (групп, кластеров, классов). Число групп может быть известно заранее или должно быть определено автоматически. Под критерием однородности разбиения понимается некоторый функционал, зависящий от описаний объектов, например показателей внутригруппового и межгруппового разброса.

В задачах классификации и прогнозирования активно развивается подход, основанный на коллективном принятии решений. При этом итоговое решение определяется на основе нескольких вариантов, полученных различными алгоритмами либо одним алгоритмом, с разными параметрами работы. Коллективный (ансамблевый) подход в кластерном анализе позволяет повышать устойчивость результатов группировки в случае неопределенности в выборе параметров, проводить обработку больших объемов данных (анализируя по отдельности сравнительно небольшие их части), а также использовать «простые» вычислительно эффективные алгоритмы (например, направленные на поиск кластеров сферической формы) для обнаружения сложных структур данных.

Существует несколько основных направлений в методах построения коллективных решений кластерного анализа. В данной работе рассматривается направление, основанное на использовании *коассоциативных матриц* (называемых также *матрицами попарных совпадений, матрицами смежности, co-occurrence matrix*), устанавливающих, как часто каждая пара объектов оказывается в одном и том же кластере (или в различных кластерах) по всем вариантам разбиения. Использование такого вида матриц позволяет решить проблему взаимного соответствия кластеров в вариантах группировки: поскольку нумерация кластеров внутри каждой кластеризации является субъективной, любые перестановки меток кластеров эквивалентны.

Элементы усредненной коассоциативной матрицы могут рассматриваться как меры попарного расстояния (сходства) между объектами: чем чаще пара объектов была объединена алгоритмами, входящими в ансамбль, в один кластер, тем более похожими являются данные объекты. Для получения итогового консенсусного разбиения используется какой-либо из алгоритмов кластерного анализа, основанный на попарном сходстве, например, спектральный алгоритм кластерного анализа.

В работе [1] проведено теоретическое и экспериментальное исследование разнородного кластерного ансамбля, основанного на наборе различных алгоритмов кластерного анализа. Коллективное решение строится путем анализа усредненной коассоциативной матрицы, при нахождении которой учитываются оценки

качества полученных вариантов группировки. Для обоснования разработанного метода предложена вероятностная модель ансамблевой классификации, учитывающая коррелированность оценочных функций. В модели делается предположение о существовании «истинных» непосредственно не наблюдаемых классов, которое позволяет вывести оценки качества работы ансамбля.

С помощью модели получены аналитические зависимости между оценками качества решения и характеристиками ансамбля (числом его элементов, ожидаемым значением и дисперсией индекса качества, показателями коррелированности алгоритмов).

Проведено исследование влияния коррелированности базовых решений ансамбля на его качество. Показано, что учет коррелированности позволяет объяснить улучшение качества ансамбля при увеличении степени разнообразия вариантов разбиения, что ранее было экспериментально установлено в ряде работ.

В рамках модели найдено выражение для оптимальных весов, для которых минимальна верхняя граница оценки вероятности ошибки классификации. Разработан алгоритм, в котором реализован метод построения ансамбля и вычисления оптимальных весов; проведено его экспериментальное исследование.

Работа поддержана грантом РНФ No. 22-21-00261.

[1] *Бериков В. Б.* Модель и метод построения разнородного кластерного ансамбля // Автоматика и телемеханика, 2022. — Т. 12, С. 89–107.

Heterogeneous cluster ensemble: probabilistic model, measure of diversity and quality estimate

Berikov Vladimir[1] berikov@math.nsc.ru

[1]Novosibirsk, Sobolev Institute of Mathematics SB RAS

In cluster analysis, it is required to obtain a partition of a certain set of objects into a relatively small number of homogeneous subsets (groups, clusters, classes). The number of groups may be known in advance or must be determined automatically. The criterion of the homogeneity for the partition is understood as a certain functional that depends on the descriptions of objects, for example, characteristics of intra-group and inter-group variance.

In the problems of classification and forecasting, an approach based on a collective decision-making is being actively developed. In this case, the final solution is determined on the basis of several variants obtained by different algorithms or by one algorithm, with different parameters. The collective (ensemble) approach makes it possible to increase the stability of the grouping results in case of uncertainty in the choice of parameters, to process large amounts of data (by separate analysis of relatively small sub-samples), and also to use "simple" computationally efficient algorithms (for example, aimed at finding spherical clusters) to discover complex data structures.

There are several main directions in the methods for constructing ensemble solutions in cluster analysis. In this paper, we consider a direction based on the use of *co-association matrices* (also called *matrices of pairwise coincidences, adjacency matrices, co-occurrence matrices*), establishing how often each pair of objects appears in the same cluster (or in different clusters) over all partitioning variants. Using this kind of matrices allows solving the problem of mutual correspondence of clusters in grouping variants: since the numbering of clusters within each partition is subjective, any permutations of cluster labels are equivalent.

The elements of the averaged co-association matrix can be considered as measures of pairwise distance (or similarity) between objects: the more often a pair of objects was combined into one cluster by the algorithms included in the ensemble, the more similar these objects are. To obtain the final consensus partition, one of the clustering algorithms based on pairwise similarity is used, for example, spectral clustering.

In [1], a theoretical and experimental study of a heterogeneous cluster ensemble based on a set of different clustering algorithms is carried out. An ensemble solution is built by analyzing the averaged co-association matrix, which takes into account the quality estimates of the obtained partition variants. To substantiate the developed method, a probabilistic model of ensemble classification is proposed that takes into account the correlation of evaluation functions. The model makes an assumption about the existence of "true" not directly observable classes, which allows us to derive estimates of the performance of the ensemble.

With the help of the model, analytical dependencies between the estimates of the quality of the solution and the characteristics of the ensemble (the number of its elements, the expected value and variance of the quality index, the correlation characteristics of algorithms) are obtained.

A study was made on the influence of the correlation of the basic elements of the ensemble on its quality. It is shown that taking into account the correlation makes it possible to explain the improvement in the quality of the ensemble with an increase in the degree of diversity of partitioning variants, which was previously experimentally established in a number of works.

Within the framework of the model, an expression is found for the optimal weights for which the minimum upper bound on the estimate of the probability of a classification error is reached.

An algorithm has been developed in which the method of constructing an ensemble and calculating the optimal weights is implemented; an experimental study of the algorithm is performed.

This research is funded by RSF, grant 22-21-00261.

[1] *Berikov V.* Model and method for constructing a heterogeneous cluster ensemble // Automation and Remote Control, 2022. — Vol. 12, p. 89–107.

О числе решений некоторых специальных задач поиска в данных частых и нечастых элементов

Дюкова Елена Всеволодовна[1] edjukova@mail.ru
Дюкова Анастасия Петровна[1]⋆ anastasia.d.95@gmail.com
[1]Москва, ФИЦ ИУ РАН

Исследуются вопросы сложности логического анализа целочисленных данных. Для специальных задач поиска в данных частых и нечастых элементов, на решении которых базируется обучение логических процедур классификации, приведены асимптотики типичного числа решений.

Рассматриваемые задачи поиска в данных частых и нечастых элементов формулируются следующим образом.

Исследуется множество объектов M. Каждый объект из M может быть представлен в виде числового вектора, полученного на основе наблюдения или измерения ряда его характеристик. Такие характеристики называют атрибутами. Предполагается, что каждый атрибут имеет ограниченное множество допустимых значений, которые кодируются целыми числами.

Пусть $X = \{x_1, \ldots, x_n\}$ — множество атрибутов; $H = \{x_{j_1}, \ldots, x_{j_r}\}$ — набор из r, $r \leqslant n$, различных атрибутов; $\sigma = (\sigma_1, \ldots, \sigma_r)$ — набор, в котором σ_i — допустимое значение признака x_{j_i}, $i = 1, 2, \ldots, r$. Пара (σ, H) *называется элементарным фрагментом (ЭФ)* длины r. Через $W(M, X)$ обозначим множество всех ЭФ.

Пусть $S = (a_1, \ldots, a_n)$ — объект из M (здесь a_j, $j \in \{1, 2, \ldots, n\}$, — значение атрибута x_j для объекта S). Объект S *содержит* ЭФ (σ, H), если $a_{j_i} = \sigma_i$ при $i = 1, 2, \ldots, r$.

Дана некоторая совокупность объектов D из M и задано число p, $0 < p \leqslant 1$. Через $|D|$ обозначается число объектов в D. ЭФ (σ, H), $(\sigma, H) \in W(M, X)$, называется (p, D)-*частым*, если не менее $p|D|$ объектов S' из D содержат ЭФ (σ, H). Иначе ЭФ (σ, H) — (p, D)-*нечастый*. ЭФ (σ, H), $(\sigma, H) \in W(M, X)$, называется $(0, D)$-*нечастым*, если ни один объект из D не содержит (σ, H).

ЭФ (σ, H), являющийся (p, D)-частым в $W(M, X)$, называется *максимальным* (p, D)-частым в $W(M, X)$, если любой ЭФ (σ', H') из $W(M, X)$ такой, что $\sigma' \supset \sigma$, $H' \supset H$, не является (p, D)-частым. ЭФ (σ, H), являющийся (p, D)-нечастым в $W(M, X)$, называется *минимальным* (p, D)-нечастым в $W(M, X)$, если любой ЭФ (σ', H') из $W(M, X)$ такой, что $\sigma' \subset \sigma$, $H' \subset H$, не является (p, D)-нечастым. Понятие минимального $(0, D)$-нечастого ЭФ полностью аналогично введённому понятию минимального (p, D)-нечастого ЭФ для $p > 0$.

Возникают две отдельные задачи: 1) для заданного p, $0 < p \leqslant 1$, найти в $W(M, X)$ все (максимальные) (p, D)-частые ЭФ; 2) для заданного q, $0 \leqslant q \leqslant 1$ найти все (минимальные) (q, D) -нечастые ЭФ. Иногда требуется совместное перечисление максимальных частых и минимальных нечастых ЭФ.

Задачи поиска в данных частых и нечастых элементов являются одними из центральных задач интеллектуального анализа данных и особенно важны в случае больших данных. Эти задачи актуальны для многих прикладных областей, среди которых следует выделить нахождение в данных ассоциативных правил и машинное обучение.

В первом случае D называют базой данных, а каждый объект из D называют транзакцией. Ассоциативное правило (АП) устанавливает зависимость между двумя частыми ЭФ, согласно которой один частый ЭФ X (посылка) с некоторой «достоверностью» влечёт другой частый ЭФ Y. При этом ЭФ X и Y порождаются одним общим частым ЭФ, обозначаемым (X, Y). Наиболее информативными считаются те АП, которые порождаются максимальными частыми ЭФ (X, Y) с «минимальной» посылкой X. Вопросы поиска ассоциативных правил наиболее изучены в случае бинарных данных [1].

Одной из главных задач машинного обучения является задача классификации на основе прецедентов. В этом случае D — обучающая выборка (заданная совокупность примеров объектов из M), а каждый объект из D — обучающий объект или прецедент. Подлежащие измерению или наблюдению свойства исследуемых объектов называются признаками. В самом простом случае прецеденты делятся на два класса (класс положительных и класс отрицательных примеров). В общем случае число классов может быть больше двух. Требуется по признаковому описанию предъявленного объекта, о котором заранее неизвестно, какому классу он принадлежит, определить (распознать) этот класс.

Хорошие результаты показывают логические классификаторы, при конструировании которых используются как основные идеи алгоритма «Кора» [2], так и алгоритмов вычисления оценок [3]. Эти классификаторы впервые предложены в [4]. В алгоритмах типа «Кора» анализ прецедентной информации проводится в предположении, что признаковые описания любых двух обучающих объектов, принадлежащих разным классам, не совпадают. На этапе обучения для каждого класса K ищутся так называемые (p, q)-представительные элементарные классификаторы, представляющие собой специальные ЭФ из $W(M, X)$.

Пусть $Q(K)$ и $Q(\overline{K})$ — множества прецедентов из класса K и не из класса K соответственно и $p > 0$, $q < p$. Тогда (p, q)-представительный элементарный классификатор класса K является одновременно (максимальным) $(p, Q(K))$-частым ЭФ и (минимальным) $(q, Q(\overline{K})$ -нечастым ЭФ в $W(M, X)$. Как правило сначала строятся минимальные $(q, Q(\overline{K}))$-нечастые ЭФ, а затем из них отбираются те, которые являются $(p, Q(K))$-частыми. На следующем этапе найденные (p, q)-представительные элементарные классификаторы класса K участвуют в процедуре «голосования» за отнесение распознаваемого объекта к этому классу. Материал обучения безошибочно классифицируется при $q = 0$. Однако нахождение $(0, Q(\overline{K}))$-нечастых ЭФ требует больших вычислительных затрат. В случае бинарных данных, когда требуется найти все минимальные $(0, Q(\overline{K}))$-

нечастые ЭФ вида $(0, \ldots, 0)$, это известная труднорешаемая перечислительная задача, называемая монотонной дуализацией.

Отметим, что к задаче совместного перечисления максимальных частых и минимальных нечастых ЭФ сводится задача расшифровки монотонной функции [5].

В целях исследования скорости решения рассматриваемых задач в случае больших данных представляет интерес получение асимптотик типичного числа частых и нечастых ЭФ, а также типичной длины таких ЭФ. В работе [6] искомые оценки приведены для множества минимальных $(0, D)$-нечастых ЭФ, называемых в этой работе тупиковыми покрытиями целочисленной матрицы. В настоящей работе так же, как и в [6] рассмотрен случай большого числа атрибутов. Требуемые оценки получены для (p, D)-частых ЭФ специального вида в предположении, что каждый атрибут имеет $k, k \geqslant 2$, допустимых значений. Сравнение полученных в настоящей работе оценок с оценками из [6] свидетельствует о перспективности применения методов поиска частых ЭФ для построения $(p, 0)$-представительных элементарных классификаторов.

Результаты настоящей работы согласуются с экспериментальными исследованиями, приведёнными в [7]. Построенные в [7] модели классификаторов с ориентацией на методы поиска частых ЭФ позволяют при определённых условиях существенно сократить время обучения.

[1] *Aggarwal C.* Frequent Pattern Mining // Springer International Publishing, 2014. P. 467.

[2] *Вайнцвайг М. Н.* Алгоритм обучения распознаванию образов «Кора» // Алгоритмы обучения распознаванию образов / Под ред. В. Н. Вапник. — М.: Советское радио, 1973. С. 110–116.

[3] *Журавлёв Ю. И.* Об алгебраическом подходе к решению задач распознавания и классификации // Проблемы кибернетики. — М.: Наука, 1978. Вып. 33. С. 5–68.

[4] *Баскакова Л. В., Журавлёв Ю. И.* Модель распознающих алгоритмов с представительными наборами и системами опорных множеств // Ж. вычисл. матем. и матем. физ., 1981. Т. 21. №5. С. 1264–1275.

[5] *Dragunov N. and Djukova E.* Finding frequent and infrequent elements of partial orders product and the problem of two-valued monotone function decoding // IEEE Proceedings of the VII International Conference on Information Technology and Nanotechnology (ITNT-2021), Samara, Russia, 2021. P. 1–5.

[6] *Дюкова Е. В., Журавлев Ю. И.* Дискретный анализ признаковых описаний в задачах распознавания большой размерности // Ж. вычисл. матем. и матем. физ., 2000. Т. 40. №8. С. 1264–1278.

[7] *Dragunov N., Djukova E. and Djukova A.* Supervised Classification and Finding Frequent Elements in Data // VIII International Conference on Information Technology and Nanotechnology (ITNT-2022), Samara, Russia, 2022. P. 1–5.

On the number of solutions to some special problems of searching for frequent and infrequent elements

Djukova Elena[1] edjukova@mail.ru
Djukova Anastasia[1]⋆ anastasia.d.95@gmail.com
[1]Moscow, FRC CSC RAS

The issues of the complexity of the integer data logical analysis are investigated. Asymptotic estimates of the typical number of solutions are given to special problems of searching for frequent and infrequent elements in data, on the solution of which the training of logical classification procedures is based.

The considered problems of searching for frequent and infrequent elements in data are formulated as follows.

Let M be the set of objects under examination. Each object from M can be represented as a numerical vector obtained by observing or measuring a number of its characteristics. Such characteristics are called attributes. It is assumed that each attribute has a limited number of admissible values, which are encoded by integers.

Let $X = \{x_1, \ldots, x_n\}$ be a set of attributes; $H = \{x_{j_1}, \ldots, x_{j_r}\}$ be a set of r, $r \leqslant n$, various attributes; $\sigma = (\sigma_1, \ldots, \sigma_r)$ be the set in which σ_i be allowed an admissible feature value x_{j_i}, $i = 1, 2, \ldots, r$. The pair (σ, H) is called an *elementary fragment (EF) of length* r. The set of all EFs is denoted by (M, X).

Let $S = (a_1, \ldots, a_n)$ be an object from M (here a_j, $j \in \{1, 2, \ldots, n\}$, be an attribute value x_j for object S). Object S *contains EF* (σ, H), if $a_{j_i} = \sigma_i$ at $i = 1, 2, \ldots, r$.

Some set of objects D from M is given and number p, $0 < p \leqslant 1$ is given. Through $|D|$ denotes the number of objects in D. EF (σ, H), $(\sigma, H) \in W(M, X)$, is called (p, D)-*frequent*, if not less $p|D|$ objects S' from D contain EF (σ, H). Otherwise EF (σ, H) is (p, D)-*infrequent*. EF (σ, H), $(\sigma, H) \in W(M, X)$, is called $(0, D)$-*infrequent*, if no object from D does not contain (σ, H).

The EF (σ, H) is called *maximal* (p, D)-frequent in $W(M, X)$ if it is (p, D)-frequent in $W(M, X)$ and any EF (σ', H') from $W(M, X)$ such that $\sigma' \supset \sigma$, $H' \supset H$ is not (p, D)-frequent. The EF (σ, H) is called *minimal* (p, D)-infrequent in $W(M, X)$ if it is (p, D)-frequent in $W(M, X)$ and any EF (σ', H') from $W(M, X)$ such that $\sigma' \subset \sigma$, $H' \subset H$ is not (p, D)-frequent. The concept of minimal $(0, D)$-infrequent EF is completely similar to the introduced concept of minimal (p, D)-infrequent EF for $p > 0$.

There are two separate tasks: 1) for a given p, $0 < p \leqslant 1$, find in $W(M, X)$ all (maximal) (p, D)-frequent EFs; 2) for a given q, $0 \leqslant q \leqslant 1$, find all (minimal) (q, D)-infrequent EFs. Sometimes a joint enumeration of the maximal frequent and minimal infrequent EFs is required.

Finding frequent and infrequent elements in data is one of the central tasks of data mining and is especially important in the case of big data. These tasks are

relevant for many application areas, among which are finding associative rules in data and machine learning.

In the first case D is called a database and each object from D is called a transaction. Association rule (AR) establishes a relationship between two frequent EFs, according to which one frequent EF X (premise) with some "probability" entails another frequent EF. Wherein EF X and Y are generated by one common frequent EF denoted (X, Y). The most informative are those AR that are generated by the maximum frequent EF (X, Y) with a "minimum" premise. The questions of finding AR are most studied in the case of binary data [1].

One of the main problems of machine learning is supervised classification. In this case D is a training set (some given set of examples of objects from M) , and each object from D is a learning object or a precedents. The properties of the objects under study that are to be measured or observed are called features. In the simplest case, the precedents are divided into two classes (the class of positive and the class of negative examples). In general case, the number of classes may be more than two. Given a description in terms of features of an unknown object, it is required to find out (recognize) the class it belongs to.

Good results are shown by logical classifiers, the construction of which uses both the main ideas of the algorithm "Cora" [2] and of calculating estimates algorithms [3]. These classifiers were first proposed in [4]. In the algorithms of "Cora" type, the analysis of precedent information is carried out under the assumption that the feature descriptions of any two training objects belonging to different classes do not coincide. At the stage of learning for each class K are looking for the so-called (p, q)-representative elementary classifiers, which are special EF from (M, X).

Let $Q(K)$ and $Q(\overline{K})$ be the sets of precedents from the class K and not from the class K, respectively, and $p > 0$, $q < p$. Then (p, q)-representative elementary classifier for class K is both (maximal) $(p, Q(K))$-frequent EF and (minimal) $(q, Q(\overline{K}))$-infrequent EF in (M, X). As a rule, the minimal $(q, Q(\overline{K}))$-infrequent EFs are constructed first, and then those that are $(p, Q(K))$-frequent are selected from them. In the next step, the found (p, q)-representative elementary classifiers of the class K take part in the "voting" procedure for assigning the recognized object to this class. The training material is unmistakably classified at $q = 0$. However, the finding of $(0, Q(\overline{K}))$-infrequent EFs is computationally expencive. In the case of binary data, when you want to find all minimal $(0, Q(\overline{K}))$-infrequent EFs of the kind $(0, 0, \ldots, 0)$, it is a well-known intractability enumeration problem called monotone dualization.

Note that the problem of decoding a monotone function is reduced to the problem of joint enumeration of the maximal frequent and minimal infrequent EFs [5].

In order to investigate the speed of solving the considered problems in the case of large data, it is of interest to obtain asymptotic estimates of the typical number of frequent and infrequent EFs, as well as the typical length of such EFs. In [6] the required estimates are given for the set of minimum $(0, D)$-infrequent EFs, which are

called irredundant coverings of an integer matrix in this work. In presented paper, as well as in [6], the case of a large number of attributes is considered. The necessary asymptotic estimates are obtained for (p, D)-frequent EFs of a special type under assumption that each attribute has k, $k \geqslant 2$, admissible values. Comparison of these estimates with those from [6] indicates that the methods of searching for frequent EFs are promising for constructing $(p, 0)$-representative elementary classifiers.

The results of this work are consistent with the experimental studies presented in [7]. The classifier models in [7] with a focus on methods for searching for frequent EFs can, under certain conditions, significantly reduce the learning time.

[1] *Aggarwal C.* Frequent Pattern Mining // Springer International Publishing, 2014. — p. 467.

[2] *Weinzweig M. N.* Algoritm obucheniya raspoznavaniyu obrazov “Kora” // Algorithms for learning pattern recognition / Edited by V. N. Vapnik. — Soviet Radio, Moscow. 1973. P. 110–116. [in Russian].

[3] *Zhuravlev Yu. I.* Ob algebraicheskom podhode k resheniyu zadach raspoznavaniya i klassifikacii // Problems of Cybernetics. — Nauka, Moscow. 1978. 33: 5–68. [in Russian].

[4] *Baskakova L. V. and Zhuravlev Yu. I.* Model of Recognition Algorithms with Representative Sampls and Systems of Supporting Sets // Comp. Math. Math. Phys. 1981. 21(5): 189–199.

[5] *Dragunov N. and Djukova E.* Finding frequent and infrequent elements of partial orders product and the problem of two-valued monotone function decoding // IEEE Proceedings of the VII International Conference on Information Technology and Nanotechnology (ITNT-2021), Samara, Russia, 2021. P. 1–5.

[6] *Djukova E. V. and Zhuravlev Yu. I.* Discrete analysis of feature descriptions in recognition problems of high dimensionality // Comp. Math. Math. Phys. 2000. 40(8): 1214–1227.

[7] *Dragunov N., Djukova E. and Djukova A.* Supervised Classification and Finding Frequent Elements in Data // VIII International Conference on Information Technology and Nanotechnology (ITNT-2022), Samara, Russia, 2022. P. 1–5.

Построение гауссовых пирамид изображений на основе решеточных уравнений Больцмана

Ильин Олег Вадимович[1] oilyin@gmail.com

[1] Москва, Федеральное государственное учреждение "Федеральный исследовательский центр "Информатика и управление"Российской академии наук"

Гауссовы пирамиды (ГП) являются одним из главных компонентов масштабно инвариантных методов обработки изображений, таких как SIFT и SURF [1]-[2]. Пирамида Гаусса представляет собой множество, состоящее из исходного изображения, его размытий по Гауссу (наложений фильтра Гаусса) и масштабных преобразований (апсемплинг и даунсэмплинг). Стандартный способ размытия по Гауссу основан на следующей дискретной свертке

$$I(x,y,\sigma) = \sum_{x',y'=-K}^{x',y'=K} A \exp\left(-\frac{(x'^2+y'^2)}{2\sigma^2}\right) I(x-x', y-y'),$$

где A — нормировочная константа для дискретного гауссова ядра, x, y — координаты пикселей двумерного изображения, $I(x,y)$ и $I(x,y,\sigma)$ — интенсивность цвета основного и размытого изображения (в градациях серого) соответственно, K зависит от σ и в большинстве случаев равно $round(3\sigma)$, также $round$ обозначает операцию округления до ближайшего целого числа.

В литературе указывается, что построение ГП занимает примерно 80% вычислительного времени для метода SIFT [4]. На практике размер окна свертки равен $2K+1 = round(3\sigma)+1$, это означает, что сложность вычислений растет с ростом σ. Например, метод SIFT требует выполнения сверток шириной $9, 11, 13, 15, 19$ на каждую октаву. Отметим также, что свертка является нелокальной процедурой, и очевидно, что необходима модификация метода (например, экстраполяция данных интенсивности I за границы изображения) вблизи границ изображения.

С другой стороны, хорошо известно, что для непрерывных переменных x, y, σ функция $I(x,y,\sigma)$ есть решение уравнения диффузии [3]

$$\frac{\partial I}{\partial \sigma^2} = D\left(\frac{\partial^2 I}{\partial x^2} + \frac{\partial^2 I}{\partial y^2}\right), \tag{1}$$

где коэффициент диффузии D равен $1/2$.

Целью настоящего исследования является разработка альтернативного подхода к гауссовскому размытию (и построению ГП) на основе решеточных моделей Больцмана (LB). Настоящая модель LB предназначена для решения уравнения диффузии без применения сверток. Подход LB основан на двух компонентах: декартовой решетке (Рис.1) с координатами (x,y), соответствующей изображению со степенью размытия σ, и модели, описывающей процесс диффузии.

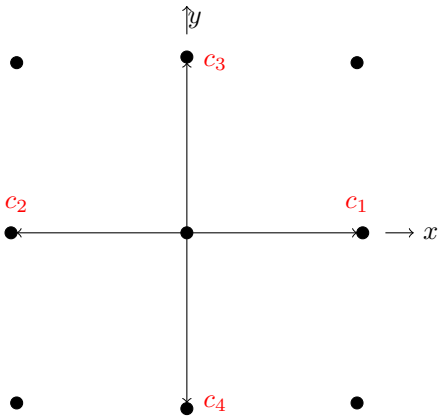


Рис. 1. Решетка модели D2Q4. Узлы решетки (расположение пикселей) обозначены кружками. Стрелки обозначают возможные направления движения частиц.

Эта модель эквивалентна уравнению, описывающему движение частиц (виртуальных) со скоростями $\boldsymbol{c}_{1,2} = (\pm c, 0), \boldsymbol{c}_{3,4} = (0, \pm c)$ (c — абсолютное значение скорости) по узлам решетки. Здесь предполагается, что расположение узлов решетки совпадает с расположением пикселей. Концентрации частиц, определяемые как $f_i(\sigma, x, y), i = 1 \dots 4$, соответствуюшие скоростям c_i, подчиняются уравнению LB $D2Q4$ (две пространственные координаты, четыре скорости) [5], в настоящей работе это уравнение переписывается в следующем виде

$$f_i(\sigma + \delta\sigma, x + c_{i,x}\delta\sigma, y + c_{i,y}\delta\sigma) = f_i\big(\sigma, x, y\big) + \frac{\delta\sigma}{4\tau}\big(I(\sigma, x, y) - f_i(\sigma, x, y)\big), \qquad (2)$$

где $i = 1 \dots 4$, также $c_{i,x}, c_{i,y}$ есть проекции скорости $\boldsymbol{c_i}$ на оси x, y, интенсивность изображения определяется из формулы

$$I(\sigma, x, y) = \sum_i f_i(\sigma, x, y),$$

где τ равна $2D/c^2 + (\delta\sigma)^2/2$. Схема (2) имеет второй порядок точности аппроксимации уравнения диффузии (1).

Применяя модель (2), можно вычислить следующее размытое изображение (со степенью размытия $\sigma + \delta\sigma$) из исходного изображения (со степенью размытия σ) для всех пикселей x, y (узлы решетки), если известны $f_i(\sigma, x, y)$. Так как для исходного изображения задана только интенсивность $I(\sigma, x, y)$, то для получения точного решения необходимо вычислить соответствующие значения $f_i(\sigma, x, y)$. Этого можно добиться, используя методы, использующиеся при моделировании гидродинамических течений [6]. Кроме того, в граничных узлах применяется экстраполяционное краевое условие второго порядка.

Предлагаемый метод высокоэффективен. Во-первых, метод LB примерно в $1,4-2$ быстрее функции GaussianBlur из пакета OpenCV (тест проводился для изображений с разрешением 500×500, результат зависит от степени размытия, программа написана на C++). Представленная схема легко программируется, предлагаемый подход значительно проще, чем метод быстрого преобразования Фурье, который часто используется для вычисления сверток (особенно с большой шириной размытия). В заключение также следует подчеркнуть, что модель LB показывает отличную масштабируемость на многоядерных компьютерах.

[1] *Lowe D.* Distinctive Image Features from Scale-Invariant Keypoints // International Journal of Computer Vision, 2004. — Vol. 60 — p. 91—110.

[2] *Bay H., Ess A., Tuytelaars T., and Van Gool L.* Speeded-up robust features (SURF) // Computer Vision and Image Understanding, 2008. — Vol. 110 — p. 346–359.

[3] *Lindeberg T.* Feature Detection with Automatic Scale Selection // International Journal of Computer Vision, 1998. — Vol. 30 — p. 79—116.

[4] *Huang F-C. et al* High-performance SIFT hardware accelerator for real-time image feature extraction // IEEE Transactions on Circuits and Systems for Video Technology, 2012. — Vol. 22 — p. 340–351.

[5] *Suga S.* Stability and accuracy of lattice Boltzmann schemes for anisotropic advection-diffusion equations //International Journal of Modern Physics C, 2009. — Vol. 20 — p. 633–650.

[6] *Ilyin O.* Discrete-velocity Boltzmann model: Regularization and linear stability // Physical Review E, 2022. — Vol. 105 — p. 045312

Development of image Gaussian pyramids using lattice Boltzmann method

Ilyin Oleg[1] oilyin@gmail.com

[1] Moscow, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences

Gaussian pyramid (GP) is one of the major components of several image processing approaches like scale-invariant feature detection SIFT and SURF methods [1]-[2]. Gaussian pyramid is a set consisting of blurred and resampled (upsampled and downsampled in a special way) images. A most common way to perform the Gauss blur is based on the following discrete convolution

$$I(x,y,\sigma) = \sum_{x',y'=-K}^{x',y'=K} A \exp\left(-\frac{(x'^2+y'^2)}{2\sigma^2}\right) I(x-x', y-y'),$$

where A is the normalization constant for the discrete Gauss kernel, x, y are the pixel coordinates in a two-dimensional image, $I(x,y)$ and $I(x,y,\sigma)$ are the base and blurred image color intensity (in grey scale) respectively, K depends on σ and in most cases equals $round(3\sigma)$, and $round$ is closest integer function.

It has been reported previously that the construction of the GP takes approximately 80% of the computation time for SIFT method [4]. In practice, the size of the convolution window is $2K+1 = round(3\sigma)+1$, this means that computation complexity is growing with σ. For instance, SIFT method requires to perform the convolutions of width equal $9, 11, 13, 15, 19$ per octave. Moreover, the convolution is non-local procedure, it is obvious that a special treatment at the vicinity of the image boundaries is needed.

On the other hand, it is well known that for the continuous x, y, σ variables the image intensity $I(x,y,\sigma)$ obeys the diffusion equation [3]

$$\frac{\partial I}{\partial \sigma^2} = D\left(\frac{\partial^2 I}{\partial x^2} + \frac{\partial^2 I}{\partial y^2}\right), \tag{1}$$

where the diffusion coefficient D equals $1/2$.

The goal of the present study is to develop an alternative approach for the gaussian blurring (and constructing GP) based on the lattice Boltzmann (LB) models. The present LB model aims to solve the diffusion equation without the application of convolutions. LB approach is based on two components: a cartesian lattice (Fig.1) with coordinates (x,y) corresponding an image with blur level σ and a model which describes the diffusion process. This model is equivalent to the equation governing the dynamics of particles (virtual) with the velocities $\boldsymbol{c}_{1,2} = (\pm c, 0), \boldsymbol{c}_{3,4} = (0, \pm c)$ (c is the velocity absolute value) hopping over lattice nodes. In here, it is assumed that the location of the lattice nodes coincide with the location of pixels. The particle

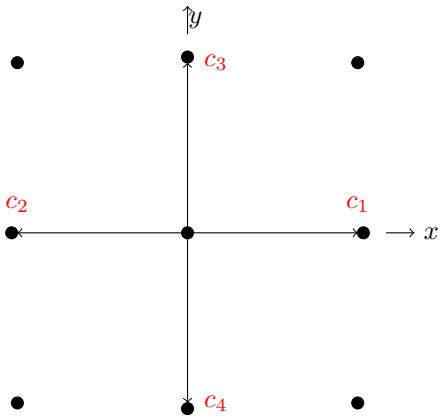


Fig. 1. Lattice for D2Q4 model. Lattice nodes (pixel locations) are denoted by circles. Arrows define possible particle motions.

concentrations defined as $f_i(\sigma, x, y), i = 1 \dots 4$ traveling with the velocities c_i are governed by $D2Q4$ (two-dimensional, four velocities) LB equation [5], in the present study this equation is rewritten in the following form

$$f_i(\sigma + \delta\sigma, x + c_{i,x}\delta\sigma, y + c_{i,y}\delta\sigma) = f_i\big(\sigma, x, y\big) + \frac{\delta\sigma}{4\tau}\big(I(\sigma, x, y) - f_i(\sigma, x, y)\big), \qquad (2)$$

where $i = 1 \dots 4$ and $c_{i,x}, c_{i,y}$ are projections of $\boldsymbol{c_i}$ on x, y, the intensity (or concentration in terms of LB variables) is given by

$$I(\sigma, x, y) = \sum_i f_i(\sigma, x, y)$$

and τ is computed as $2D/c^2 + (\delta\sigma)^2/2$. It can be shown that the scheme (2) is second order accurate approximation of the diffusion equation (1). Applying LB model (2) one can compute the next blurred image in GP (with the blur level $\sigma + \delta\sigma$) from the initial image (with blur σ) for all pixels x, y (lattice nodes) if $f_i(\sigma, x, y)$ are known. Since for the initial image only the intensity $I(\sigma, x, y)$ is given, then to obtain accurate solution one needs to deduce appropriate values of $f_i(\sigma, x, y)$. This can be performed by adopting the techniques widely used in modeling of hydrodynamic flows [6]. In addition, at the boundary nodes second-order extrapolation boundary condition is applied.

The proposed method is efficient. First, LB method is approximately $1.4 - -2$ times faster than GaussianBlur function from OpenCV package (the test was performed for the images with the resolution 500×500, the result also depends on the blurring levels, the program is written in C++). Next, the presented scheme can be be easily programmed, the proposed method is significantly simpler than the

fast Fourier transform method which is used for computing the convolutions. As a final remark, one should emphasize that LB model shows excellent scalability on multicore computers.

[1] *Lowe D.* Distinctive Image Features from Scale-Invariant Keypoints // International Journal of Computer Vision, 2004. — Vol. 60 — p. 91—110.

[2] *Bay H., Ess A., Tuytelaars T., and Van Gool L.* Speeded-up robust features (SURF) // Computer Vision and Image Understanding, 2008. — Vol. 110 — p. 346–359.

[3] *Lindeberg T.* Feature Detection with Automatic Scale Selection // International Journal of Computer Vision, 1998. — Vol. 30 — p. 79—116.

[4] *Huang F-C. et al* High-performance SIFT hardware accelerator for real-time image feature extraction // IEEE Transactions on Circuits and Systems for Video Technology, 2012. — Vol. 22 — p. 340–351.

[5] *Suga S.* Stability and accuracy of lattice Boltzmann schemes for anisotropic advection-diffusion equations //International Journal of Modern Physics C, 2009. — Vol. 20 — p. 633–650.

[6] *Ilyin O.* Discrete-velocity Boltzmann model: Regularization and linear stability // Physical Review E, 2022. — Vol. 105 — p. 045312

Парето-оптимальные решения в задаче мультимодальной кластеризации

Богатырев Михаил Юрьевич[1] okkambo@mail.ru
Орлов Дмитрий Александрович[1]⋆ di-orl@mail.ru
[1]Тула, ТулГУ

В докладе рассматривается применение многокритериальной оптимизации в задаче мультимодальной кластеризации. Мультимодальная кластеризация выполняется на *мультимодальных данных* - данных, представленных в виде нескольких множеств. Особенность мультимодальной кластеризации в том, что она выполняется одновременно на всех множествах данных. В результате мультимодальный кластер представляет собой сочетание экземпляров данных из разных множеств.

Известной проблемой в кластерном анализе является проблема интерпретации полученных кластеров, решение которой связано с интеллектуализацией обработки данных. Любой алгоритм кластеризации использует некоторую меру близости, определенную на множестве кластеризуемых объектов. Поэтому «смысл» получаемых кластеров определяется используемой мерой близости: объекты входят в один кластер потому, что они близки друг другу согласно выбранной, как правило, числовой мере близости.

Специальный подход к мультимодальной кластеризации используется в Анализе формальных понятий (АФП) [1]. Алгоритмы АФП работают на тензорных представлениях многомерных данных – *формальных контекстах*, и строят на них *решётки понятий* или *мультимодальные кластеры*.

Многомерный формальный контекст представляет собой отношение

$$\mathbb{K} =< K_1, K_2, \ldots, K_n, R > \tag{1}$$

на доменах данных $D_1, D_2, \ldots, D_n$, $K_i \subseteq D_i$. Мультимодальные кластеры на контексте (1) строятся в виде

$$\mathbb{C} =< X_1, X_2, \ldots, X_n > \tag{2}$$

$X_i \subseteq K_i$, и обладают следующим свойством замыкания:

$$\forall u = (x_1, x_2, \ldots, x_n) \in X_1, X_2, \ldots, X_n, u \in R, \tag{3}$$

при этом $\forall j = 1, 2, \ldots, n, \forall x_j \in D_j \backslash X_j < X_1, \ldots, X_j \cup \{x_j\}, \ldots, X_m >$ не удовлетворяет условию (3).

Мультимодальный кластер - это подмножество в виде комбинаций элементов из разных наборов K_i. Он также определяется как замкнутое m-множество, поскольку свойство замыкания (3) обеспечивает его «самодостаточность»: кластер не может быть увеличен без нарушения условия замыкания.

Модальность кластера – это количество подмножеств его образующих, $m \leqslant n$.

Размерность кластера – это количество объектов в кластере, как числовых, так и нечисловых. Формальное понятие – это такой мультимодальный кластер, в котором для всех его элементов выполняется условие

$$u = (x_1, x_2, \ldots, x_k) \in X_1, X_2, \ldots, X_k, u \in R. \quad (4)$$

Формальное понятие представляют собой максимально возможный m-мерный гиперкуб, полностью заполненный элементами из множеств данных $K_i \subseteq D_i$. В АФП введено понятие *плотности мультимодального кластера* и формальные понятия интерпретируются как абсолютно плотные кластеры [2].

Доклад основан на содержании наших работ [3, 4] и содержит новые экспериментальные результаты. Принципиальными положениями, раскрываемыми в докладе, являются следующие.

1. Формальные контексты (1), строящиеся на реальных данных, являются их объектно-признаковыми представлениями. Одно из множеств $K_i \subseteq D_i$ - это объекты реального мира, а другие множества – это атрибуты объектов и множества иных, внешних по отношению к объектам данных. Интерпретация мультимодальных кластеров (2) выполняется в соответствии с двумя их характеристиками: плотностью и размерностью (объемом). Эти характеристики противоречат друг другу: как следует из экспериментов на объектно-признаковых данных, плотные кластеры имеют небольшой объем и наоборот. Поэтому задача мультимодальной кластеризации формулируется как задача многокритериальной оптимизации.

2. Среди решений задачи многокритериальной оптимизации предпочтительны Парето-оптимальные решения. Фронт Парето содержит множество равнозначных решений, на котором исследуются различные варианты интерпретации кластеров.

3. Для построения Парето-оптимальных решений задачи мультимодальной кластеризации целесообразно применять эволюционные алгоритмы. Преимуществами эволюционных алгоритмов являются присущая им множественность решений и возможность управлять эволюцией решений в современных реализациях таких алгоритмов.

В работах [3, 4] используется эволюционный алгоритм многокритериальной оптимизации, который применен в решениях задач мультимодальной кластеризации на нескольких наборах данных. Алгоритм относится к семейству алгоритмов NSGA-II, в которых применяется свойство элитизма в операторах отбора генетических алгоритмов. Также в нашем алгоритме используется *недоминируемая сортировка решений*, что позволяет строить фронт Парето. Применение указанных свойств позволяет управлять эволюцией решений с целью поиска кластеров с заданными свойствами.

Мультимодальная кластеризация применена в анализе данных осложнений инфаркта миокарда [4]. Использовались данные о 1700 пациентах, имеющие 123 признака для каждого пациента. Данные признаков содержат сведения об анамнезе пациентов, анализах, примененной терапии и результатах лечения.

Парето-оптимальные решения строятся оптимальными для двух критериев: плотности

$$d(\mathbb{C}) = \frac{|R \cap (X_1 \times X_2 \times \ldots \times X_n)|}{|X_1| \times |X_2| \times \ldots \times |X_n|} \tag{5}$$

и объема кластеров

$$v(\mathbb{C}) = |X_1| \times |X_2| \times \ldots \times |X_n|. \tag{6}$$

Получаемый в результате фронт Парето содержит множество кластеров, среди которых особый интерес представляют кластеры двух типов.

1. Кластеры с большим числом пациентов и достаточно плотные интерпретируются как содержащие информацию, отражающую общие закономерности, например, стандартную терапию, применяемую для большинства пациентов.

2. Кластеры с небольшим числом пациентов, в пределе – с единственным пациентом, содержат нетипичную информацию и интерпретируются как факты.

В докладе приводятся результаты исследования данных осложнений инфаркта миокарда и результаты исследования производительности разработанного алгоритма: его вычислительная сложность, масштабируемость, временные характеристики. Обсуждаются возможности применения Парето-оптимальных решений задачи мультимодальной кластеризации в прикладных системах поддержки принятия решений.

Работа поддержана грантами РФФИ № 19-07-01178, № 19-47-710007 и № 20-07-00055.

[1] Formal Concept Analysis: Foundations and Applications. // Lecture Notes in Artificial Intelligence. Eds. Ganter B., Stumme G., Wille R. No. 3626. Berlin: Springer-Verlag, 2005. doi: 10.1007/978-3-540-31881-1

[2] Ignatov D.I., Gnatyshak D.V., Kuznetsov S.O., Mirkin B.G. Triadic Formal Concept Analysis and triclustering: searching for optimal patterns. // Mach. Learn. 2015. V. 101. P. 271–302.

[3] Bogatyrev, M., Orlov, D., Shestaka, T. On the Pareto-Optimal Solutions in the Multimodal Clustering Problem. In: Recent Trends in Analysis of Images, Social Networks and Texts. AIST 2021. // Communications in Computer and Information Science, vol 1573. Springer, Cham. Pp 179–194. 2022. https://doi.org/10.1007/978-3-031-15168-2_15

[4] Богатырев М.Ю., Шестака Т.В. Мультимодальная кластеризация в анализе данных осложнений инфаркта миокарда. // Доклады Международной конференции “Математическая биология и биоинформатика”. Под ред. В.Д. Лахно. Том 9. Пущино: ИМПБ РАН, 2022. doi: 10.17537/icmbb22.45

Pareto-optimal solutions in the multimodal clustering problem

Bogatyrev Michael[1] okkambo@mail.ru
Orlov Dmitriy[1]⋆ di-orl@mail.ru
[1]Tula, Tula State University

The report discusses the application of multi-criteria optimization in the task of multimodal clustering. Multimodal clustering is performed on *multimodal data* - data presented as multiple sets. The peculiarity of multimodal clustering is that it is performed simultaneously on all data sets. As a result, a multimodal cluster is a combination of data instances from different sets.

A well-known problem in cluster analysis is the problem of interpretation of the obtained clusters, the solution of which is associated with the intellectualization of data processing. Any clustering algorithm uses some measure of proximity defined on the set of objects to be clustered. Therefore, the "meaning" of the resulting clusters is determined by the proximity measure used: objects are included in the same cluster because they are close to each other according to the numerical proximity measure chosen, as a rule.

A special approach to multimodal clustering is used in the Formal Concepts Analysis (FCA) [1]. FCA algorithms work on tensor representations of multidimensional data – *formal contexts*, and build *lattices of concepts* or *multimodal clusters* on them.

A multidimensional formal context is a relation

$$\mathbb{K} =< K_1, K_2, \ldots, K_n, R > \tag{1}$$

on data domains $D_1, D_2, \ldots, D_n$, $K_i \subseteq D_i$. Multimodal clusters on the context (1) are constructed as

$$\mathbb{C} =< X_1, X_2, \ldots, X_n > \tag{2}$$

$X_i \subseteq K_i$, and have the following closure property:

$$\forall u = (x_1, x_2, \ldots, x_n) \in X_1, X_2, \ldots, X_n, u \in R, \tag{3}$$

at the same time $\forall j = 1, 2, \ldots, n, \forall x_j \in D_j \backslash X_j < X_1, \ldots, X_j \cup \{x_j\}, \ldots, X_m >$ does not satisfy condition (3).

Multimodal cluster is a subset in the form of combinations of elements from different sets of K_i. It is also defined as a closed m-set, since the closure property (3) ensures its "self-sufficiency": the cluster cannot be enlarged without violating the closure condition.

The modality of a cluster is the number of subsets of its constituents, $m \leqslant n$.

Cluster dimension is the number of objects in the cluster, both numeric and non–numeric. A formal concept is a multimodal cluster in which the condition is met for all its elements

$$u = (x_1, x_2, \ldots, x_k) \in X_1, X_2, \ldots, X_k, u \in R. \quad (4)$$

The formal concept is the maximum possible m-dimensional hypercube, completely filled with elements from the data sets $K_i \subseteq D_i$. The FCA introduces the concept of *multimodal cluster density* and formal concepts are interpreted as absolutely dense clusters [2].

The report is based on the content of our papers [3, 4] and contains new experimental results. The principal provisions disclosed in the report are the following.

1. Formal contexts (1) based on real data are their object-attribute representations. One of the sets $K_i \subseteq D_i$ are objects of the real world, and the other sets are attributes of objects and many others external to data objects. The interpretation of multimodal clusters (2) is performed in accordance with their two characteristics: density and dimension (volume). These characteristics contradict each other: as follows from the experiments on object-based data, dense clusters have a small volume and vice versa. Therefore, the task of multimodal clustering is formulated as a multi-criteria optimization problem.

2. Among the solutions to the multi-criteria optimization problem, Pareto-optimal solutions are preferred. The Pareto front contains a set of equivalent solutions, on which various variants of cluster interpretation are investigated.

3. To construct Pareto-optimal solutions to the problem of multimodal clustering, it is advisable to use evolutionary algorithms. The advantages of evolutionary algorithms are their inherent multiplicity of solutions and the ability to control the evolution of solutions in modern implementations of such algorithms.

In [3, 4], an evolutionary algorithm of multicriteria optimization is used, which is applied in solving multimodal clustering problems on several data sets. The algorithm belongs to the NSGA-II family of algorithms, in which the elitism property is applied in the selection operators of genetic algorithms. Our algorithm also uses *nondominable sorting of solutions*, which allows us to build a Pareto front. The use of these properties allows you to control the evolution of solutions in order to search for clusters with the specified properties.

Multimodal clustering has been applied in the analysis of data on complications of myocardial infarction [4]. Data on 1,700 patients with 123 signs for each patient were used. These signs contain information about the history of patients, tests, the therapy used and the results of treatment.

Pareto-optimal solutions are constructed optimal for two criteria: density

$$d(\mathbb{C}) = \frac{|R \cap (X_1 \times X_2 \times \ldots \times X_n)|}{|X_1| \times |X_2| \times \ldots \times |X_n|} \quad (5)$$

and clusters volume

$$v(\mathbb{C}) = |X_1| \times |X_2| \times \ldots \times |X_n|. \quad (6)$$

The resulting Pareto front contains many clusters, among which two types of clusters are of particular interest.

1. Clusters with a large number of patients and sufficiently dense are interpreted as containing information reflecting general patterns, for example, standard therapy used for most patients.

2. Clusters with a small number of patients, in the limit – with a single patient, contain atypical information and are interpreted as facts.

The report presents the results of the study of these complications of myocardial infarction and the results of the study of the performance of the developed algorithm: its computational complexity, scalability, time characteristics. The possibilities of using Pareto-optimal solutions to the problem of multimodal clustering in applied decision support systems are discussed.

The work was supported by RFBR grants No. 19-07-01178, No. 19-47-710007 and No. 20-07-00055.

[1] Formal Concept Analysis: Foundations and Applications. // Lecture Notes in Artificial Intelligence. Eds. Ganter B., Stumme G., Wille R. No. 3626. Berlin: Springer-Verlag, 2005. doi: 10.1007/978-3-540-31881-1

[2] Ignatov D.I., Gnatyshak D.V., Kuznetsov S.O., Mirkin B.G. Triadic Formal Concept Analysis and triclustering: searching for optimal patterns. // Mach. Learn. 2015. V. 101. P. 271–302.

[3] Bogatyrev, M., Orlov, D., Shestaka, T. On the Pareto-Optimal Solutions in the Multimodal Clustering Problem. In: Recent Trends in Analysis of Images, Social Networks and Texts. AIST 2021. // Communications in Computer and Information Science, vol 1573. Springer, Cham. Pp 179–194. 2022. https://doi.org/10.1007/978-3-031-15168-2_15

[4] Bogatyrev M.Y., Shestaka T.V. Multimodal clustering in the analysis of data on complications of myocardial infarction. // Reports of the International Conference “Mathematical Biology and Bioinformatics”. Edited by V.D. Lakhno. Volume 9. Pushchino: IMPB RAS, 2022. doi: 10.17537/icmbb22.45

Поиск частых элементов в данных и обучение по прецедентам

Драгунов Никита Аркадьевич[1]⋆ nikitadragunovjob@gmail.com
Дюкова Елена Всеволодовна[1] edjukova@mail.ru
[1]Москва, ФИЦ ИУ РАН

В работе рассматривается подход к задаче классификации по прецедентам, основанный на применении аппарата дискретной математики [1]. Предложена новая вычислительно эффективная модель корректного классификатора, базирующаяся на поиске частых элементов в данных [2]. Приведены результаты экспериментов на модельных и реальных задачах.

Задача классификации по прецедентам формулируется следующим образом. Пусть M — исследуемое множество объектов. Известно, что множество M представимо в виде объединения непересекающихся подмножеств $K_1, \ldots, K_l$, называемых *классами*. Объекты из M описываются некоторой системой числовых *признаков* $x_1, \ldots, x_n$. Каждый признак имеет ограниченное число допустимых значений, которые кодируются целыми числами. Имеются примеры объектов из множества M, про каждый из которых известно, какому классу он принадлежит. Это обучающие объекты или *прецеденты*. Требуется по предъявленному набору значений признаков, описывающему некоторый объект из M, о котором заранее не известно, какому классу он принадлежит, определить этот класс.

Введем основные понятия. *Элементарным классификатором* (ЭК) ранга r называется элементарная конъюнкция над переменными $x_1, \ldots, x_n$ вида $x_{j_1}^{\sigma_1} \ldots x_{j_r}^{\sigma_r}$, где $x_{j_i} \neq x_{j_u}$ $(i \neq u)$, σ_i — допустимое значение признака x_{j_i} при $i = 1, \ldots, r$ и $x_{j_i}^{\sigma_i}$ принимает значение 1 при $x_{j_i} = \sigma_i$ и значение 0 иначе. Через N_B обозначается интервал истинности ЭК B. Объект $S \in M$ *содержит* ЭК B, если $S \in N_B$. ЭК B *порождает* ЭК B', если $N_B \subset N_{B'}$.

Пусть $K \in \{K_1, \ldots, K_l\}$, $\bar{K} = \{K_1, \ldots, K_l\} \setminus \{K\}$. Положим $Q(K)$ и $Q(\bar{K})$ — множества всех прецедентов из K и из $\bar{K}$ соответственно.

ЭК B называется *p-частым* в $Q(K)$, если $|N_B \cap Q(K)| \geqslant p$. ЭК B, являющийся p-частым в $Q(K)$, называется *максимальным p-частым* в $Q(K)$, если в $Q(K)$ не существует ни одного p-частого ЭК, который порождает B.

ЭК B называется *покрытием* для $Q(\bar{K})$, если $N_B \cap Q(\bar{K}) = \varnothing$. ЭК B, являющийся покрытием для $Q(\bar{K})$, называется *минимальным покрытием* для $Q(\bar{K})$, если B не порождает никакое другое покрытие для $Q(\bar{K})$.

ЭК B называется *(тупиковым) p-представительным* для класса K, если B — p-частый в $Q(K)$ и B —(минимальное) покрытие для $Q(\bar{K})$. ЭК B, являющийся p-представительным для класса K, называется *максимальным p-представительным* для класса K, если B — максимальный p-частый в $Q(K)$.

Рассматриваются три модели классификаторов. Первая — это хорошо известная модель голосования по тупиковым p-представительным ЭК [3, 4], в которой поиск искомых ЭК для каждого класса K производится в два этапа. На

первом этапе на основе анализа множества $Q(\bar{K})$ строятся минимальные покрытия для $Q(\bar{K})$. При этом решается сложная перечислительная задача дискретной математики, называемая монотонной дуализацией [5]. На втором этапе из найденных ЭК отбираются те, которые являются p-частыми в $Q(K)$. Основная вычислительная сложность в этой модели заключается в необходимости решать задачу монотонной дуализации, для решения которой не существует эффективного алгоритма (алгоритма с полиномиальной задержкой).

Вторая модель строит множество максимальных p-представительных элементарных классификаторов для каждого класса K [6]. Поиск таких ЭК также осуществляется в два этапа. На первом этапе анализируется множество $Q(K)$. В результате формируются все максимальные p-частые ЭК. На втором этапе из найденных ЭК выбираются только те, которые являются (минимальными) покрытиями для $Q(\bar{K})$.

Вычислительная сложность второй модели в основном заключается в необходимости поиска максимальных p-частых ЭК. Для решения этой задачи не существует алгоритма с полиномиальной задержкой. Однако по сравнению с первой моделью, эта модель является более эффективной на практике, особенно в случае большого числа классов, так как вместо $Q(\bar{K})$ алгоритм рассматривает меньшее по мощности множество $Q(K)$.

Третья модель также работает в два этапа. На первом этапе строится множество p-частых в $Q(K)$ элементарных классификаторов ранга p. ЭК такого вида называются *p-правильными* в $Q(K)$. На втором этапе из найденных ЭК выбираются только те, которые являются (минимальными) покрытиями для $Q(\bar{K})$. Поиск p-правильных ЭК на первом этапе осуществляется алгоритмом ADR, предложенным в настоящей работе. Алгоритм ADR является модификацией алгоритма поиска частых элементов в бинарных данных DepthProject [2].

На вход алгоритму ADR подается матрица L, строками которой являются описания объектов класса K, бинаризованные с помощью известного метода one-hot кодирования. Несложно видеть, что поиск всех p-правильных элементарных классификаторов эквивалентен поиску всех наборов из p столбцов матрицы L, которые в пересечении не менее чем с p строками этой матрицы образуют подстроку, состоящую из единичных элементов. Такой набор столбцов называется p-правильным. Работу алгоритма ADR можно представить в виде обхода в глубину дерева решений, вершинами которого являются наборы столбцов матрицы L, причем на глубине d находятся все наборы мощности d и только они. На выходе алгоритм ADR возвращает все p-правильные наборы столбцов матрицы L в порядке их обхода

В работе проведено экспериментальное сравнение трех моделей поиска корректных элементарных классификаторов. Рассматриваемые модели реализованы на языке C++. Первая модель строит тупиковые p-предстательные ЭК, при этом задача монотонной дуализации решается с помощью асимптотически оптимального алгоритма RUNC-M [7], который является лидером по скорости

счёта среди других известных алгоритмов монотонной дуализации. Вторая модель строит максимальные p-представительные ЭК с применением алгоритма DepthProject [2]. Третья модель строит p-правильные ЭК с применением предлагаемого в настоящей работе алгоритма ADR. Экспериментально исследованы количественные свойства множеств искомых ЭК и выявлены условия применимости новой модели поиска p-правильных ЭК.

[1] *Журавлёв Ю. И.* Об алгебраическом подходе к решению задач распознавания и классификации // Проблемы кибернетики, Москва: Наука, 1978. Вып. 33. — С. 5–68.

[2] *Aggarwal C.* Frequent Pattern Mining // Springer International Publishing, 2014. — 471 p.

[3] *Баскакова Л. В., Журавлев Ю. И.* Модель распознающих алгоритмов с представительными наборами и системами опорных множеств // Ж. вычисл. матем. и матем. физ., 1981. Т. 21. No.5 — С. 1264–1275.

[4] *Дюкова Е. В.* Алгоритмы распознавания типа «Кора»: сложность реализации и метрические свойства // Распознавание, классификация, прогноз (матем. методы и их применение), Москва: Наука, 1989. Вып. 2. — С. 99–125.

[5] *Дюкова Е. В., Журавлев Ю. И.* Задача монотонной дуализации и её обобщения: асимптотические оценки числа решений // Ж. вычисл. матем. и матем. физ., 2018. Т. 58. No.12. — С. 5–25.

[6] *Dragunov N., Djukova E. and Djukova A.* Supervised Classification and Finding Frequent Elements in Data // VIII International Conference on Information Technology and Nanotechnology (ITNT), 2022. — p. 1–5.

[7] *Дюкова Е. В., Прокофьев П. А.* Об асимптотически оптимальных алгоритмах дуализации // Ж. вычисл. матем. и матем. физ., 2015. Т. 16. No.1. — С. 895–910.

Finding frequent elements in data and supervised learning

Dragunov Nikita[1]⋆ nikitadragunovjob@gmail.com
Djukova Elena[1] edjukova@mail.ru
[1]Moscow, FRC CSC RAS

The paper considers an approach to the problem of supervised classification based on the application of the apparatus of discrete mathematics [1]. A new computationally efficient model of correct classifier based on finding frequent elements in data is proposed [2]. The results of experiments on model and real tasks are presented.

The task of supervised classification is formulated as follows. Let M be a nonempty finite set of objects. The set M is known to be represented as a union of disjoint subsets $K_1, \ldots, K_l$, called *classes*. Objects from M are described by some system of integer *features* $x_1, \ldots, x_n$. There are examples of objects from the set of M, about each of which it is known which class it belongs to. These are training objects or *precedents*. It is required to determine the class based on the presented set of feature values describing some object from M, about which it is not known in advance to which class it belongs.

Let's introduce the basic concepts. An *elementary classifier* (EC) of rank r is an elementary conjunction over variables $x_1, \ldots, x_n$ of the form $B = x_{j_1}^{\sigma_1} \ldots x_{j_r}^{\sigma_r}$, where $x_{j_i} \neq x_{j_u}$ $(i \neq u)$, σ_i is the allowed value of the attribute x_{j_i} when $i = 1, \ldots, r$ and $x_{j_i}^{\sigma_i}$ takes the value 1 when $x_{j_i} = \sigma_i$ and the value 0 otherwise. N_B denotes the interval of truth of EC B. The object $S \in M$ *contains* EC B if $S \in_B$. EC B *generates* EC B' if $N_B \subset N_{B'}$.

Let $K \in \{K_1, \ldots, K_l\}$, $\bar{K} = \{K_1, \ldots, K_l\} \setminus \{K\}$. Assume $Q(K)$ and $Q(\bar{K})$ are the sets of all precedents from K and from $\bar{K}$, respectively.

EC B is called *p-frequent* in $Q(K)$ if $|N_B \cap Q(K)| \geqslant p$. EC B that is p-frequent in $Q(K)$ is called *maximal p-frequent* in $Q(K)$ if there are no p-frequent ECs in $Q(K)$ that generate B.

EC B is called a *covering* for $Q(\bar{K})$ if $N_B \cap Q(\bar{K}) = \varnothing$. An EC B that is a covering for $Q(\bar{K})$ is called a *minimal covering* for $Q(\bar{K})$ if B does not generate any other covering for $Q(\bar{K})$.

EC B is called *(irredundant) p-representative* for the class K if B is p-frequent in $Q(K)$ and B is a (minimal) covering for $Q(\bar{K})$. An EC B that is p-representative for the class K is called *maximal p-representative* for the class K if B is maximal p-frequent in $Q(K)$.

Three classification models are considered. The first one is a well-known model of voting by irredundant p-representative ECs [3, 4], in which finding of target ECs for each class K is performed in two stages. At the first stage, the minimal coverings for $Q(\bar{K})$ are constructed based on the analysis of $Q(\bar{K})$. Herewith a complex enumerative problem of discrete mathematics, called monotone dualization, is solved [5]. At the second stage, ECs that are p-frequent in $Q(K)$ are selected from the found ECs. The main computational complexity in this model lies in the need to solve

the monotone dualization problem, for which there is no effective algorithm (an algorithm with a polynomial delay).

The second model builds a set of maximal p-representative elementary classifiers for each class K [6]. Finding of such ECs is also performed in two stages. At the first stage, the set $Q(K)$ is analyzed. As a result, the set of all maximal p-frequent ECs is formed. At the second stage, only those that are (minimal) coverings for $Q(\bar{K})$ are selected from the found ECs.

The computational complexity of the second model mainly lies in the need to find maximal p-frequent ECs. There is no polynomial-delay algorithm for solving this problem. However, compared to the first model, this model is more efficient in practice, especially in the case of a large number of classes, since instead of $Q(\bar{K})$, the algorithm considers a smaller set of $Q(K)$.

The third model works in two stages. At the first stage the set of p-frequent in $Q(K)$ elementary classifiers of rank p is constructed. ECs of this kind are called *p-regular* in $Q(K)$. At the second stage, only those that are (minimal) coverings for $Q(\bar{K})$ are selected from the found ECs. Finding p-regular ECs at the first stage is carried out by the ADR algorithm proposed in this paper. The ADR algorithm is a modification of the algorithm of finding frequent elements in binary data DepthProject [2].

The input of ADR algorithm is the matrix L whose rows are descriptions of objects of class K, binarized using the well-known one-hot encoding method. It is easy to see that finding of all p-regular elementary classifiers is equivalent to finding of all sets of p columns of the matrix L, which in intersection with at least p rows of this matrix form a substring consisting of unit elements. Such a set of columns is called p-regular. The ADR algorithm can be described as a traversal into the depth of the decision tree, the vertices of which are the sets of columns of the matrix L, and at a depth of d are all the sets of power d and only them. At the output, the ADR algorithm returns all p-regular sets of columns of the matrix L in the order of their traversal.

The paper presents an experimental comparison of three models of correct elementary classifiers finding. The models are implemented in C++. The first model builds irredundant p-representative ECs by solving the monotone dualization problem using the asymptotically optimal RUNC-M [7] algorithm, which is the leader in counting speed among other well-known monotone dualization algorithms. The second model builds maximal p-representative ECs using the DepthProject [2] algorithm. The third model is proposed in this paper and builds p-regular ECs using the ADR algorithm described above. The paper experimentally investigates the quantitative properties of the sets of the desired ECs and identifies the conditions for the applicability of a new model of p-regular ECs finding.

[1] *Zhuravlev Yu. I.* Ob algebraicheskom podhode k resheniyu zadach raspoznavaniya i klassifikacii // Problems of Cybernetics, Moscow: Nauka, 1978. No.33 — p. 5–68.

[2] *Aggarwal C.* Frequent Pattern Mining // Springer International Publishing, 2014. — 471 p.

[3] *Baskakova L. V., Zhuravlev Yu. I.* Model' raspoznayushchih algoritmov s predstavitel'nymi naborami i sistemami opornyh mnozhestv // Zh. vychisl. matem. i matem. fiz., 1981. V. 21. No.5 — p. 1264–1275.

[4] *Djukova E. V.* Algoritmy raspoznavaniya tipa "Kora": slozhnost' realizacii i metricheskie svojstva // Raspoznavanie, klassifikaciya, prognoz (matem. metody i ih primenenie), Moscow: Nauka, 1989. No.2. — p. 99–125.

[5] *Djukova E. V., Zhuravlev Yu. I.* Zadacha monotonnoj dualizacii i eyo obobshcheniya: asimptoticheskie ocenki chisla reshenij // Zh. vychisl. matem. i matem. fiz., 2018. V. 58. No.12. — p. 5–25.

[6] *Dragunov N., Djukova E. and Djukova A.* Supervised Classification and Finding Frequent Elements in Data // VIII International Conference on Information Technology and Nanotechnology (ITNT), 2022. — p. 1–5.

[7] *Djukova E. V., Prokof'ev P. A.* Ob asimptoticheski optimal'nyh algoritmah dualizacii // Zh. vychisl. matem. i matem. fiz., 2015. V. 16. No.1. — p. 895–910.

Принципы построения и функционирования многоуровневых моделей распознавания

Краснопрошин Виктор Владимирович[1,2] krasnoproshin@bsu.by
Образцов Владимир Алексеевич[2]⋆ obraztsov@bsu.by
[1]Минск, Белорусский государственный университет
[2]Минск, Белорусский государственный университет

Задача распознавания образов является индуктивной по построению, так как заданная в ней информация априори неполная. Поэтому все алгоритмы, используемые для ее решения, по своей сути являются эвристическими, то есть не имеют строгого математического обоснования.

Большинство индуктивных задач решается по следующей схеме. С помощью эвристических алгоритмов получают одно из возможных решений, которое затем пытаются улучшить с помощью допустимых математических средств. Такая схема позволяет не только упростить требования к эвристическим алгоритмам, но и повысить качество финального решения. Построенные в результате алгоритмы назвали многоуровневыми. На рис. 1 приведена схема двухуровневых алгоритмов.

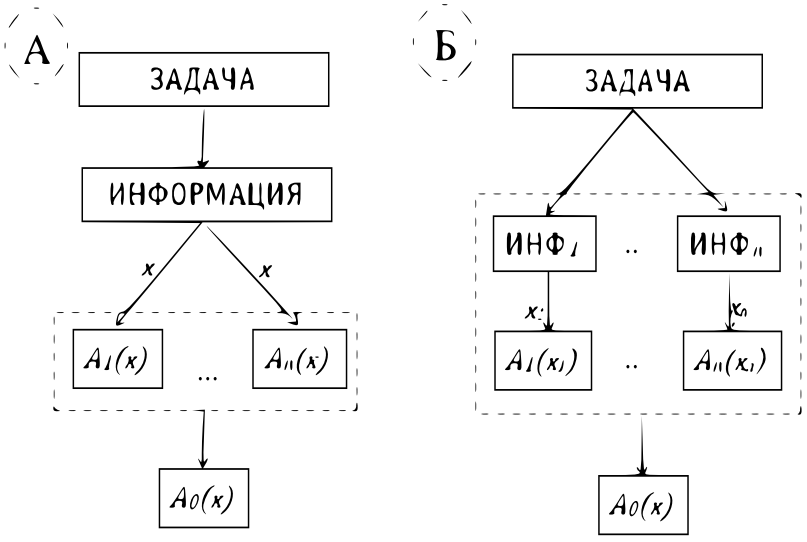


Рис. 1. Общая схема построения двухуровневых моделей распознавания

Принципы построения моделей для приведенных на схеме вариантов А и Б различны. В первом случае строится суперпозиция эвристических алгоритмов $A_1, ..., A_n$ и некоторого алгоритма A_0 . Все эти алгоритмы работают с одним и тем же входным объектом x. Во втором – информация о задаче декомпози-

руется. Каждый из алгоритмов A_i работает со своим объектом x_i. При этом алгоритм A_0 применяется для синтеза результата на объекте x , который является суперпозицией объектов $x_1, ..., x_n$.

Активное использование первого их них было инициировано Ю.И.Журавлевым [1]. Смысл его предложения был достаточно прост. Так как эвристические алгоритмы являются плохо управляемыми, то получив с их помощью какое-либо решение, его можно затем корректировать. То есть управление решением в таком случае передавалось на последующие шаги. При этом требования к эвристическим алгоритмам существенно понижались.

Также было замечено [1], что любой алгоритм распознавания можно рассматривать в виде суперпозиции распознающего оператора и решающего правила. Исследова-ния таких схем, в зависимости от результатов работы алгоритмов $A_1, ..., A_n$, развивались в двух направлениях. В первом случае предполагалось, что финальный результат строится алгоритмами из указанного набора и формируется в пространствах $B_2 = \{0, 1\}$ либо $B_3 = \{0, 1, 2\}$. Такой подход назвали логической корректировкой [2]. Во втором случае в качестве $A_1, ..., A_n$ использовались распознающие операторы. Алгоритмы, в общем случае, формировали результат на базе пространства действительных чисел R . Для построения алгоритма A_0 в этом случае использовались обычные алгебраические операции. Такое подход, в свою очередь, назвали алгебраической корректировкой.

В результате проведенных исследований выяснилось, что возможности логических корректоров, в силу "бедности" пространства финальных решений, сильно ограничены [3]. В свою очередь исследования алгебраической корректировки были направлены на упрощение требований к эвристическим алгоритмам. Предельный (в некотором смысле) для них результат [4] можно сформулировать следующим образом. Если объекты контрольной выборки не совпадают, а эвристический алгоритм не слишком "вычурный", то в билинейном замыкании единственного распознающего оператора всегда можно построить корректный (точный для заданной выборки) алгоритм. Следствием этого является утверждение: если корректный алгоритм невозможно построить в двухуровневой модели, то его невозможно построить для рассматриваемой задачи в принципе.

Алгебраическая, как и логическая корректировка не лишена недостатков:

- корректные алгоритмы строились только для контрольных выборок, что обеспечивало необходимые условия разрешимости задачи на всем множестве допустимых объектов;

- за пределами контрольной выборки алгоритм A_0 вел себя не совсем адекватно.

Перейдем теперь к обсуждению варианта Б. С необходимостью построения таких алгоритмов авторы столкнулись в задачах медицинской диагностики [5]. При решении практических задач часть информации, как правило, берется из медицинской литературы (в виде логических правил), а часть - из личного опыта врача. Причем последняя часть информации в большинстве случаев не фор-

мализована. Часто она представляться в виде примеров (прецедентов), описанных в своем признаковом пространстве. Для постановки диагноза с использованием правил можно применять стандартный алгоритм резолюций A_1 . А для диагностики по примерам, естественно, использовать алгоритм распознавания A_2 . Однако для получения общего результата $A_0(x)$ возникает проблема, связанная с объединением решений $A_1(x)$ и $A_2(x)$. Для ее решения разработан оригинальный подход, основанный на модификации двухуровневой модели варианта Б.

Общая схема такого подхода включает следующие основные этапы.

- Строится декартово произведение пространств, в которых определены объекты x_1 и x_2 .

- Алгоритмами A_1 (с использованием процедуры построения КНФ) в указанном про-странстве генерируется подмножество объектов.

- На основе примеров в декартовом произведении строится проекция объектов, а все объекты приводятся к одной размерности.

- В пространстве $X \subseteq B_2^n$ вводится мера прецедентности $s : X \times X \to [-1, 1]$ такая, что

$$\forall x_1, x_2 \in X \begin{cases} s(x_1, x_2) = 1 \Leftrightarrow x_1 = x_2; \\ s(x_1, x_2) = s(x_2, x_1); \\ s(x_1, x_2) + s(x_1, \overline{x_2}) = 0. \end{cases}$$

где $\overline{x_2}$ - логическое отрицание объекта x_2.

- На базе меры прецедентности строится семейство алгоритмов A_0 , в виде расширения алгоритмов A_2 . При этом семейство должно удовлетворять условию: на данных, генерируемых из логической части информации, результаты алгоритмов A_0 и A_1 должны совпадать. Аналогичное условие накладывается на выборки и алгоритмы A_0 и A_2.

Предложенный подход успешно применялся на практике для решения прикладных задач из области ортопедии и спортивной медицины.

Работа поддержана грантом БРФФИ Ф21АРМ-005.

[1] *Журавлев В.* Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики, Москва, 1978. — С. 5–68.

[2] *Zhuravlev Yu.* Algorithms for Algebraic and Logical Correction and Their Applications // Pattern Recognition and Image Analysis, 2010. — p. 155–169.

[3] *Краснопрошин В.* Об оптимальном корректоре совокупности алгоритмов распознавания // Журнал вычислительной математики и математической физики, 1979. — С. 204–215.

[4] *Krasnoproshin V.* The Problem of Algorithms Choosing in Pattern Recognition // Pattern Recognition and Image Analysis, 1996. — p. 188–199.

[5] *Krasnoproshin V.* Decision-Making in Sports Traumatology // Sports Management as an Emerging Economic Activity Trends and Best Practices, 2017. — p. 207–219.

Construction and operation principles of multilevel recognition models

Krasnoproshin Victor[1,2] krasnoproshin@bsu.by
Obraztsov Vladimir[2]⋆ obraztsov@bsu.by
[1]Belarus, Minsk, Belarusian State University
[2]Belarus, Minsk, Belarusian State University

The problem of pattern recognition is inductive by construction, since the information given in it is a priori incomplete. Therefore, all the algorithms used to solve it are inherently heuristic which means that they do not have a strict mathematical justification.

Most inductive problems are solved according to the following scheme. Using heuristic algorithms, one of the possible solutions is obtained, which they then try to improve with the help of valid mathematical means. Such a scheme allows not only simplifying the requirements for heuristic algorithms, but also improving the quality of the final solution. The resulting algorithms are called multilevel algorithms. Fig. 1 shows a diagram of two-level algorithms.

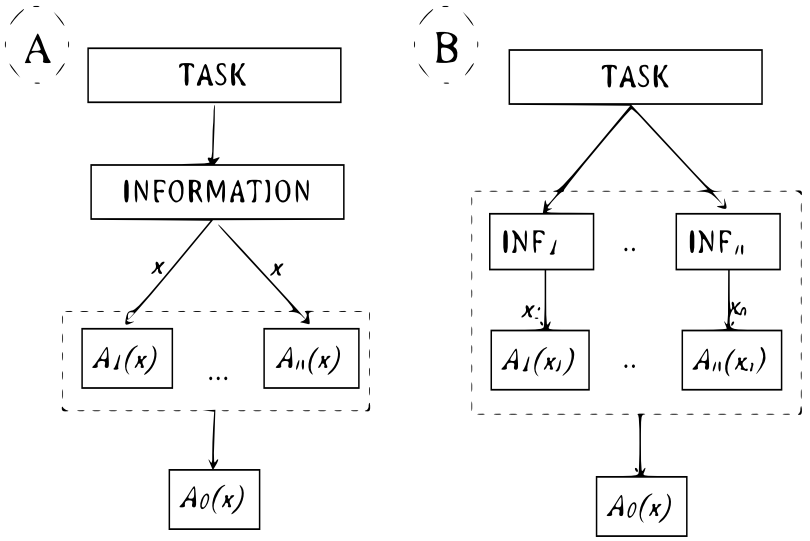


Fig. 1. General scheme for constructing two-level recognition models

The principles of building models for options A and B shown in the diagram are different. In the first case, a superposition $A_1, ..., A_n$ of heuristic algorithms is constructed as well as algorithm A_0. All these algorithms work with the same input object x. In the second case, information about the task is decomposed. Each of the

algorithms A_i works with its own object x_i. In this case, the algorithm A_0 is used to synthesize the result on object x, which is a superposition of objects $x_1, ..., x_n$.

The active use of the first algorithm was initiated by Yuriy Zhuravlev [1]. His idea was quite simple. Since heuristic algorithms are poorly controlled we can correct any solution being obtained by such algorithms later on. This means that decision control is transferred to subsequent steps. At the same time, the requirements for heuristic algorithms were significantly reduced.

It was also noted [1] that any recognition algorithm can be considered as a superposition of a recognition operator and a decision rule. Studies of such schemes, depending on the results of the algorithms $A_1, ..., A_n$, developed in two directions. In the first case, it was assumed that the final result is built by algorithms from the specified set and is formed in the spaces $B_2 = \{0, 1\}$ or $B_3 = \{0, 1, 2\}$. This approach was called logical correction [2]. In the second case, recognition operators were used as $A_1, ..., A_n$. Algorithms, in the general case, formed the result based on the space of real numbers R . In this case, ordinary algebraic operations were used to construct the algorithm A_0. This approach was called algebraic correction.

As a result of the research, it turned out that the possibilities of logical corrections, due to the "poverty" of the final solutions space are severely limited [3]. In turn, studies of algebraic correction were aimed at simplifying the requirements for heuristic algorithms. The limit (in some sense) result [4] for them can be formulated as follows. If the objects of the control sample do not coincide, and the heuristic algorithm is not too "pretentious", then in the bilinear closure of the only recognizing operator, it is always possible to construct a correct (exact for a given sample) algorithm. The consequence of this is the statement: if a correct algorithm cannot be constructed using a two-level model, then it cannot be constructed for the problem being under consideration at all.

Algebraic, as well as logical correction has its drawbacks:

- correct algorithms were built only for control samples, which provided the necessary conditions for the solvability of the problem on the entire set of admissible objects;

- outside the control sample, the algorithm did not behave quite adequately.

Let us now proceed to the discussion of the option B. The authors encountered the need to construct such algorithms to solve medical diagnostics problems[5]. When solving practical problems, part of the information, as a rule, is taken from the medical literature (in the form of logical rules), and part is taken from the doctor's personal experience. Moreover, the last part of the information in most cases is not formalized. Often it is presented in the form of examples (precedents) described in its feature space. To make a diagnosis using the rules, you can use the standard resolution algorithm A_1. And for diagnostics by examples, of course, use the recognition algorithm A_2. However, to obtain a general result $A_0(x)$ there is a problem associated with combining solutions found with $A_1(x)$ and $A_2(x)$. To solve it, an

original approach has been developed based on a modification of the two-level model of option B.

The general scheme of this approach includes the following main stages.

- Construction of the Cartesian product of the spaces in which the objects x_1 and x_2 are defined.

- Algorithms A_1 (using the CNF construction procedure) generate a subset of objects in the specified space.

- Based on the examples in the Cartesian product, a projection of objects is constructed, and all objects are reduced to the same dimension.

- In space $X \subseteq B_2^n$ a measure of precedence is introduced $s : X \times X \to [-1, 1]$ such as

$$\forall x_1, x_2 \in X \begin{cases} s(x_1, x_2) = 1 \Leftrightarrow x_1 = x_2; \\ s(x_1, x_2) = s(x_2, x_1); \\ s(x_1, x_2) + s(x_1, \overline{x_2}) = 0. \end{cases}$$

where $\overline{x_2}$ - logical negation of an object x_2.

- Based on the measure of precedence, a family of algorithms A_0 , is built in the form of an extension of algorithms A_2 . In this case, the family must satisfy the condition: on the data generated from the logical part of the information, the results of the algorithms A_0 and A_1 must match. A similar condition is imposed on samples and algorithms A_0 and A_2.

The proposed approach was successfully used in practice to solve applied problems in the field of orthopedics and sports traumatology.

This work was supported by the BRFFR grant F21ARM-005.

[1] *Zhuravlev Yu.* On an algebraic approach to solving problems of recognition or classification // Problems of Cybernetics, Moscow, 1978. — p. 5–68.

[2] *Zhuravlev Yu.* Algorithms for Algebraic and Logical Correction and Their Applications // Pattern Recognition and Image Analysis, 2010. — p. 155–169.

[3] *Krasnoproshin V.* On the optimal corrector of a set of recognition algorithms // Journal of Computational Mathematics and Mathematical Physics, 1979. — p. 204–215.

[4] *Krasnoproshin V.* The Problem of Algorithms Choosing in Pattern Recognition // Pattern Recognition and Image Analysis, 1996. — p. 188–199.

[5] *Krasnoproshin V.* Decision-Making in Sports Traumatology // Sports Management as an Emerging Economic Activity Trends and Best Practices, 2017. — p. 207–219.

Об одном робастном методе главных компонент

Шибзухов Заур Мухадинович[1,2] intellimath@mail.ru
[1]Москва, Институт математики и информатики МПГУ
[2]Москва, Московский физико-технический институт

Классический метод главных компонент первоначально рассматривался как задача наилучшей аппроксимации конечного множества точек прямыми и плоскостями [1]. Рассмотрим один вариант робастной постановки этой задачи. Он основан на применении дифференцируемых агрегирующих функций $\mathsf{M}\{z_1,\dots,z_N\}$ [3, 2], которые являются нечувствительными к выбросам.

Дано конечное множество точек $\{x_1,\dots,x_N\}\subset\mathbb{R}^n$. Центр $a_0\in\mathbb{R}^n$ ищется как решение задачи:

$$a_0=\arg\min_{a\in\mathbb{R}^n}\mathsf{M}\big\{\|x_1-a\|^2,\dots,\|x_N-a\|^2\big\}.$$

Для ее решения используется следующая итерационная процедура:

$$a^{t+1}=\sum_{k=1}^{N}v_k^t x_k,$$

где $v_k^t=\dfrac{\partial\mathsf{M}\{\|x_1-a^t\|^2,\dots,\|x_N-a^t\|^2\}}{\partial z_k}$.

После нахождения a_0 осуществляется центрирование:

$$x_k\to x_k-a_0,\quad k=1,\dots,N.$$

Очередная главная компонента a_j $(1\leqslant j<n)$ ищется как решения задачи:

$$a_j=\arg\min_{\|a\|=1}\mathsf{M}\left\{\|x_1\|^2-(a,x_1)^2,\dots,\|x_N\|^2-(a,x_N)^2\right\}.$$

Для ее решения используется следующая итерационная процедура:

$$a_j^{t+1}=\arg\min_{\|a\|=1}\sum_{k=1}^{N}v_k^t\left(\|x_k\|^2-(a,x_k)^2\right)$$

где $v_k^t=\dfrac{\partial\mathsf{M}\left\{\|x_1\|^2-(a^t,x_1)^2,\dots,\|x_N\|^2-(a^t,x_N)^2\right\}}{\partial z_k}$.

Ее можно также привести к следующей форме:

$$a^{t+1}=\frac{1}{\lambda^t}\left(S^t a^t\right),$$

где

$$\lambda^t = \frac{(a^t)^\top S^t a^t}{(a^t, a^t)}, \quad S^t = X^\top \begin{pmatrix} v_1^t & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & v_N^t \end{pmatrix} X, \quad X = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{Nn} \end{pmatrix}.$$

На наглядных примерах показывается устойчивость предложенных методов к выбросам.

[1] Pearson K. On lines and planes of closest fit to systems of points in space. Philosophical Magazine. 1901. Vol. 2. PP. 559–572.

[2] Шибзухов З., М. Принцип минимизации эмпирического риска и усредняющие агрегирующие функции. – В: Итоги науки и техники. Серия: Современная математика и ее приложения. Тематические обзоры. ВИНИТИ РАН. 2018, Т.152.

[3] Shibzukhov, Z. Machine Learning Based on the Principle of Minimizing Robust Mean Estimates Advances in Intelligent Systems and Computing, Springer International Publishing. 2020. V.1310, PP.472–477.

About one robust principal component method

Shibzukhov Zaur[1,2] intellimath@mail.ru

[1]Moscow, Institute of mathematoca and informatics MPSU

[2]Moscow, Moscow Institute of Physics Technologies

The classical principal component method was initially considered as the problem of the best approximation of a finite set of points by straight lines and planes [1]. Let's consider one variant of the robust formulation of this problem. It is based on the application of differentiable aggregating functions $\mathsf{M}\{z_1,\ldots,z_N\}$ [2, 3], which are insensitive to outliers.

Given a finite set of points $\{x_1,\ldots,x_N\}\subset\mathbb{R}^n$. The center $a_0\in\mathbb{R}^n$ is sought as a solution to the problem:

$$a_0 = \arg\min_{a\in\mathbb{R}^n} \mathsf{M}\big\{\|x_1-a\|^2,\ldots,\|x_N-a\|^2\big\}.$$

To solve it , the following iterative procedure is used:

$$a^{t+1} = \sum_{k=1}^{N} v_k^t x_k,$$

where $v_k^t = \dfrac{\partial \mathsf{M}\{\|x_1-a^t\|^2,\ldots,\|x_N-a^t\|^2\}}{\partial z_k}$.

After finding a_0, centering is performed:

$$x_k \to x_k - a_0, \quad k=1,\ldots,N.$$

The next main component a_j $(1\leqslant j<n)$ is sought as a solution to the problem:

$$a_j = \arg\min_{\|a\|=1} \mathsf{M}\left\{\|x_1\|^2-(a,x_1)^2,\ldots,\|x_N\|^2-(a,x_N)^2\right\}.$$

To solve it , the following iterative procedure is used:

$$a_j^{t+1} = \arg\min_{\|a\|=1} \sum_{k=1}^{N} v_k^t\left(\|x_k\|^2-(a,x_k)^2\right)$$

where $v_k^t = \dfrac{\partial \mathsf{M}\left\{\|x_1\|^2-(a^t,x_1)^2,\ldots,\|x_N\|^2-(a^t,x_N)^2\right\}}{\partial z_k}$.

It can also be reduced to the following form:

$$a^{t+1} = \frac{1}{\lambda^t}\left(S^t a^t\right),$$

where

$$\lambda^t = \frac{(a^t)^\top S^t a^t}{(a^t, a^t)}, \quad S^t = X^\top \begin{pmatrix} v_1^t & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & v_N^t \end{pmatrix} X, \quad X = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{Nn} \end{pmatrix}.$$

The stability of the proposed methods to outliers is shown by illustrative examples.

[1] Pearson K. On lines and planes of closest fit to systems of points in space. Philosophical Magazine. 1901. Vol. 2. PP. 559–572.

[2] Shibzukhov Z.M. The empirical risk minimization principle and averaging aggregating functions. – In: Results of science and technology. Sries: Modern Mathematics and its applications. Thematic reviews. VINITI RAS. 2018, Vol.152.

[3] Shibzukhov Z. Machine Learning Based on the Principle of Minimizing Robust Mean Estimates Advances in Intelligent Systems and Computing, Springer International Publishing. 2020. V.1310, PP.472–477.

Сравнение средств AutoML на примере задачи регресии

Попова Инна Андреевна[1]⋆ popovai1@student.bmstu.ru
Гапанюк Юрий Евгеньевич[1] gapyu@bmstu.ru
[1]Москва, МГТУ им. Н.Э.Баумана

Автоматизированное машинное обучение (AutoML) можно назвать современной тенденцией в сфере машинного обучения. Данное направление активно исследуется научным сообществом, что подтверждает существование и разработку различных программных систем автоматизированного машинного обучения. В качестве примера можно привести достаточно распространенные системы AutoML: Light AutoML (LAMA), TPOT, Auto-Sklearn, H2O AutoML, Mljar. Автоматизированное машинное обучение направлено на автоматическую настройку алгоритмов машинного обучения и их компоновку в общее (программное) решение – конвейер машинного обучения – с учетом поставленной задачи обучения (набора данных) [1].

Данные разработки позволяют специалистам в области машинного обучения подобрать подходящий алгоритм с оптимальными гиперпараметрами для описания исследуемого набора данных. Зачастую исследователи выполняют и оценивают множество конфигураций модели методом проб и ошибок. Последние достижения в области исследований AutoML решают эту проблему путем автоматического поиска подходящего алгоритма с соответствующими гиперпараметрами, то есть автоматизируют практически все этапы разработки модели машинного обучения [2]. AutoML пытается создать единую систему, которая может устранить необходимость вмешательства человека на каждом этапе процесса моделирования и обучения.

Прогресс в области AutoML привел к появлению множества систем, которые автоматизируют проектирование и разработку преимущественно моделей машинного обучения с учителем на разных этапах.

Мы исследовали ряд современных разработок в области AutoML: Light AutoML (LAMA), Tree-Based Pipeline Optimization Tool (TPOT), Auto-Sklearn, H2O AutoML, Mljar, решая при этом задачу машинного обучения с учителем. В исследовании результаты предсказания модели линейной регрессии сравнивается с результатами моделей, которые предложили перечисленные системы. В экспериментальной части работы мы использовали два набора данных, содержащих пропуски, категориальные и числовые признаки: cars [3] и plants [4].

По результатам эксперимента было определено, что модель линейной регрессии показала наилучшие результаты по метрикам точности MAE, $RMSE$, $MedAE$, однако метрика R^2 наиболее точный результат получался у модели, построенной системой Light AutoML.

Также результаты измерения процесса построения и обучения модели для рассмотренных систем показали, что среднее время, необходимое для подбора наилучшей конфигурации модели системой AutoML составляет около одной

минуты для небольших (порядка 10 тыс. образцов) наборов данных. Все же процесс подготовки данных и обучения на них модели может сильно варьироваться в зависимости от предметной области и решаемой задачи.

В процессе анализа систем было обнаружено, что не все системы полностью автоматизируют все этапы машинного обучения: система TPOT не может работать с данными, содержащими пропуски, а также категориальные признаки должны быть заранее перекодированы в числовые. Многие системы используют на этапе предварительной обработки данных только статистические методы для очистки пропущенных значений (Mljar исключает строки с пропусками либо заполняет их средним значением, модой, минимальным значением; Auto-Sklearn заполняет пропуск в числовых данных средним значением, медианой, самым частым значением), но не удаляют аномальные значения и не снижают размерность признакового пространства. Методы машинного обучения без учителя можно также использовать для предварительной обработки данных, однако для этого потребуется значительно усложнить имплементацию модулей системы, отвечающих за предварительную обработку данных.

Таким образом, можно сделать вывод, что рассмотренные системы AutoML сложно назвать универсальными. Исследователи в области данных в итоге должны принимать выбор относительно того, какую систему использовать либо прибегнуть к самостоятельному анализу, обработке данных и подбору наиболее эффективной конфигурации модели машинного обучения.

Можно предложить ряд перспективных направлений развития для систем AutoML: внедрение алгоритмов машинного обучения без учителя для подготовки данных; явный выбор алгоритма оптимизации гиперпараметров; необходимость более интеллектуальной предварительной обработки данных, ввиду того, что для различных наборов данных унифицированный метод обработки может не подойти.

Подводя итог, следует отметить, что системы AutoML являются достаточно важным направлением для бизнеса (в частности, Light AutoML была разработана для нужд финансовой компании) [5]. Исследования в данном направлении помогут разработать общую методику построения систем автоматизированного машинного обучения, которая помогла спроектировать систему для нужд конкретной отрасли либо предложить более общее решение.

[1] *Karmaker, S., Hassan, M.M., Smith, M.J., Xu, L., Zhai, C., Veeramachaneni, K.* AutoML to Date and Beyond: Challenges and Opportunities // ACM Computing Surveys (CSUR) 54, 2022.— P.1–36.

[2] *Koroteev, M.V.* Review of some modern trends in machine learning technology // E-Management 1(1), 2018.— P.26–35.

[3] Car Dekho Data, `https://www.kaggle.com/datasets/shindenikhil/car-dekho-data`. Last accessed 30 October 2022.

[4] Combined Cycle Power Plant Dataset, `https://archive.ics.uci.edu/ml/datasets/Combined+Cycle+Power+Plant`. Last accessed 30 October 2022.

[5] *Vakhrushev A. et al.* LightAutoML: AutoML Solution for a Large Financial Services Ecosystem // CoRR, 2021.

The Comparison of AutoML Tools in Relation to the Regression Problem

Popova Inna[1]⋆ popovai1@student.bmstu.ru
Gapanyuk Yuriy[1] gapyu@bmstu.ru
[1] Moscow, Bauman Moscow State Technical University

Automated machine learning (AutoML) can be called the current trend in machine learning. This direction is being actively explored by the scientific community, which confirms the existence and development of various software systems for automated machine learning. Quite common AutoML systems can be cited as an example: Light AutoML (LAMA), TPOT, Auto-Sklearn, H2O AutoML, Mljar. Automated machine learning is aimed at automatically setting up machine learning algorithms and linking them into a common (software) solution – a machine learning pipeline – taking into account the set learning task (data set) [1].

These developments allow specialists in the field of machine learning to choose the appropriate algorithm with optimal hyperparameters to describe the data set under study. Often, researchers perform and evaluate multiple model configurations through trial and error. Recent advances in AutoML research solve this problem by automatically searching for a suitable algorithm with appropriate hyperparameters, i.e., automating almost all stages of machine learning model development [2]. AutoML is trying to create a single system that can eliminate the need for human intervention at every step of the modeling and training process.

Advances in AutoML have resulted in many systems that automate the design and development of predominantly supervised machine learning models at various stages. We explored a number of modern AutoML developments: Light AutoML (LAMA), Tree-Based Pipeline Optimization Tool (TPOT), Auto-Sklearn, H2O AutoML, Mljar while solving a supervised machine learning problem. In the study, the prediction results of the linear regression model are compared with the results of the models proposed by the listed systems. In the experimental part of the work, we used two data sets containing gaps, categorical and numerical features: cars [3] and plants [4].

According to the results of the experiment, it was determined that the linear regression model showed the best results in terms of MAE, $RMSE$, $MedAE$ accuracy metrics, however, the R^2 metric was the most accurate result for the model built by the Light AutoML system.

Also, the timings of the process of building and training the model for the considered systems showed that the average time required to select the best model configuration by the AutoML system is about one minute for small (about 10 thousand samples) data sets. Nevertheless, the process of preparing data and training a model on them can vary greatly depending on the subject area and the problem being solved.

During the analysis of systems, it was found that not all systems fully automate all stages of machine learning: the TPOT system cannot work with data containing gaps, and categorical features must be pre-coded into numerical ones. Many systems use only statistical methods at the data preprocessing stage to clean up missing values (Mljar excludes rows with gaps or fills them with mean, mode, minimum value; Auto-Sklearn fills a gap in numeric data with mean, median, most frequent value), but do not remove anomalous values and do not reduce the dimension of the feature space. However, unsupervised machine learning methods can also be used for data preprocessing, however, this will require significantly complicating the implementation of the system modules responsible for data preprocessing.

Thus, we can conclude that the considered AutoML systems can hardly be called universal. Data scientists ultimately have to make a choice about which system to use or resort to their own analysis, data processing and selection of the most effective configuration of a machine learning model.

We can suggest a number of promising areas of development for AutoML systems: introduction of unsupervised machine learning algorithms for data preparation; explicit choice of hyperparameter optimization algorithm; the need for more intelligent data pre-processing, since a unified processing method may not be suitable for different datasets.

In the end, it should be noted that AutoML systems are quite an important area for business (Light AutoML, which was developed for the needs of a financial company) [5]. Research in this area will help develop a general methodology for building automated machine learning systems, which helped to design a system for the needs of a particular industry or offer a more general solution.

[1] *Karmaker, S., Hassan, M.M., Smith, M.J., Xu, L., Zhai, C., Veeramachaneni, K.* AutoML to Date and Beyond: Challenges and Opportunities // ACM Computing Surveys (CSUR) 54, 2022.— P.1–36.

[2] *Koroteev, M.V.* Review of some modern trends in machine learning technology // E-Management 1(1), 2018.— P.26–35.

[3] Car Dekho Data, `https://www.kaggle.com/datasets/shindenikhil/car-dekho-data`. Last accessed 30 October 2022.

[4] Combined Cycle Power Plant Dataset, `https://archive.ics.uci.edu/ml/datasets/Combined+Cycle+Power+Plant`. Last accessed 30 October 2022.

[5] *Vakhrushev A. et al.* LightAutoML: AutoML Solution for a Large Financial Services Ecosystem // arXiv preprint arXiv:2109.01528, 2021.

Тестирование инвариантами в применении к задаче тестирования рекомендательных систем

Якушева Софья Федоровна[1]⋆ yakusheva.sf@phystech.edu
Хританков Антон Сергеевич[2] akhritankov@hse.ru
[1]Долгопрудный, Московский физико-технический институт
[2]Москва, НИУ ВШЭ

Рекомендательные системы в наше время активно используются в онлайн-сервисах – магазинах, кинотеатрах, новостных порталах. Одной из рассматриваемых в этой сфере задач является задача "многорукого бандита". В ней в качестве алгоритма рекомендации используется алгоритм для автомата с несколькими ручками, одну или несколько из которых пользователь может “сыграть”, за что алгоритм получит награду. Цель алгоритма – подобрать стратегию выбора ручек для пользователя так, чтобы минимизировать потери.

Обеспечение высокого уровня доверия к работе алгоритма рекомендаций является жизненно важной задачей для подобных систем – в противном случае ими просто не будут пользоваться. Вводимый в эксплуатацию сервис должен приносить благо обществу, удоблетворять требованиям Кодекса этики в сфере ИИ и соответствовать принятым стандартам, например, ISO-24028. Чтобы не допустить попадания к конечным пользователям систем, не соответствующих заявленным к ним требованиям, необходимо разрабатывать точные и объективные методы проверки соответствия и оценки уровня доверия. Одним из вариантов проверки является автоматическое тестирование. Однако привычные методы тестирования путём сравнения ответов системы с эталонными не работают, поскольку эталонов попросту не существует. Иногда даже нельзя определить, что вообще является ответом. Эта проблема известна как проблема тестового оракула [1].

В подобных ситуациях можно использовать metamorphic testing, или тестирование инвариантами [2]. Основной идеей этого метода является не проверка правильности ответа на каждом конкретном тесте, а проверка выполнения тестовых инвариантов (metamorphic relations) — отношений вида

$$\mathcal{R}(x_1, x_2, ..., x_n, f(x_1), f(x_2), \ldots, f(x_n)) \longrightarrow \{0, 1\}, \tag{1}$$

где $f(x_i)$- результат работы программы на i-м тестовом входе. В более кратком виде инвариант можно записать как функцию от матрицы X размера n:

$$\mathcal{R}(X, f(X)) \longrightarrow \{0, 1\}. \tag{2}$$

Однако детерминированные инварианты могут быть неэффективны для тестирования систем, допускающих возникновение некоторого количества ошибок в процессе работы. В подобных случаях можно использовать стохастические инварианты вида

$$\mathcal{S}(X^1, \ldots, X^n, f(X^1), \ldots, f(X^n)) \longrightarrow (\mathcal{R}(X^1, f(X^1)), \ldots, \mathcal{R}(X^n, f(X^n)), \tag{3}$$

то есть рассматривать большое количество проверок как единый инвариант. Это позволит количественно оценить ошибки, а также более детально изучить области входных данных, на которых эти ошибки происходят.

Стохастический инвариант можно при необходимости сделать детерминированными, если рассмотреть его в терминах статистических гипотез.

H_0: Инвариант $\mathcal{S}$ выполняется.

H_1: Инвариант $\mathcal{S}$ не выполняется.

Обозначим функцию детерминирования инварианта как $\mathcal{T}(\mathcal{S})$. В результате проверки гипотеза либо будет отвергнута, либо не будет отвергнута с заданным уровнем значимости.

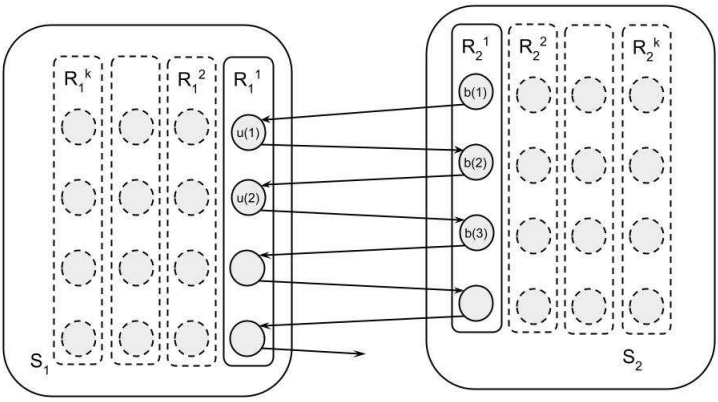


Рис. 1. Пример составления стохастических инвариантов $\mathcal{S}_1$ и $\mathcal{S}_2$ из детерминированных $\mathcal{R}_1$ и $\mathcal{R}_2$, вычисленных для k параллельных запусков системы из пользователя (u) и многорукого бандита (b). u(i) и b(i) - данные, получаемые на шаге i.

В данной работе рассматривается вопрос применимости и реализуемости тестирования стохастическими инвариантами для различных алгоритмов работы с многорукими бандитами.

Постановка задачи. Разработать инварианты для обнаружения некорректного поведения системы и проверить их реализуемость.

Модель системы. Алгоритмы бандита и пользователя рассматриваются как “чёрные ящики” с известными входными параметрами и принципами работы. Параметрами задачи являются количество ручек (N), число выбираемых ручек (A), длина эксперимента (L) и параметры пользователя (F).

Инварианты. Для оценки качества работы системы используется следующий инвариант, получаемый из общих соображений о работоспособности системы. *Если пользователь меняет свои предпочтения не чаще чем раз в x шагов, то бандит успевает обучаться до такой степени, что контрольная метри-*

ка становится не ниже доли y от метрики, достигаемой на пользователе без изменения предпочтений. Это свойство сложно проверить традиционными метриками наподобие ROC-AUC или эмпирического риска, поскольку трудно сформулировать критерии для их поведения в процессе обучения. Проверка этого инварианта сводится к индивидуальным проверкам пользователя и бандита.

Байесовская сеть. Чтобы оценить зависимости между инвариантами для частей системы, построим байесовскую сеть. Входами сети будут все входные параметры запуска, а выходами - итоговые оценки работы системы. В качестве латентных переменных будем рассматривать инварианты компонент системы - $\mathcal{R}_{i,j}$. Это позволит как рассмотреть сложные взаимосвязи между зависимыми переменными, так и оценить вероятности выполнения инвариантов компонент.

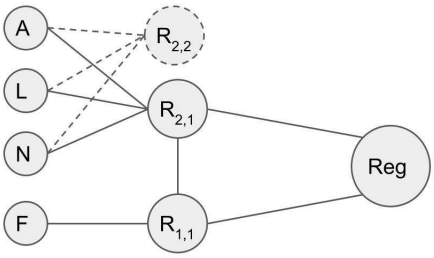


Рис. 2. Пример байесовской сети, построенной для системы из пользователя и многорукого бандита. $\mathcal{R}_{1,1}$ - инвариант пользователя, $\mathcal{R}_{2,\{1,2\}}$ - инварианты бандита.

Заключение. В работе предложен новый способ анализа сложных систем, к которым не применимо тестирование с оракулом. Преимуществом метода является возможность строгой проверки наличия необходимых свойств [3], которые нельзя оценить традиционными метриками качества. Также метод позволяет оценивать взаимосвязи между ошибками в разных частях.

[1] *Barr E.T., Harman M., McMinn P., Shahbaz M., Yoo S.*, The Oracle Problem in Software Testing: A Survey// IEEE Transactions on Software Engineering, vol. 41, n. 5, 2015. — p. 507-525.

[2] *Chen T.Y., Kuo F.-C., Liu H., Poon P.-L., Towey D., Tse T.H., Zhou Z.Q.*, Metamorphic Testing: A Review of Challenges and Opportunities // New York: Association for Computing Machinery, vol. 51, n. 1, 2018.

[3] *Sculley D., Holt G., Golovin D., Davydov E. et al.* Hidden Technical Debt in Machine Learning Systems. NIPS, January 2015. —p. 2494-2502.

Testing by invariants as applied to the problem of testing recommender systems

Yakusheva Sofiya[1]⋆ yakusheva.sf@phystech.edu
Khritankov Anton[2] akhritankov@hse.ru
[1]Dolgoprudny, Moscow Institute of Physics and Technology
[2]Moscow, Higher School of Economics

Recommender systems in our time are actively used in online services - shops, cinemas, news portals. One of the problems considered in this area is the problem of the "many-armed bandit". It uses as a recommendation algorithm an algorithm for an automaton with several handles, one or more of which the user can "play", for which the algorithm will receive a reward. The purpose of the algorithm is to choose a handle selection strategy for the user in such a way as to minimize losses.

Ensuring a high level of confidence in the operation of the recommendation algorithm is a vital task for such systems - otherwise they simply will not be used. The service being deployed must be socially beneficial, comply with the AI Code of Ethics, and comply with accepted standards such as ISO-24028. Accurate and objective methods for verifying compliance and evaluating the level of trust must be developed to prevent end users from reaching systems that do not meet their stated requirements. One of the verification options is automated testing. However, the usual testing methods by comparing the responses of the system with the reference ones do not work, since there are simply no standards. Sometimes you can't even tell what the answer is. This problem is known as the test oracle problem [1].

In such situations, you can use metamorphic testing, or testing by invariants [2]. The main idea of this method is not to check the correctness of the answer on each specific test, but to check the fulfillment of test invariants (metamorphic relations) — relations of the form

$$\mathcal{R}(x_1, x_2, ..., x_n, f(x_1), f(x_2), \ldots, f(x_n)) \longrightarrow \{0, 1\}, \tag{1}$$

where $f(x_i)$ is the result of the program on the i-th test input. In a shorter form, the invariant can be written as a function of the matrix X of size n:

$$\mathcal{R}(X, f(X)) \longrightarrow \{0, 1\}. \tag{2}$$

However, deterministic invariants may not be effective for testing systems that allow a certain number of errors to occur during operation. In such cases, one can use stochastic invariants of the form

$$\mathcal{S}(X^1, \ldots, X^n, f(X^1), \ldots, f(X^n)) \longrightarrow (\mathcal{R}(X^1, f(X^1)), \ldots, \mathcal{R}(X^n, f(X^n)), \tag{3}$$

that is, consider a large number of checks as a single invariant. This will allow you to quantify the errors, as well as to study in more detail the areas of the input data on which these errors occur.

The stochastic invariant can be made deterministic, if necessary, if we consider it in terms of statistical hypotheses.

H_0: $\mathcal{S}$ invariant holds.

H_1: $\mathcal{S}$ invariant fails.

Denote the invariant determination function as $\mathcal{T}(\mathcal{S})$. As a result of testing, the hypothesis will either be rejected or will not be rejected with a given level of significance.

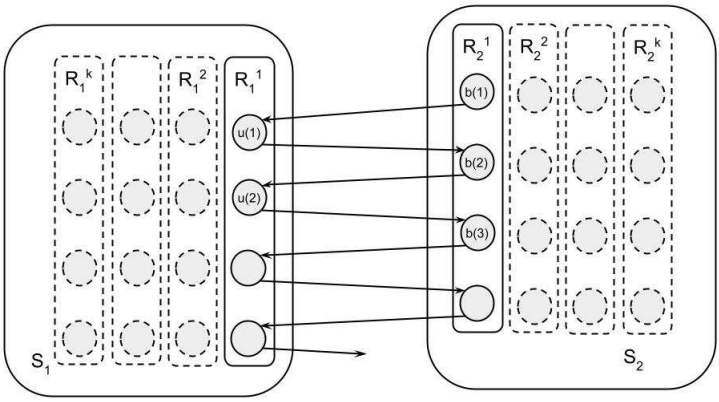


Fig. 1. An example of compiling stochastic invariants $\mathcal{S}_1$ and $\mathcal{S}_2$ from deterministic $\mathcal{R}_1$ and $\mathcal{R}_2$ computed for k parallel system launches from the user (u) and the multi-armed bandit (b). u(i) and b(i) are data received at step i.

This paper considers the applicability and feasibility of testing by stochastic invariants for various algorithms for working with multi-armed bandits.

Problem statement. Develop invariants for detecting incorrect behavior of the system and check their feasibility.

System model. Bandit and user algorithms are considered as "black boxes" with known input parameters and operating principles. The task parameters are the number of handles (N), the number of selectable handles (A), the length of the experiment (L) and the user parameters (F).

Invariants. The following invariant is used to assess the quality of the system's operation, which is obtained from general considerations about the system's performance. *If the user changes his preferences no more than once every x steps, then the bandit has time to learn to such an extent that the control metric becomes at least the fraction y of the metric achieved on the user without changing preferences.* This property is difficult check with traditional metrics like ROC-AUC or empirical risk, as it is difficult to formulate criteria for their behavior in the learning process. Checking this invariant is reduced to individual checks of the user and the bandit.

Bayesian network. To evaluate dependencies between invariants for parts of the system, let's build a Bayesian network. The inputs of the network will be all the input parameters of the launch, and the outputs will be the final estimates of the system. As latent variables we will consider the invariants of the system components - $\mathcal{R}_{i,j}$. This will allow both considering complex relationships between dependent variables and estimating the probabilities of fulfilling component invariants.

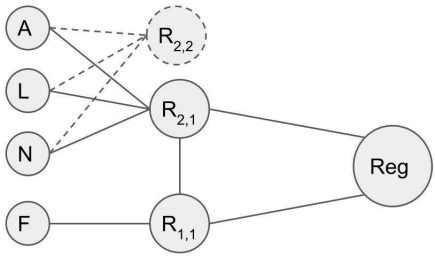


Fig. 2. An example of a Bayesian network built for a system of a user and a multi-armed bandit. $\mathcal{R}_{1,1}$ - user invariant, $\mathcal{R}_{2,\{1,2\}}$ - bandit invariants.

Conclusion. The paper proposes a new way to analyze complex systems that cannot be tested with an oracle. The advantage of the method is the ability to rigorously check the presence of the necessary properties [3], which cannot be assessed by traditional quality metrics. Also, the method allows you to evaluate the relationship between errors in different parts.

[1] *Barr E.T., Harman M., McMinn P., Shahbaz M., Yoo S.*, The Oracle Problem in Software Testing: A Survey// IEEE Transactions on Software Engineering, vol. 41, n. 5, 2015. — p. 507-525.

[2] *Chen T.Y., Kuo F.-C., Liu H., Poon P.-L., Towey D., Tse T.H., Zhou Z.Q.*, Metamorphic Testing: A Review of Challenges and Opportunities // New York: Association for Computing Machinery, vol. 51, n. 1, 2018.

[3] *Sculley D., Holt G., Golovin D., Davydov E. et al.* Hidden Technical Debt in Machine Learning Systems. NIPS, January 2015. —p. 2494-2502.

Теоретико-информационная нижняя граница погрешности для оценки параметра плотности распределения вероятностей

Ланге Михаил Михайлович[1]⋆ lange_mm@mail.ru
Ланге Андрей Михайлович[1] lange_am@mail.ru
[1]Москва, Федеральный исследовательский центр «Информатика и управление» Российской академии наук

Задача исследования. Исследуется наименьшая средняя погрешность оценивания параметра плотности распределения по выборке независимых наблюдений как функция средней взаимной информации между выборками и значениями оценок. При заданной плотности распределения параметра и квадратичной мере погрешности указанная зависимость строится в форме обращения известной в теории информации функции «скорость-погрешность» (rate distortion function) [1].

Пусть $p_X(x|\Theta)$ – условная плотность распределения случайной величины x на множестве X, где θ — неизвестный случайный параметр, принимающий значения на множестве Θ с априорной плотностью $p_\Theta(\theta)$. Полагая, что оценка $\hat{\theta}$ строится по выборке независимых наблюдений $x^n = (x_1, \ldots, x_n)$, а погрешность вычисляется по квадратичной мере $(\hat{\theta} - \theta)^2$, вводятся средняя погрешность

$$E_{q_{\hat{\Theta}}}(X^n;\Theta) = \int_{X^n} p_{X^n}(x^n)dx^n \int_{\hat{\Theta}} q_{\hat{\Theta}}(\hat{\theta}|x^n)d\hat{\theta} \int_{\Theta} (\theta - \hat{\theta})^2 p_\Theta(\theta|x^n)d\theta \qquad (1)$$

и средняя взаимная информация

$$I_{q_{\hat{\Theta}}}(X^n;\Theta) = \int_{X^n} p_{X^n}(x^n)dx^n \int_{\hat{\Theta}} q_{\hat{\Theta}}(\hat{\theta}|x^n) \ln(q_{\hat{\Theta}}(\hat{\theta}|x^n)/q_{\hat{\Theta}}(\hat{\theta}))d\hat{\theta}, \qquad (2)$$

которые зависят от условной плотности $q_{\hat{\Theta}}(\hat{\theta}|x^n)$. Здесь $p_{X^n}(x^n)$ и $q_{\hat{\Theta}}(\hat{\theta})$ — безусловные плотности распределений на множествах X^n и $\hat{\Theta}$. Функционалы (1) и (2) позволяют минимизировать $I_{q_{\hat{\Theta}}}(X^n;\Theta)$ по плотности $q_{\hat{\Theta}}(\hat{\theta}|x^n)$ при условии $E_{q_{\hat{\Theta}}}(X^n;\Theta) \leqslant \varepsilon$, где ε — заданная допустимая погрешность. Указанный минимум дает функцию $R_n(\varepsilon)$, которая аналогична функции «скорость-погрешность» для для моделей кодирования независимых непрерывных сообщений, переданных по каналу с аддитивным гауссовым шумом [2]. Задача состоит в построении монотонно убывающей нижней границы $\underline{R}_n(\varepsilon)$ для функции $R_n(\varepsilon)$. Тогда обратная функция $\underline{R}_n^{-1}(I)$ дает нижнюю границу средней погрешности при любом фиксированном значении средней взаимной информации $I_{q_{\hat{\Theta}}}(X^n;\hat{\Theta}) = I$.

Нижняя граница функции $R_n(\varepsilon)$. Применяя технику вычисления функции «скорость-погрешность», изложенную в монографии [1], получена нижняя граница

$$R_n(\varepsilon) \geqslant \underline{R}_n(\varepsilon) = h(p_{\Theta_n}) - 2^{-1}\ln 2\pi e(\varepsilon - \varepsilon_{\min}^{(n)}), \qquad (3)$$

где $\varepsilon_{\min}^{(n)} < \varepsilon \leqslant \varepsilon_{\max}^{(n)}$. Здесь $h(p_{\Theta_n}) = -\int_{\Theta_n} p_{\Theta_n}(\theta_n) \ln p_{\Theta_n}(\theta_n) d\theta_n$ — дифференциальная энтропия от плотности распределения $p_{\Theta_n}(\theta_n)$ значений $\theta_n(x^n) = \int_{\Theta} \theta p_{\Theta}(\theta|x^n) d\theta$ на множестве Θ_n; $p_{\Theta}(\theta|x^n)$ — апостериорная плотность на множестве Θ; $\sigma_n^2(x^n) = \int_{\Theta} \theta^2 p_{\Theta}(\theta|x^n) d\theta - \left(\int_{\Theta} \theta p_{\Theta}(\theta|x^n) d\theta\right)^2$ — дисперсия плотности $p_{\Theta}(\theta|x^n)$ и $\varepsilon_{\min}^{(n)} = \int_{X^n} p_{X^n}(x^n) \sigma_n^2(x^n) dx^n$. Граница (3) монотонно убывает, $\underline{R}_n(\varepsilon_{\min}^{(n)}) \to \infty$ и $\underline{R}_n(\varepsilon_{\min}^{(n)}) = 0$. В [3] аналогичный подход использован для построения нижней границы вероятности ошибки классификации объектов в пространстве представлений с заданным расстоянием.

Пример вычисления границы $\underline{R}_n(\varepsilon)$. Граница вида (3) вычислена для гауссовой плотности $p_X(x|\theta) \sim N(\theta, \sigma^2)$ с неизвестным средним значением θ и известной дисперсией σ^2. Значения параметра θ имеют нормальную плотность распределения $p_{\Theta}(\theta) \sim N(\theta_0, \sigma_0^2)$ со средним $\theta_0 = 0$ и дисперсией σ_0^2. В этом случае апостериорная плотность является гауссовой $p_{\Theta}(\theta|x^n) \sim N(\theta_n, \sigma_n^2)$ со средним значением $\theta_n = a_n \sum_{k=1}^n x_k$ и дисперсией $\sigma_n^2 = \sigma_0^2\sigma^2/(n\sigma_0^2 + \sigma^2)$, где $a_n = \sigma_0^2/(n\sigma_0^2 + \sigma^2)$ [4]. Используя систему преобразований $\theta_n = a_n \sum_{k=1}^n x_k$, $n = 1, 2, \ldots$, и вычисляя $(n-1)$– кратную свертку нормальных плотностей, получена гауссова плотность распределения $p_{\Theta_n}(\theta_n) \sim N(0, a_n^2 n(\sigma_0^2 + \sigma^2))$ значений наилучшей оценки и дифференциальная энтропия $h(p_{\Theta_n}) = 2^{-1} \ln 2\pi e a_n^2 n(\sigma_0^2 + \sigma^2)$. Подстановка найденной дифференциальной энтропии в (3) дает границу

$$\underline{R}_n(\varepsilon) = 2^{-1} \ln a_n^2 n(\sigma_0^2 + \sigma^2) - 2^{-1} \ln(\varepsilon - \varepsilon_{\min}^{(n)}), \tag{4}$$

где $\varepsilon_{\min}^{(n)} = \sigma_0^2\sigma^2/(n\sigma_0^2 + \sigma^2)$, $\varepsilon_{\max}^{(n)} = \varepsilon_{\min}^{(n)} \left(1 + n\sigma_0^2(\sigma_0^2 + \sigma^2)/(n\sigma_0^2 + \sigma^2)\sigma^2\right) \leqslant \sigma_0^2$. Граница вида (4) сохраняется и в случае $\theta_0 \neq 0$, когда $\theta_n = a_n \sum_{k=1}^n x_k + (\sigma^2/(n\sigma_0^2 + \sigma^2))\theta_0$.

При $n = 1$ имеем $\varepsilon_{\min}^{(1)} \geqslant \varepsilon_{\min}^{(n)}$ и $\varepsilon_{\max}^{(1)} = \sigma_0^2$. В этом случае формула (4) дает нижнюю границу скорости кодирования независимых и одинаково распределенных гауссовых величин θ, переданных по каналу с аддитивным гауссовым шумом $x - \theta$ [2]. В случае $n = 1$ и $\sigma^2 = 0$ имеем $x = \theta$, $\varepsilon_{\min}^{(1)} = 0$, $\varepsilon_{\max}^{(1)} = \sigma_0^2$ и граница (4) совпадает с известной формулой $\underline{R}_1(\varepsilon) = 2^{-1} \ln(\sigma_0^2/\varepsilon)$ [1]. Графические иллюстрации границы (4) при $n = 1$ и $n > 1$ даны на рисунке 1.

Практическая значимость границы $\underline{R}_n(\varepsilon)$ состоит в возможности ее применения для вычисления избыточности погрешности при различных способах построения оценки параметра по выборкам квантованных наблюдений $\hat{X}^n = \{\hat{x}^n = (\hat{x}_1, \ldots, \hat{x}_n)\}$. Полагая, что при фиксированном размере выборки выбранный способ порождает множество оценок $\hat{\Theta} = \{\hat{\theta}(\hat{x}^n), \forall \hat{x}^n \in \hat{X}^n\}$ со средней погрешностью $E(\hat{X}^n; \hat{\Theta})$ и средней взаимной информацией $I(\hat{X}^n; \hat{\Theta})$, избыточность средней погрешности относительно нижней границы определяется величиной $r = E - \underline{R}_n^{-1}(I)$.

Основные выводы. В рамках вероятностной модели оценивания непрерывного параметра по выборке независимых наблюдений получена аналитическая

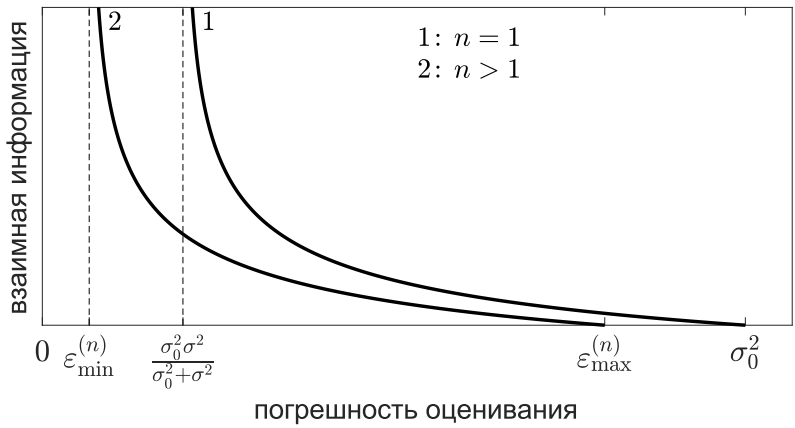


Рис. 1. Поведение функции $\underline{R}_n(\varepsilon)$ для гауссовой модели.

нижняя граница средней погрешности как функция средней взаимной информации между множеством наблюдений и множеством возможных оценок. Приведен пример границы для гауссовой модели. Отмечено практическое применение нижней границы для вычисления избыточности средней погрешности при любом способе построения оценки. Рассмотренный подход допускает обобщение для векторного параметра.

[1] *Berger T.* Rate Distortion Theory: A Mathematical Basis for Data Compression // Englewood Cliffs, New Jersey: Prentice-Hall Inc., 1971.

[2] *Dobrushin R. L., Tsybakov B. S.* // IRE Trans. on Inform. Theory, 1962. —8(5). — p. 293–304.

[3] *Lange M. M., Lange A. M.* Information-theoretic lower bounds to error probability for the models of noisy discrete source coding and object classification // Pattern Recognition and Image Analysis, 2022. — 32(3). — p. 570–574.

[4] *Duda R. O., Hart P. E., and Stork D. G.* Pattern Classification, 2nd ed. // New York: Wiley & Sons, 2001.

Information-theoretic lower bound to estimation error for a parameter of a given probability distribution density

Lange Mikhail[1]⋆ lange_mm@mail.ru
Lange Andrey[1] lange_am@mail.ru
[1]Moscow, Federal Reseach Center "Computer Science and Control" of the Russian Academy of Sciences

Research task. Given probability distribution density with a random parameter, a minimal average error to estimate the parameter via a sample of the independent observations as a function of an average mutual information between the values of the observations and the estimates is investigated. For a known probability density of the parameter and a square distortion measure, the specified function is made by an inversion of the rate distortion function that is known in the information theory [1].

Let $p_X(x|\Theta)$ be a conditional distribution density of a random variable x in a set X, where θ is an unknown random parameter which is taken on the values in a set Θ by a prior density $p_\Theta(\theta)$. Taking into account that an estimate $\hat{\theta}$ is constructed over a sample $x^n = (x_1, \ldots, x_n)$ of independent observations and an estimation error is calculated by the square distortion measure $(\hat{\theta}-\theta)^2$, we define the average estimation error

$$E_{q_{\hat{\Theta}}}(X^n;\Theta) = \int_{X^n} p_{X^n}(x^n)dx^n \int_{\hat{\Theta}} q_{\hat{\Theta}}(\hat{\theta}|x^n)d\hat{\theta} \int_{\Theta} (\theta-\hat{\theta})^2 p_\Theta(\theta|x^n)d\theta \qquad (1)$$

and the average mutual information

$$I_{q_{\hat{\Theta}}}(X^n;\Theta) = \int_{X^n} p_{X^n}(x^n)dx^n \int_{\hat{\Theta}} q_{\hat{\Theta}}(\hat{\theta}|x^n) \ln(q_{\hat{\Theta}}(\hat{\theta}|x^n)/q_{\hat{\Theta}}(\hat{\theta}))d\hat{\theta} \qquad (2)$$

that are the functionals depending on a free conditional distribution density $q_{\hat{\Theta}}(\hat{\theta}|x^n)$. Here, $p_{X^n}(x^n)$ and $q_{\hat{\Theta}}(\hat{\theta})$ are the unconditional distribution densities in the sets X^n and $\hat{\Theta}$ respectively. The functionals (1) and (2) allow us to minimize $I_{q_{\hat{\Theta}}}(X^n;\Theta)$ over the density $q_{\hat{\Theta}}(\hat{\theta}|x^n)$ when $E_{q_{\hat{\Theta}}}(X^n;\Theta) \leqslant \varepsilon$ subject to a given admissible estimation error ε. This conditional minimum yields a function $R_n(\varepsilon)$ which is similar to the rate distortion function for encoding the independent continues letters with the additive Gaussian noise [2]. The task is to construct a strictly decreasing lower bound $\underline{R}_n(\varepsilon)$ to the function $R_n(\varepsilon)$. Then the inverse function $\underline{R}_n^{-1}(I)$ yields the lower bound to the average estimation error subject to any fixed value of the average mutual information $I_{q_{\hat{\Theta}}}(X^n;\hat{\Theta}) = I$.

Lower bound to the function $R_n(\varepsilon)$. Using a technique to calculate the rate distortion function [1], we have obtained the following lower bound

$$R_n(\varepsilon) \geqslant \underline{R}_n(\varepsilon) = h(p_{\Theta_n}) - 2^{-1}\ln 2\pi e(\varepsilon - \varepsilon_{\min}^{(n)}), \qquad (3)$$

where $\varepsilon^{(n)}_{\min} < \varepsilon \leqslant \varepsilon^{(n)}_{\max}$. Here, $h(p_{\Theta_n}) = -\int_{\Theta_n} p_{\Theta_n}(\theta_n) \ln p_{\Theta_n}(\theta_n) d\theta_n$ is a differential entropy for a distribution density $p_{\Theta_n}(\theta_n)$ of the estimates $\theta_n(x^n) = \int_\Theta \theta p_\Theta(\theta|x^n) d\theta$ in the set Θ_n; $p_\Theta(\theta|x^n)$ is a posterior distribution density in the set Θ; $\sigma^2_n(x^n) = \int_\Theta \theta^2 p_\Theta(\theta|x^n) d\theta - \left(\int_\Theta \theta p_\Theta(\theta|x^n) d\theta\right)^2$ is a dispersion of the density $p_\Theta(\theta|x^n)$, and $\varepsilon^{(n)}_{\min} = \int_{X^n} p_{X^n}(x^n) \sigma^2_n(x^n) dx^n$ is a minimal average estimation error. The bound (3) is decreased strictly, $\underline{R}_n(\varepsilon^{(n)}_{\min}) \to \infty$, and $\underline{R}_n(\varepsilon^{(n)}_{\min}) = 0$. In [3], the similar approach has been used to construct the lower bound to an error probability for classifying objects in a space of their representations with a given distance.

Example of calculating the bound $\underline{R}_n(\varepsilon)$. The calculation of the bound (3) has been performed for the Gaussian density $p_X(x|\theta) \sim N(\theta, \sigma^2)$ with an unknown mean value θ and a given dispersion σ^2. Also, the values of the parameter θ are distributed by the normal density $p_\Theta(\theta) \sim N(\theta_0, \sigma^2_0)$ with the zero mean $\theta_0 = 0$ and the dispersion σ^2_0. In this case, the posterior density is Gaussian $p_\Theta(\theta|x^n) \sim N(\theta_n, \sigma^2_n)$ with the mean $\theta_n = a_n \sum_{k=1}^n x_k$ and the dispersion $\sigma^2_n = \sigma^2_0 \sigma^2/(n\sigma^2_0 + \sigma^2)$, where $a_n = \sigma^2_0/(n\sigma^2_0 + \sigma^2)$ [4]. Using the transformations $\theta_n = a_n \sum_{k=1}^n x_k$, $n = 1, 2, \ldots$, and calculating the appropriate $(n-1)$-times composition of the normal densities, we obtain the Gaussian distribution density $p_{\Theta_n}(\theta_n) \sim N(0, a^2_n n(\sigma^2_0 + \sigma^2))$ for the finest estimate and the corresponding differential entropy $h(p_{\Theta_n}) = 2^{-1} \ln 2\pi e a^2_n n(\sigma^2_0 + \sigma^2)$. The next substitution of this entropy in (3) yields the following bound

$$\underline{R}_n(\varepsilon) = 2^{-1} \ln a^2_n n(\sigma^2_0 + \sigma^2) - 2^{-1} \ln(\varepsilon - \varepsilon^{(n)}_{\min}), \tag{4}$$

where $\varepsilon^{(n)}_{\min} = \sigma^2_0 \sigma^2/(n\sigma^2_0 + \sigma^2)$ and $\varepsilon^{(n)}_{\max} = \varepsilon^{(n)}_{\min}\left(1 + n\sigma^2_0(\sigma^2_0 + \sigma^2)/(n\sigma^2_0 + \sigma^2)\sigma^2\right) \leqslant \sigma^2_0$. Notice that the bound of the form (**??**) is true when $\theta_0 \neq 0$ and $\theta_n = a_n \sum_{k=1}^n x_k + (\sigma^2/(n\sigma^2_0 + \sigma^2))\theta_0$.

For $n = 1$, we have $\varepsilon^{(1)}_{\min} \geqslant \varepsilon^{(n)}_{\min}$ and $\varepsilon^{(1)}_{\max} = \sigma^2_0$. In this case, the form (4) yields the lower bound to a code rate in a scheme of coding the independent and identically distributed Gaussian values θ after their transmission over a channel with the additive Gaussian noise $x - \theta$ [2]. When $n = 1$ and $\sigma^2 = 0$, we have $x = \theta$, $\varepsilon^{(1)}_{\min} = 0$, $\varepsilon^{(1)}_{\max} = \sigma^2_0$ and the bound (4) gives the well known formula $\underline{R}_1(\varepsilon) = 2^{-1} \ln(\sigma^2_0/\varepsilon)$ [1]. For $n = 1$ and $n > 1$, the graphic illustrations of the bound (4) are shown in the following figure.

In practice, the bound $\underline{R}_n(\varepsilon)$ can be applied in calculating a redundancy of the estimation error for the different techniques of making the parameter estimates via the quantized observations $\hat{X}^n = \{\hat{x}^n = (\hat{x}_1, \ldots, \hat{x}_n)\}$. If, for a fixed sample size, a given technique produces the set of the estimates $\hat{\Theta} = \{\hat{\theta}(\hat{x}^n), \forall \hat{x}^n \in \hat{X}^n\}$ with the average estimation error $E(\hat{X}^n; \hat{\Theta})$ and the average mutual information $I(\hat{X}^n; \hat{\Theta})$, the redundancy of the average estimation error relative to the lower bound is defined by $r = E - \underline{R}^{-1}_n(I)$.

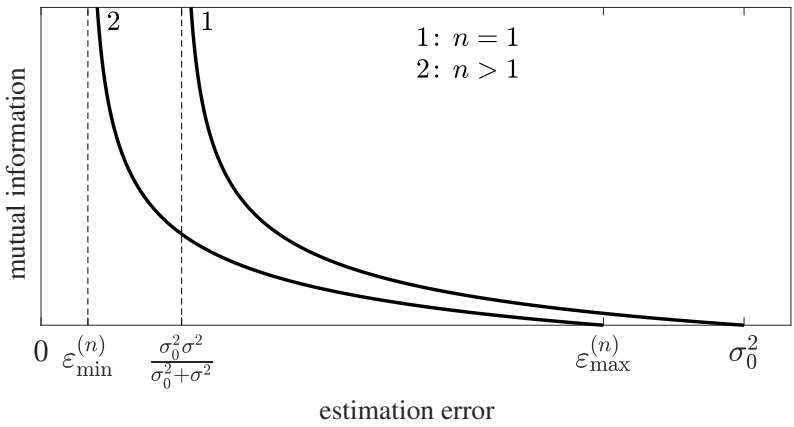


Fig. 1. The sketches of the function $\underline{R}_n(\varepsilon)$ for the Gaussian model.

Main terminals. Within the probabalistic model of estimating a continuous parameter via a sample of the independent observations, the analytic lower bound to the average estimation error depending on the average mutual information between the observations and the estimates has been constructed. For the Gaussian model, the example of the bound has been calculated. In general, the obtained bound gives a possibility to calculate the redundancy of the estimation error for any technique of making the parameter estimates via the quantized samples of the observations. The developed approach can be extended for the case of a vector parameter.

[1] *Berger T.* Rate Distortion Theory: A Mathematical Basis for Data Compression // Englewood Cliffs, New Jersey: Prentice-Hall Inc., 1971.

[2] *Dobrushin R. L., Tsybakov B. S.* // IRE Trans. on Inform. Theory, 1962. —8(5). — p. 293–304.

[3] *Lange M. M., Lange A. M.* Information-theoretic lower bounds to error probability for the models of noisy discrete source coding and object classification // Pattern Recognition and Image Analysis, 2022. — 32(3). — p. 570–574.

[4] *Duda R. O., Hart P. E., and Stork D. G.* Pattern Classification, 2nd ed. // New York: Wiley & Sons, 2001.

Новый ансамблевый метод прогнозирования свойств химических соединений

Сенько Олег Валентинович[1]⋆ senkoov@mail.ru
Докукин Александр Александрович[1] dalex@ccas.ru
Киселёва Надежда Николаевна [2] kis-japan@mail.ru
Кузнецова Юлиана Олеговна[2] jul1998@ya.ru
Дударев Виктор Анатольевич[2] vicdudarev@mail.ru
[1]Москва, ФИЦ ИУ РАН
[2]Москва, ИМЕТ им. Байкова РАН

Ансамблевые методы получили значительное распространение в машинном обучении. Они успешно используются при решении прикладных задач в самых различных областях. Наибольшее распространение получили методы, использующие ансамбли решающих деревьев. Наиболее известными ансамблевыми методами являются случайный лес, а также различные варианты метода градиентный бустинг. Работу ансамблевого метода всегда можно представить в виде двухуровневой схемы. Первый уровень состоит из так называемых слабых предикторов, предсказывающих значения целевой переменной Y по исходным признаковым описаниям. На втором уроне находится алгоритм, вычисляющий коллективное решения по прогнозам, рассчитанным на первом уровне. Распространённым способом получения коллективного решения является использование арифметического среднего. Однако могут использоваться и более сложные стэкинговые схемы, в которых выходы алгоритмов первого уровня используются в качестве входных признаков для алгоритма второго уровня. Источником высокой обобщающей способности ансамблевых методов является взаимное расхождение прогнозов, вычисляемых алгоритмами ансамбля. Состоятельность данного предположения следует из разложения ошибки выпуклой комбинации алгоритмов некоторого ансамбля $A_1, \ldots, A_r$. Пусть $\hat{A} = \sum_{i=1}^{r} c_i A_i$, где $\sum_{i=1}^{r} c_i = 1$, $c_i \geqslant 0, i = 1, \ldots, r$. Тогда

$$\mathbb{E}(Y - \hat{A})^2 = \frac{1}{m} \sum_{i=1}^{r} c_i \delta_i - \frac{1}{r} \sum_{i=1}^{r} c_i \mathbb{E}(A_i - \hat{A})^2, \qquad (1)$$

где $\delta_i = \mathbb{E}(Y - A_i)^2$.

Целью работы является разработка методов построения по обучающей выборке $S = \{(y_1, \mathbf{x}_1), \ldots, (y_j, \mathbf{x}_j)\}$ ансамбля алгоритмов, для которого достигается минимум ошибки в смыле разложения (1). При этом предполагается, что все алгоритмы имеют при вычислении коллективного решения одинаковый вес. Для достижения данной цели используется процедура, заключающаяся в добавлении в ансамбль на шаге k алгоритма A_k, минимизирующего функционал

$$D(S, \hat{A}_{k-1}) = \frac{1}{m}\sum_{i=1}^{m}[y_j - A_k(\mathbf{x}_j)]^2 - \frac{1}{m}\sum_{i=1}^{m}(A_k(\mathbf{x}_j) - \hat{A}_{k-1}(\mathbf{x}_j))^2$$

Алгоритм A_k ищется как сумма деревьев B_k и T_k, где B_k строится по бутстрэп репликации обучающей выборки S и случайному подмножеству исходного набора признаков. Дерево T_k строится по обучающей выборке $\{(d_1, \mathbf{x}_1), \ldots, (d_m, \mathbf{x}_m)\}$, где $d_1, \ldots, d_m$ вещественный вектор смещений прогнозов, вычисляемых B_k, при котором достигается минимум $D(S, \hat{A}_{k-1})$. Указанная схема формирования ансамбля первого уровня была применена в двухуровневом ансамблевом методе с использованием случайного регрессионного леса в качестве агрегирующего алгоритмоа второго уровня.

Было произведенно сравнение разработанного двухуровневого метода с градиентным бустингом и случайным регрессионным лесом на представительном наборе задач прогнозирования количественных свойств неорганических соединений. Поведённые исследования показали, что на значительной доле задач эффективность разработанного двухуровневого метода превосходит эффективность обоих базовых алгоритмов.

Настоящее исследование поддержано РФФИ, гранты 21-51-53019, 20-01-00609.

[1] *Журавлев Ю.И.*, *Сенько О.В.*, *Докукин А.А.*, *Киселёва Н.Н.*, *Саенко И.А.* Двухуровневый метод регрессионного анализа, использующий ансамбли деревьев с оптимальной дивергенцией // Докл. РАН. Матем., информ., проц. упр., 499, 2021. — С. 63–66.

A new ensemble method for predicting the properties of chemical compounds

Senko Oleg[1]⋆ senkoov@mail.ru
Dokukin Alexandr[1] alex-dok@mail.ru
Kiseleva Nadezhda[2] kis-japan@mail.ru
Kuznetsova Juliana[2] jul1998@ya.ru
Dudarev Victor[2] vic-dudarev@mail.ru
[1]Moscow, FRC Computer Science and Control of RAS
[2]Moscow, Baikov IMET RAS

Ensemble methods have gained significant popularity in machine learning. They are successfully used in solving applied problems in a variety of areas. The most widely used methods use ensembles of decision trees. The most well-known ensemble methods are random forest, as well as various variants of the gradient boosting method. The operation of the ensemble method can always be represented as a two-level scheme. The first level consists of the so-called weak predictors that predict the values of the target variable Y by the initial feature descriptions. At the second level there is an algorithm that calculates a collective solution by the forecasts that have been calculated at the first level. A common way to obtain a collective solution is to use the arithmetic mean. However, more complex stacking schemes can be used, in which the outputs of the first level algorithms are used as input features for the second level algorithm.The source of the high generalizing ability of ensemble methods is the mutual discrepancy between the forecasts calculated by the ensemble algorithms. The validity of this assumption follows from the decomposition of the error of a convex combination of algorithms from some ensemble $A_1, \ldots, A_r$. Let $\hat{A} = \sum_{i=1}^{r} c_i A_i$, where $\sum_{i=1}^{r} c_i = 1$, $c_i \geqslant 0, i = 1, \ldots, r$. Then

$$\mathbb{E}(Y - \hat{A})^2 = \frac{1}{m} \sum_{i=1}^{r} c_i \delta_i - \frac{1}{r} \sum_{i=1}^{r} c_i \mathbb{E}(A_i - \hat{A})^2, \tag{1}$$

where $\delta_i = \mathbb{E}(Y - A_i)^2$. The aim of this work is to develop methods for constructing an ensemble of algorithms by training sample $S = \{(y_1, \mathbf{x}_1), \ldots, (y_j, \mathbf{x}_j)\}$ with the minimal error according decomposition (1).

To achieve this goal, a procedure is used that consists in adding to the ensemble at step k the algorithm A_k, which minimizes the functional

$$D(S, \hat{A}_{k-1}) = \frac{1}{m} \sum_{i=1}^{m} [y_j - A_k(\mathbf{x}_j)]^2 - \frac{1}{m} \sum_{i=1}^{m} (A_k(\mathbf{x}_j) - \hat{A}_{k-1}(\mathbf{x}_j))^2$$

Algorithm A_k is searched as the sum of trees B_k and T_k, where B_k is constructed from the bootstrap replication of the training sample S and a random subset of the original feature set. The T_k tree is constructed from the training

set $\{(d_1, \mathbf{x}_1), \ldots, (d_m, \mathbf{x}_m)\}$, where $d_1, \ldots, d_m$ is a real vector of biases of prognoses calculated by B_k, for which $D(S, \hat{A}_{k-1})$ is minimal. The specified scheme for the formation of the ensemble of the first level was applied in a two-level ensemble method using a random regression forest as an aggregating algorithm of the second level.

The developed two-level method was compared with gradient boosting and random regression forest on a representative set of problems for predicting the quantitative properties of inorganic compounds. It was shown that developed method outperforms two reference methods at great part of tasks.

This research is funded by RFBR, grants 21-51-53019, 20-01-00609.

[1] *Zhuravlev Yu.I.*, *Sen'ko O.V.*, *Dokukin A.A.*, *Kiselyova O.V.*, *Saenko O.V.* Two-level regression method using ensembles of trees with optimal divergence // Dokl. Math., 104:1 (2021), 212–215.

Аналитические выражения для разложения ошибки метода kNN на смещение и разброс

Неделько Виктор Михайлович ⋆nedelko@math.nsc.ru
Новосибирск, Институт математики СО РАН

Метод ближайших соседей (kNN) является одним из наиболее исследованных методов построения решающих функций. В частности, для него общеизвестно аналитическое выражение для разложения ошибки регрессионной модели на смещение и разброс.

Однако упомянутое выражение справедливо только для классической постановки задачи регрессионного анализа, в которой случайной является только целевая переменная, а «объясняющие» переменные неслучайны. Вместе с тем, больший практический интерес представляет постановка задачи, когда все переменные являются случайными.

Пусть X — пространство значений «объясняющих» переменных, а Y — множество значений целевой переменной. Все переменные (в общей постановке) являются случайными величинами с некоторой функцией совместного распределения.

Решающая функция есть отображение $f : X \to Y$, которое строится на основе обучающей выборки объёма N

$$S_N = \left((x^\omega, y^\omega), \omega = \overline{1, N}\right).$$

Качество решающей функции будем оценивать критерием среднего квадрата отклонения (MSE)

$$R(f(\cdot)) = \mathsf{E}_{x,y}(y - f(x))^2.$$

В классической постановке переменные X не случайны, а целевая переменная описывается моделью

$$y(x) = \hat{f}(x) + \delta, \tag{1}$$

где $\hat{f}(x)$ — некоторая неизвестная функция, а δ — случайная величина с нулевым средним и дисперсией σ^2.

Для произвольных независимых случайных величин u и v (с конечными вторыми моментами) справедливо тождество

$$\mathsf{E}(u - v)^2 = \mathsf{D}u + (\mathsf{E}u - \mathsf{E}v)^2 + \mathsf{D}v.$$

Зафиксируем некоторую точку x пространства X и подставим $u = y \,|\, x$, $v = f(x)$. Получаем

$$\mathsf{E}_{S_N, y \,|\, x}(y - f(x))^2 = \mathsf{D}_{y \,|\, x} y + (\mathsf{E}_{y \,|\, x} y - \mathsf{E}_{S_N} f(x))^2 + \mathsf{D}_{S_N} f(x). \tag{2}$$

Обозначение $\mathsf{E}_{S_N, y \,|\, x}$ подразумевает взятие математическиго ожидания по всем выборкам объёма N и по условному распределению переменной y в точке x. Таким образом, нижний индекс при операторах E и D задаёт область усреднения.

Выражение 2 является разложением MSE на «шум», смещение и разброс.

Разложение получено в точке x. При необходимости, выражение 2 можно усреднить по всему пространству X.

Во многих истониках (например [1]) можно найти следующее выражение для ошибки метода kNN

$$\mathsf{E}_{S_N,y\,|\,x}(y-f(x))^2 = \left(f(x) - \frac{1}{k}\sum_{i=1}^{k}\hat{f}(\xi_i(x))\right)^2 + \frac{\sigma^2}{k} + \sigma^2, \tag{3}$$

где $\xi_i(x)$ — координаты i–го «соседа» точки x.

Второе слагаемое этого разложения интерпретируется как разброс.

Выражение получено в предположении, что все координаты x^ω в обучающей выборке фиксированы, т.е. в рамках постановки 1.

Здесь разброс монотонно уменьшается с ростом k, иными словами, монотонно увеличивается с ростом сложности решения, поскольку сложность решения методом kNN уменьшается с ростом k.

Перейдём теперь к постановке задачи построения регрессионной модели в случае, когда «объясняющие» переменные случайны.

Пусть $X = [0,1]^n$ и $y = x_1 + \delta$, где $x = (x_1, ..., x_n) \in X$.

Рассмотрим модель линейной регрессии $\hat{f}(x) = x_1$. Для удобства коэффициент регрессии положили равным 1, поскольку это не ограничивает общность.

Пусть x_j являются независимыми случайными величинами с равномерным распределением, т. е. $x_j \sim U(0,1)$.

Утверждение. Для внутренних точек пространства X для рассматриваемой модели имеет место разложение:

$$\mathsf{E}_{S_N,y\,|\,x}(y-f(x))^2 = \mathsf{D}\left[\frac{1}{k}\sum_{i=1}^{k}\xi_i(x)\right] + \frac{\sigma^2}{k} + \sigma^2, \tag{4}$$

где

$$\mathsf{D}\left[\frac{1}{k}\sum_{i=1}^{k}\xi_i(x)\right] \approx \frac{(N^*)^{-\frac{2}{n}}}{2n^2k^2}\sum_{m=0}^{k-1}\frac{k-m}{m!}\Gamma\left(m+\frac{2}{n}\right) \tag{5}$$

и

$$N^* = NV_0, \quad V_0 = \frac{\pi^{\frac{n}{2}}}{2^n\Gamma\left(1+\frac{n}{2}\right)}.$$

Разложение является точным при $N \to \infty$.

Здесь V_0 есть объём n–мерного шара диаметра 1. Если X — не единичный, а произвольный интервал из R^n, то нужно взять $N^* = N\frac{V_0}{V}$, где V — объём (мера) множества X.

В данном разложении 4 смещение отсутствует, первые два слагаемых относятся к компоненте разброса, последнее — «шум».

Заметим, что формула 3 не перестала быть справедливой для случайных x_j, однако первое слагаемое в ней перестало быть смещением, и таким образом формула перестала быть разложением на искомые компоненты.

Формула 5 является достаточно громоздкой, однако в некоторых частных случаях принимает простой вид, например, при $n = 1$ получаем

$$\mathsf{D}\left[\frac{1}{k}\sum_{i=1}^{k}\xi_i(x)\right] = \frac{(k+1)(k+2)}{12N^2k}.$$

В отличие от постановки 1, здесь компонента разброса растёт с ростом k, т.е. уменьшается с ростом сложности. Это очень нетипичное поведение разброса.

При $n = 2$ из 5 получаем

$$\mathsf{D}\left[\frac{1}{k}\sum_{i=1}^{k}\xi_i(x)\right] = \frac{k+1}{4\pi Nk}.$$

В итоге можем наблюдать, что для метода kNN компонента разброса демонстрирует различное поведение при разной размерности пространства: при $n = 1$ разброс практически линейно увеличивается с ростом k; при $n = 2$ он сходится к положительной константе при $k \to \infty$; при $n > 2$ разброс стремится к 0 при $k \to \infty$.

Подобное поведение разложения на смещение и разброс является нежелательным его свойством при использовании для объяснения структуры ошибок обучения.

В связи с этим представляется целесообразным использовать разложение ошибки на погрешность аппроксимации и статистическую погрешность, что было предложено авторами статьи [3]. Компоненты этого разложения всегда монотонны.

Работа выполнена в рамках госзадания Института математики им. С.Л. Соболева СО РАН, проект FWNF-2022-0015, и при частичной поддержке РФФИ, грант No. 19–29–01175.

[1] *T. Hastie, R. Tibshirani, H. Friedman, Jerome.* The Elements of Statistical Learning, 2009.

[2] *V. Nedel'ko.* On decompositions of decision function quality measure // Bulletin of Irkutsk State University. Series Mathematic. No 33. 2020. — p. 64–79.

[3] *Лбов Г.С., Старцева Н.Г.* Сложность распределений в задачах классификации // Докл. РАН. 1994. Т. 338. No. 5. С. 592–594.

Some Properties of Bias-Variance Decomposition for kNN Classifier

Nedel'ko Victor ⋆nedelko@math.nsc.ru
Novosibirsk, Institute of mathematics SB RAS

When choosing the optimal complexity of the method for constructing decision functions, an important tool is the decomposition of the quality criterion into bias and variance.

In this paper, we obtain an expression for the variance component for the kNN method for the linear regression problem in the formulation when the "explanatory" features are random variables. In contrast to the well-known result obtained for non-random "explanatory" variables, in the considered case, the variance may increase with the growth of k.

Let X be the space of values of variables used for forecasting, and Y be the set of values of the predicted variable.

All variables are random variables with some joint distribution function.

Decision function is a mapping $f : X \to Y$.

The decision function is constructed based on some training sample of size N

$$S_N = \left((x^\omega, y^\omega), \omega = \overline{1, N}\right).$$

For the decision function as a whole, the quality criterion will be MSE, i.e.

$$R(f(\cdot)) = \mathsf{E}_{x,y}(y - f(x))^2.$$

By this criterion, the optimal solution will be a regression function, i.e. a conditional mathematical expectation.

In the classical statement of regression ploblem, the values of X are not random. Only the target variable is random, which is represented as

$$y(x) = \hat{f}(x) + \delta, \tag{1}$$

where $\hat{f}(x)$ is some unknown function, and δ is a random variable with zero mean and variance σ^2.

For arbitrary independent random variables u and v (if the corresponding moments exist), the identity holds

$$\mathsf{E}(u - v)^2 = \mathsf{D}u + (\mathsf{E}u - \mathsf{E}v)^2 + \mathsf{D}v,$$

where D denotes variance, i.e. $\mathsf{D}u \equiv \mathsf{E}u^2 - (\mathsf{E}u)^2$.

Let's fix a point x of the feature space and substitute $u = y \,|\, x$, $v = f(x)$. Since $f(x)$ is constructed on a random sample, v is a random variable. Then we get

$$\mathsf{E}_{S_N, y \,|\, x}(y - f(x))^2 =$$

$$\mathsf{D}_{y\,|\,x}y + (\mathsf{E}_{y\,|\,x}y - \mathsf{E}_{S_N}f(x))^2 + \mathsf{D}_{S_N}f(x). \tag{2}$$

The notation $\mathsf{E}_{S_N,y\,|\,x}$ means that the expectation is taken over all samples of size N and over the conditional distribution on the target variable y at the point x. So, a subscript at operators E or D indicates the domain for averaging.

We obtain that 2 in this formulation is the decomposition of MSE into "noise", bias and variance.

Note that this decomposition is done for each point x. If necessary, 2 can be additionally averaged over X.

A number of sources (e.g. [1]) provide the following decomposition formula for the kNN method

$$\mathsf{E}_{S_N,y\,|\,x}(y - f(x))^2 = \left(f(x) - \frac{1}{k}\sum_{i=1}^{k}\hat{f}(\xi_i(x))\right)^2 + \frac{\sigma^2}{k} + \sigma^2, \tag{3}$$

where $\xi_i(x)$ is the coordinates of the i–th "neighbor" of a point x.

The second term in this decomposition is proposed to be interpreted as a variance.

This expression is obtained for the case when the coordinates of x^ω in the training sample are fixed, i.e. for the statement 1.

Note that the conventional definition of the decomposition components assumes complete averaging over random samples.

The variance component in 3 decreases monotonically with the growth of k, i.e. it increases with increasing complexity, since the complexity characteristic for kNN is opposite to k and can be, for example, $\frac{1}{k}$.

Consider now "explanatory" features to be random. Let $X = [0,1]^n$ and $y = x_1 + \delta$, where $x = (x_1, ..., x_n) \in X$.

We consider the model $\hat{f}(x) = x_1$ as a linear regression model without loss of generality because any linear model may be converted to it via proper transformation of features.

Suppose that x_j are independent random variables, $x_j \sim U(0,1)$.

Proposition. For the inner points of X, there is:

$$\mathsf{E}_{S_N,y\,|\,x}(y - f(x))^2 = \mathsf{D}\left[\frac{1}{k}\sum_{i=1}^{k}\xi_i(x)\right] + \frac{\sigma^2}{k} + \sigma^2, \tag{4}$$

where

$$\mathsf{D}\left[\frac{1}{k}\sum_{i=1}^{k}\xi_i(x)\right] \approx \frac{(N^*)^{-\frac{2}{n}}}{2n^2k^2}\sum_{m=0}^{k-1}\frac{k-m}{m!}\Gamma\left(m + \frac{2}{n}\right) \tag{5}$$

and

$$N^* = NV_0, \quad V_0 = \frac{\pi^{\frac{n}{2}}}{2^n\Gamma\left(1 + \frac{n}{2}\right)}.$$

The decomposition is asymptotically exact as $N \to \infty$.

Here V_0 is the n-dimensional volume of a Euclidean ball of diameter 1.

If X is an arbitrary finite interval from R^n, then one need to take $N^* = N\frac{V_0}{V}$, where V is the volume (measure) of X.

The first two terms in 4 are the variance, the last one is noise.

For $n = 1$ the formula 5 get

$$\mathsf{D}\left[\frac{1}{k}\sum_{i=1}^{k}\xi_i(x)\right] = \frac{(k+1)(k+2)}{12N^2k}.$$

In contrast to 1, the resulting decomposition has a monotonically increasing (close to linear growth) term in the variance component. This term provides the possibility of decreasing variance with increasing complexity.

For $n = 2$ the formula 5 get simple

$$\mathsf{D}\left[\frac{1}{k}\sum_{i=1}^{k}\xi_i(x)\right] = \frac{k+1}{4\pi Nk}.$$

We can see that the variance for kNN demonstrates different behavior depending on dimensionality. By $n = 1$ the variance increases as k increases. By $n = 2$ the variance tends to a positive constant as $k \to \infty$. By $n > 2$ the variance tends to zero.

Such undesired properties of bias-variance decomposition encourage to search alternatives. As such alternative might be considered another decomposition of the error: into a measure of adequacy and a measure of stability, that was proposed in [3]. Since the cited references are hardly accessible, one can read brief statement of the concept in [2]. The idea of the approach is to decompose the error into the approximation error and the statistical error. The components of this decomposition are obviously monotonic.

The study was carried out within the framework of the state contract of the Sobolev Institute of Mathematics (project no FWNF-2022-0015) and with partial support by RFBF grant 19–29–01175.

[1] *T. Hastie, R. Tibshirani, H. Friedman, Jerome.* The Elements of Statistical Learning, 2009.

[2] *V. Nedel'ko.* On decompositions of decision function quality measure // Bulletin of Irkutsk State University. Series Mathematic. No 33. 2020. — p. 64–79.

[3] *G.S. Lbov, N.G. Startseva.* Complexity of distributions in classification problems // Russ. Acad. Sci., Dokl., Math. 50 (2) (1994)

Применение нейросетевого многомерного шкалирования для построения векторных представлений разнородных данных

Колосов Алексей Михайлович[1]⋆ akolosov@cs.msu.ru
Майсурадзе Арчил Ивериевич[1] maysuradze@cs.msu.ru
[1]Москва, МГУ им. М.В.Ломоносова

Векторные представления объектов активно используются в задачах, связанных с визуализацией разнородной информации. Предложенный в работе [1] метод позволяет по матрице порядков близостей, заданной на парах объектов, решить задачу обобщённого неметрического многомерного шкалирования, Generalized Non-metric Multidimensional Scaling, GNMDS для получения векторных представлений объектов, сохраняющих заданные порядки близостей (1).

$$d(i,\ j) < d(k,\ l) => e(g(i),\ g(j)) < e(g(k),\ g(l)) \tag{1}$$

где $d(i, j)$ — исходные близости, $e(g(i),\ g(j))$ — некоторая функция расстояния, $g(i)$ — искомое векторное представление объекта.

Предложенный метод реализован в виде библиотеки на языке python [2] и апробирован в настоящей работе на следующих типах данных:

1. *Синтетические данные*
   Для синтетических данных рассмотрен пример, когда данные представлены точками на прямой. Сложность нахождения решения для такого примера заключается в сохранении размерности, равной единице, с полным сохранением порядков близостей. Метод справляется с этой задачей, находя подходящее расположение точек на прямой, без увеличения размерности пространства. Стоит отметить, что корректным решением также является инверсия порядка расположения точек на прямой.

2. *Данные, содержащих сведения о семантической близости в парах слов*
   Для построения векторных представлений слов использованы экспертные данные, содержащие сведения о семантической близости в парах слов: SimLex-999 [3], WordSim353 [4], MEH [5]. Для размерности 1 и евклидовой функции расстояния качество применения метода, определямое как ранговая корреляция между величинами экспертных близостей и расстояний между векторными представлениями слов в парах слов, составило: 0.959 для SimLex-999, 0.962 для WordSim-353 и 0.741 для MEN. Качество для размерности 2 и евклидовой функции расстояния составило: 0.984 для SimLex-999, 0.976 для WordSim-353 и 0.921 для MEN. Стоит отметить, что показатель качества растёт с увеличением размерности целевых представлений, однако более эффективным для прикладных применений являются векторные представления наименьшей допустимой размерности (см. Рис. 1).

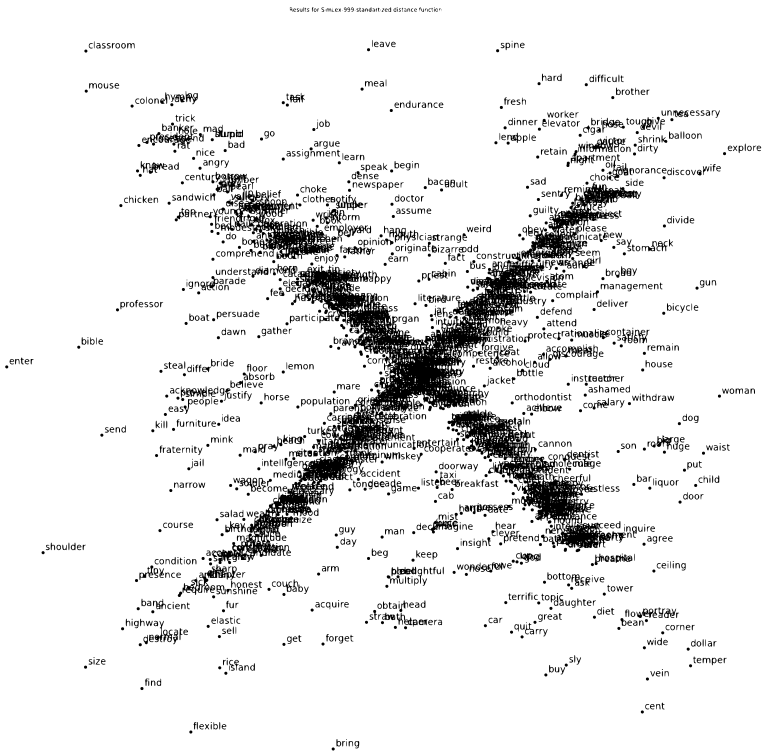


Рис. 1. Визуализация векторных представлений слов для набора SimLex-999

3. *Данные, содержащие сведения о группе пользователей социальной сети*
   Для группы пользователей социальной сети были собраны сведения о друзьях каждого пользователя группы внутри этой группы. Затем для каждой пары пользователей была рассчитана мере сходства, определяемая как отношения мощности пересечения к мощности объединения множеств друзей этих двух пользователей, также известная как мера Жаккара. Применение метода для полученной матрицы попарного сходства пользователей позволило получить векторные представления группы пользователей социальной сети (см. Рис. 2).

Актуальным для дальнейшего исследования является вопрос о расположении полученных векторных представлений вдоль направляющих линий. Также вопросом для дальнейшего исследования является определение условий, при выполнении которых используемый метод будет получать точное решение.

Рис. 2. Визуализация векторных представлений группы пользователей социальной сети

Работа выполнена при поддержке НОШ МГУ «Мозг, когнитивные системы, искусственный интеллект», НИР МГУ 5.1.21, гранта РФФИ No. 20-01-00664.

[1] *Maysuradze, A., Kolosov, A.* Neural network method for solving the problem of generalized non-metric multidimensional scaling // Abstracts of the 20th Russian National Conference “Mathematical Methods for Pattern Recognition”, 111–116.

[2] *Kolosov, A.* obj2vec // github.com/obj2vec/gnmds

[3] *Felix Hill et al* SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation // Computational Linguistics. 2015.

[4] *Lev Finkelstein et al* Placing Search in Context: The Concept Revisited // ACM Transactions on Information Systems, 20(1):116-131, 2002.

[5] *E. Bruni et al* Multimodal Distributional Semantics // Journal of Artificial Intelligence Research 49: 1-47.

Application of neural network multidimensional scaling for constructing vector representations of heterogeneous data

Maysuradze Archil[1] maysuradze@cs.msu.ru
Kolosov Alexey[1]⋆ akolosov@cs.msu.ru
[1]Moscow, Lomonosov Moscow State University

Vector representations of objects are actively used in tasks related to the visualization of heterogeneous data. The method proposed in the work [1] allows solving the problem of generalized non-metric multidimensional scaling, GNMDS, using the matrix of similarity orders given on pairs of objects to obtain vector representations of objects that preserve the given similarity orders (1).

$$d(i,\ j) < d(k,\ l) => e(g(i),\ g(j)) < e(g(k),\ g(l)) \tag{1}$$

where $d(i, j)$ are the original similarity, $e(g(i),\ g(j))$ is some distance function, $g(i)$ is the desired vector representation of the object.

The proposed method is implemented as a library in the python language [2] and tested in this work on the following data types:

1. *Synthetic data*
   For synthetic data, an example is considered when the data is represented by points on a straight line. The complexity of finding a solution for such an example lies in keeping the dimension equal to one, with full preservation of the similarity orders. The method copes with this task by finding a suitable location of points on a straight line, without increasing the dimension of space. It should be noted that the correct solution is also the inversion of the order of points on the line.

2. *Data containing information about semantic similarity in word pairs*
   To build vector representations of words, expert data were used that contain information about semantic similarity in pairs of words: SimLex-999 [3], WordSim353 [4], MEH [5]. For dimension 1 and the Euclidean distance function, the quality of the method application, defined as the rank correlation between the values of expert similarity and distances between vector representations of words in word pairs, was: 0.959 for SimLex-999, 0.962 for WordSim-353 and 0.741 for MEN. The quality for dimension 2 and the Euclidean distance function was: 0.984 for SimLex-999, 0.976 for WordSim-353 and 0.921 for MEN. It should be noted that the quality indicator grows with the increase in the dimension of the target representations, however, vector representations of the smallest allowable dimension are more efficient for applied applications (see Fig. 1).

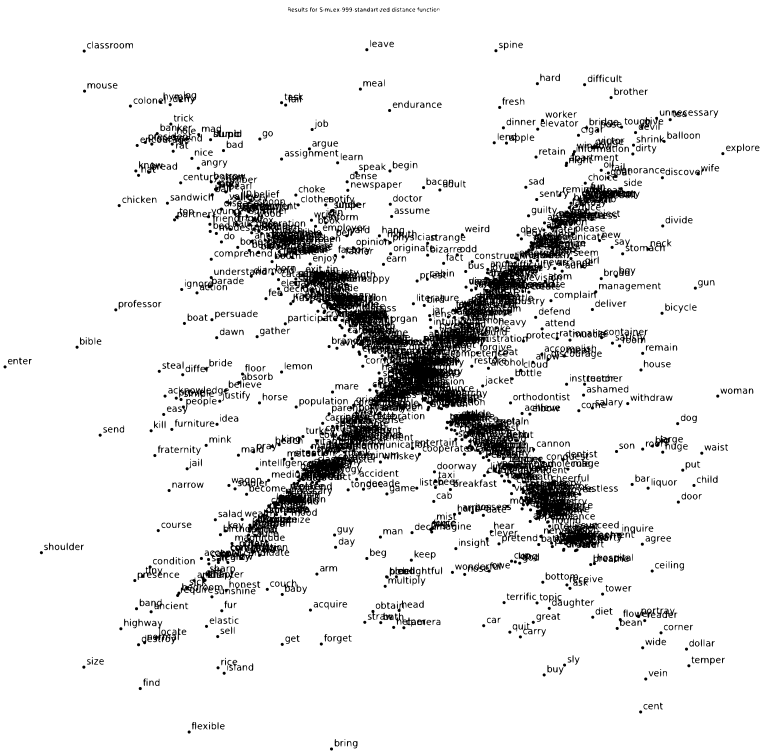


Fig. 1. Visualization of vector representations of words for the SimLex-999

3. *Data containing information about a group of social network users*
   For a group of social network users, information was collected about the friends of each user of the group within this group. Then, for each pair of users, a measure of similarity was calculated, defined as the ratio of the power of the intersection to the power of the union of the sets of friends of these two users, also known as the Jaccard measure. Application of the method for the resulting matrix of pairwise similarity of users made it possible to obtain vector representations of a group of users of a social network (see Fig. 2).

Relevant for further research is the question of the location of the obtained vector representations along the guide lines. It is also a question for further research to determine the conditions under which the method used will obtain an exact solution.

Fig. 2. Visualization of vector representations for a group of social network users

This work was supported by the NOSH MSU "Brain, cognitive systems, artificial intelligence", Research and Development MSU 5.1.21, RFBR grant No. 20-01-00664.

[1] *Maysuradze, A., Kolosov, A.* Neural network method for solving the problem of generalized non-metric multidimensional scaling // Abstracts of the 20th Russian National Conference "Mathematical Methods for Pattern Recognition", 111–116.

[2] *Kolosov, A.* obj2vec // github.com/obj2vec/gnmds

[3] *Felix Hill et al* SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation // Computational Linguistics. 2015

[4] *Lev Finkelstein et al* Placing Search in Context: The Concept Revisited // ACM Transactions on Information Systems, 20(1):116-131, 2002.

[5] *E. Bruni et al* Multimodal Distributional Semantics // Journal of Artificial Intelligence Research 49: 1-47.

Поиск согласованных нейросетевых моделей в задаче мультидоменного обучения

Яковлев Константин Дмитриевич[1]⋆ iakovlev.kd@phystech.edu
Бахтеев Олег Юрьевич[1,2] bakhteev@phystech.edu
Стрижов Вадим Викторович[1,2] strijov@phystech.edu
[1]Москва, Московский физико-технический институт
[2]Москва, Вычислительный центр им. А.А. Дородницына ФИЦ ИУ РАН

В работе исследуется проблема выбора структуры модели глубокого обучения для мультидоменных данных. Модель представляет собой ориентированный ациклический граф, в котором ребра являются нелинейными функциями, дифференцируемыми по параметрам [1]. Структура модели определяется набором векторов, задающих категориальное распределение на ребрах графа. Предлагается рассматривать задачу как задачу мультимоделирования: для каждого домена оптимизируется отдельная структура. Рассматриваются два метода регуляризации: структурная и регуляризация пространства скрытых представлений модели. Структурная регуляризация основана на приближении вероятностных распределений, задающих структуру. Регуляризация пространства скрытых представлений модели основана на приближении скрытых представлений схожих объектов из разных доменов, получаемых под действием нелинейных функций. В работе показано, что предложенная оптимизационная задача позволяет найти компромисс между сложностью итоговой модели и ее прогностическими характеристиками.

Базовый эксперимент проводится на выборке MNIST. Рассматриваются модели, обученные на своем домене, модель, обученная на всех доменах, а также предложенная мультимодель, обученная на всех доменах с предлагаемыми методами регуляризации. Показано, что предложенная модель обладает наибольшей обобщающей способностью.

[1] Yakovlev, K., Grebenkova, O., Bakhteev, O. & Strijov, V. Neural Architecture Search with Structure Complexity Control. *Recent Trends In Analysis Of Images, Social Networks And Texts.* pp. 207-219 (2022)

Concordant neural architecture search on multi-domain data

Yakovlev Konstantin[1]⋆ iakovlev.kd@phystech.edu
Bakhteev Oleg[1,2] bakhteev@phystech.edu
Strijov Vadim[1,2] strijov@phystech.edu

[1]Moscow, Moscow Institute of Physics and Technology
[2]Moscow, FRCCSC of the Russian Academy of Sciences

The paper investigates the problem of selection the structure of a deep learning model for multi-domain data. The model is a directed acyclic graph in which the edges are nonlinear functions differentiable by parameters [1]. The structure of the model is determined by a set of vectors that define a categorical distribution on the edges of the graph. It is proposed to consider the problem as a multimodeling problem: a separate structure is optimized for each domain. Two regularization methods are considered: structural and regularization of the space of hidden representations of the model. Structural regularization is based on the approximation of probability distributions, which defines the structure. Regularization of the space of hidden representations of the model is based on the approximation of hidden representations obtained under the action of nonlinear functions. The paper shows that the proposed optimization problem makes it possible to find a compromise between the complexity of the final model and its performance.

The basic experiment is performed on a MNIST dataset. Considered models trained on its domain, a model trained on all domains, as well as a proposed multimodel trained on all domains with proposed regularizations. It is shown that the proposed model has the best performance.

[1] Yakovlev, K., Grebenkova, O., Bakhteev, O. & Strijov, V. Neural Architecture Search with Structure Complexity Control. *Recent Trends In Analysis Of Images, Social Networks And Texts*. pp. 207-219 (2022)

Дистилляция моделей глубокого обучения на многодоменных выборках

Баязитов Камил Маратович[1] baiazitov.km@phystech.edu
Грабовой Андрей Валериевич[1,2]⋆ grabovoy.av@phystech.edu
Стрижов Вадим Викторович[1,3] strijov@phystech.edu

[1]Москва, Московский физико-технический институт (национальный исследовательский университет)
[2]Москва, Антиплагиат
[3]Москва, Федеральный исследовательский центр «Информатика и управление» Российской академии наук (ФИЦ ИУ РАН)

Исследуется проблема снижения сложности аппроксимирующих моделей машинного обучения с учетом многодоменности данных в выборке. Для решения задачи снижения сложности используются методы, которые основаны на дистилляции моделей глубокого обучения.

Предлагается, что задана предобученная модель на выборке большей мощности — называемая учителем. Требуется обучить модель на выборке меньшей мощности — называемую учеником. Предполагается, что большая и меньшая выборка принадлежат различным доменам. В свою очередь объекты из рассматриваемых доменов должны принадлежать близким генеральным совокупностям. Генеральная совокупность объектов B называется близкой к совокупности A, если существует инъективное отображение $\varphi : A \to B$. Заметим, что в общем случае не накладывается никакое ограничение на гладкость функции φ, что позволяет использовать широкий класс функций.

В ходе экспериментов проведен анализ качества предложенного метода на реальных и синтетических данных. В качестве реальных данных рассматривались выборки MNIST, FashionMNIST и ImageNet. В эксперименте показано, что качество модели ученика повышается при использовании ответов учителя при обучении на многодоменных выборках.

[1] *Грабовой А. В., Стрижов В. В.* Байесовская дистилляция моделей глубокого обучения // Автоматика и телемеханика, 2021.

[2] *Грабовой А. В., Стрижов В. В.* Вероятностная интерпретация задачи дистилляции // Автоматика и телемеханика, 2022.

Multi-Domain Distillation of Deep Learning Model

Bayazitov Kamil[1] baiazitov.km@phystech.edu
Grabovoy Andrey[1,2]⋆ grabovoy.av@phystech.edu
Strijov Vadim[1,3] strijov@phystech.edu
[1]Moscow, Moscow Institute of Physics and Technology
[2]Moscow, Antiplagiat
[3]Moscow, Federal Research Center "Informatics and management" Russian Academy of Sciences

This research studies the approximating machine learning models' complexity reduction problem for multi-domain datasets. We use knowledge distillation methods to solve the problem of reducing the complexity of deep learning models.

It is assumed that a pre-trained model on a large dataset is given. We called this model a teacher. It is required to train the model on a smaller dataset. We called this model a student. We assumed that the larger and smaller datasets belong to different domains. But objects from domains must be similar enough. Objects from domain B are similar enough to the domain A if there exists an injective mapping $\varphi : A \to B$.

The computational experiment is carried out on synthetic data, as well as on MNIST, FashionMNIST, and ImageNet real datasets. The experiment shows that the quality of the student model increases when we were using the teachers trained in a similar enough domain.

[1] *Grabovoy A. V., Strijov V. V.* Bayesian Distilation of Deep Learning Model // Automation and Remote Control, 2021.

[2] *Grabovoy A. V., Strijov V. V.* Probabilistic Interpretation of the Distillation Problem // Automation and Remote Control, 2021.

Интеллектуализация анализа выполнения запросов в колоночной СУБД

Рябцев Антон Борисович[1]⋆ ryabtsev.ab@phystech.edu
Дулин Сергей Константинович[2] skdulin@mail.ru
[1]Москва, МФТИ (НИУ)
[2]Москва, ФИЦ ИУ РАН

В данной работе рассматривается задача оптимизации планов выполнения SQL запросов [1] с помощью машинного обучения. В работе подробно описан традиционный подход к решению данной задачи (рисунок 1), рассмотрены его недостатки [2]. Также приведён анализ существующих методов на основе ма-

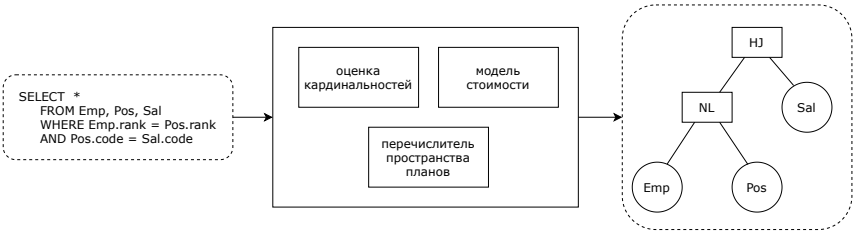


Рис. 1. Архитектура традиционного оптимизатора запросов.

шинного обучения, которые призваны устранить ряд недостатков традиционного оптимизатора [3, 4, 5, 6, 7].

Сформулированы рекомендации к выбору метода в зависимости от особенностей СУБД и от доступных вычислительных ресурсов. Показана актуальность проблемы для массово-параллельных колоночных СУБД и отмечено отсутствие исследования возможностей применения известных подходов на основе машинного обучения к оптимизации данного типа СУБД.

	Время выполнения 100 запросов.			
	Табличный режим СУБД		**Колоночный режим СУБД**	
	Традиционный оптимизатор	**“Умный” оптимизатор**	**Традиционный оптимизатор**	**“Умный” оптимизатор**
Вариант из оригинальной статьи.	**2.5 часа**	**5 часов**	**1.5 часа**	**3 часа**
Вариант с рядом модификаций.		**1.5 часа**		**2 часа**

Таблица 1. Анализ результатов подхода Neo

В рамках исследования применимости к колоночной СУБД подходы на основе оценок кардинальности показали себя плохо – одни не приводили к улучшениям, другие оказались слишком "тяжеловесны" для использования в СУБД [8]. Подходы на основе аппроксимации функции стоимости также показали себя не лучшим образом. Для достижения положительных результатов (таблица 1) с их использованием потребовался ряд модификаций [9, 10, 11].

Основные результаты работы:

— Произведена оценка методов предсказания кардинальности (NeuroCard [4], PostgreSQL AQO) и методов аппроксимации функции стоимости (DQN, Neo [7]).
— Экспериментально показано, что левые-глубокие планы проигрывают ветвистым.
— Предложен новый способ получения целевых значений для обучения моделей – вместо времени выполнения запроса используется его фактическая стоимость (вычисленная с использованием фактических кардинальностей), что делает возможным обучение моделей в продуктовых системах.
— Предложена модификация архитектуры системы Neo – рисунок 2.

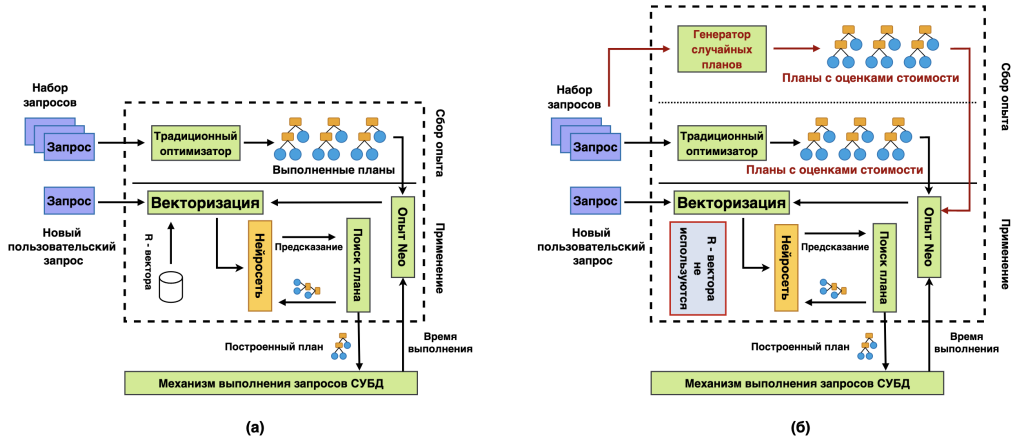


Рис. 2.
(а) Дизайн системы Neo в оригинале.
(б) Модифицированный дизайн системы Neo – модификации подсвечены бордовым цветом.

— Признаковые описания были модифицированы, так как планы в массово-параллельных колоночных СУБД синтаксически отличаются от планов в централизованных СУБД.

— Предложен способ регуляризации для предотвращения переобучения, так как значения селективностей базовых отношений при перезапуске исследованной СУБД флуктуируют, а эксперименты показали, что модель переобучается на конкретных значениях.
— Предложена модификация архитектуры нейросети Neo (transformer [10] вместо Dynamic Pooling).
— Показана польза от получения оценок неопределённости предсказаний [12, 13].

Эти результаты представляются достаточно интересными как с практической, так и с теоретической точки зрения.

[1] *Зильбершац, Авраам и Корт, Генри Ф. и Сударшан, Шашанк* Концепции систем баз данных // McGraw-Hill New York, 2002.

[2] *Лейс* Насколько хороши оптимизаторы запросов на самом деле? // Труды фонда VLDB, 2015. — С. 204–215.

[3] *Хармуш, Хазар и Науманн, Феликс* Оценка кардинальности: экспериментальный обзор // Труды фонда VLDB, Фонд VLDB, 2017. — С. 499–512.

[4] *Янг* NeuroCard: один оценщик кардинальностей для всех таблицы // Труды фонда VLDB, 2020. — С. 61–73.

[5] *Кай, Балазинска, Сучиу, Дэн* Пессимистическая оценка кардинальности: более жесткие верхние границы для кардинальностей промежуточных соединений // Труды международной конференции по управлению данными 2019 г., 2019. — С. 18–35.

[6] *Кифер, Хеймель, Бресс, Маркл, Фолькер* Оценка избирательности соединения с использованием моделей плотности ядра с оптимизированной пропускной способностью // Труды фонда VLDB, Фонд VLDB, 2017. — С. 2085–2096.

[7] *Райан Маркус* Neo: обученный оптимизатор запросов // Труды фонда VLDB, 2021.

[8] *Дулин С.К., Розенберг И.Н., Уманский В.И.* О проблеме интеграции информационных ресурсов // Системы и средства информатики, 2019. — С. 127–138.

[9] *Сюй, Цзинцзин* Понимание и улучшение нормализации слоя // NIPS, 2019.

[10] *Васвани* Внимание – это всё, что Вам нужно // NIPS, 2017.

[11] *Хасселът, Хадо* Двойное Q-обучение // NIPS, 2010. — С. 204–215.

[12] *Галь, Ярин, Гахрамани, Зубин* Дропаут как байесовское приближение: представление неопределенности модели в глубоком обучении // ICML, PMLR, 2016. — С. 1050–1059.

[13] *Сривастава, Нитиш, Хинтон, Суцкевер* Дропаут: простой способ предотвратить переобучение нейронных сетей // The journal of machine learning research, JMLR.org, 2014. — p. 1929–1958.

[14] *Эйке Хюллермайер, Виллем Вагеман* Алеаторическая и эпистемическая неопределенность в машинном обучении: введение в концепции и методы // Machine Learning, Springer, 2021. — p. 457–506.

Intellectualization of query execution analysis in a columnar DBMS

Ryabtsev Anton[1]⋆ ryabtsev.ab@phystech.edu
Dulin Sergey[2] skdulin@mail.ru
[1]Moscow, MIPT
[2]Moscow, FRCCSC

In this paper, we consider the problem of optimizing SQL query execution plans [1] using machine learning. The paper describes in detail the traditional approach to solving this problem. (picture 1), it's shortcomings are considered [2]. It

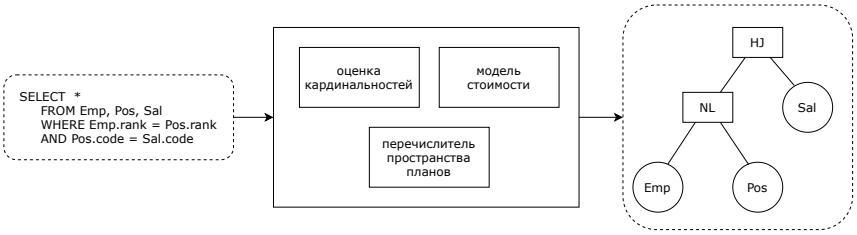


Fig. 1. Traditional query optimizer architecture.

also provides an analysis of existing methods based on machine learning, which are designed to eliminate a number of shortcomings of the traditional optimizer[3, 4, 5, 6, 7].

Recommendations are formulated for choosing a method depending on the features of the DBMS and on the available computing resources. The relevance of the problem for massively parallel columnar DBMS is shown and the lack of research into the possibilities of using known approaches based on machine learning to optimize this type of DBMS is noted.

As part of the study of applicability to a columnar DBMS, approaches based on cardinality estimates performed poorly - some did not lead to improvements, others turned out to be too "heavy" for use in a DBMS [8]. Approaches based on the approximation of the cost function also proved to be not the best. To achieve positive results (table 1) with their use, a number of modifications were required. [9, 10, 11].

The main results of the work:

— Evaluation of cardinality estimation methods (NeuroCard [4], PostgreSQL AQO) d cost function approximation methods (DQN, Neo [7]) was made.
— It has been experimentally shown that left-deep plans lose out to bushy ones.
— A new way to obtain target values for model training is proposed - instead of the query execution time, its actual cost (calculated using actual cardinalities) is used, which makes it possible to train models in product systems.

	Execution time of 100 queries.			
	Table mode of DBMS		Columnar mode of DBMS	
	Traditional optimizer	"Smart" optimizer	Traditional optimizer	"Smart" optimizer
Original paper approach.	2.5 hours	5 hours	1.5 hours	3 hours
Approach with modifications.		1.5 hours		2 hours

Table 1. Analysis of the results of the Neo approach

— A modification of the architecture of the Neo system is proposed – picture 2.

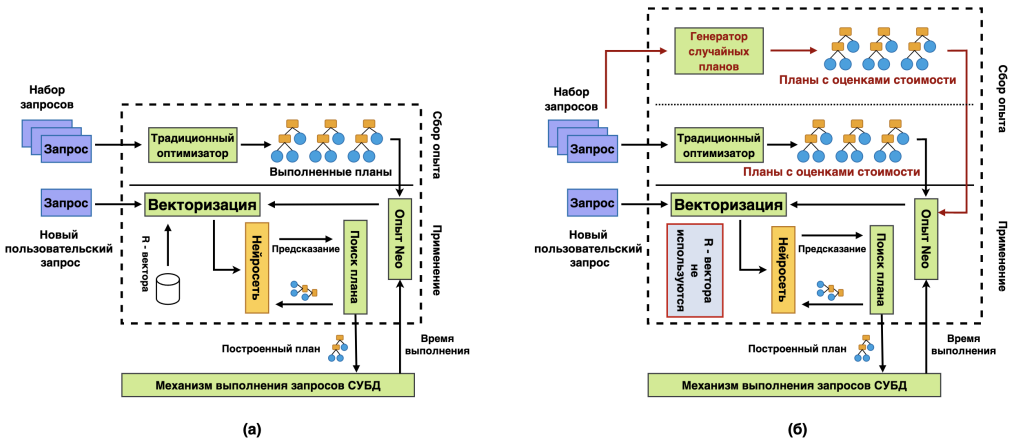


Fig. 2.
(a) The design of the Neo system from the original paper.
(b) The modified design of the Neo system - modifications are highlighted in burgundy.

— Feature descriptions have been modified, since plans in massively parallel columnar DBMSs are syntactically different from plans in centralized DBMSs.
— A regularization method is proposed to prevent overfitting, since the selectivity values of the base ratios fluctuate when the studied DBMS is restarted, and experiments have shown that the model overfits on specific values.
— A modification of the Neo neural network architecture is proposed(transformer [10] instead of Dynamic Pooling).

— The benefits of obtaining estimates of the uncertainty of predictions are shown [12, 13].

These results are quite interesting from both practical and theoretical points of view.

[1] *Zilbershatz, Avraham & Kort, Henry F. & Sudarshan, Shashank* Database system concepts // McGraw-Hill New York, 2002.

[2] *Leis* How good are query optimizers, really? // Proceedings of VLDB Endowment, 2015. — p. 204–215.

[3] *Kharmush, Khazar and Naumann, Felix* Cardinality estimation: experimental survey // Proceedings of VLDB Endowment, VLDB Endowment, 2017. — p. 499–512.

[4] *Yang* NeuroCard: one cardinality estimator for all tables // Proceedings of VLDB Endowment, 2020. — p. 61–73.

[5] *Kai, Balazinska, Suchiu, Dan* Pessimistic cardinality estimation: Tighter upper bounds for intermediate join cardinalities // Proceedings of the International Conference on Data Management 2019, 2019. — p. 18–35.

[6] *Kiefer, Heimel, Bress, Markle, Volker* Estimating join selectivities using bandwidth-optimized kernel density models // Proceedings of VLDB Endowment, VLDB Endowment, 2017. — p. 2085–2096.

[7] *Ryan Marcus* Neo: A learned query optimizer // Proceedings of VLDB Endowment, 2021.

[8] *Dulin, Rosenberg, Umanskiy.* On the problem of integration of information resources // Systems and means of informatics, 2019. — p. 127–138.

[9] *Xu, Jingjing* Understanding and improving layer normalization // NIPS, 2019.

[10] *Vaswani* Attention is all you need // NIPS, 2017.

[11] *Hasselt, Hado* Double Q-learning // NIPS, 2010. — p. 204–215.

[12] *Gal, Yarin, Gahramani, Zubin* Dropout as a bayesian approximation: Representing model uncertainty in deep learning // ICML, PMLR, 2016. — p. 1050–1059.

[13] *Srivastava, Nitish, Hinton, Sutzkever* Dropout: a simple way to prevent neural networks from overfitting // The journal of machine learning research, JMLR.org, 2014. — p. 1929–1958.

[14] *Eike Hüllermeier, Willem Wagemann* Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods // Machine Learning, Springer, 2021. — p. 457–506.

Обеспечение синхронизации в системах совместной разработки ML-решений

Решетков Андрей Эдуардович[1]⋆ reshetkov.aeh@phystech.edu
Хританков Антон Сергеевич[1] anton.khritankov@phystech.edu
[1]Долгопрудный, Московский физико-технический институт

Воспроизводимость экспериментов - одно из ключевых требований к современным исследованиям в области машинного обучения. В условиях, когда проектирование, отладка и эксперименты в области машинного обучения ведутся большими командами инженеров и вычислительных агентов, возникает потребность в методах организации совместной работы, в программных средствах распределённой разработки и исследований. Благодаря им инженеры и исследователи данных могли бы заниматься своими задачами параллельно и независимо друг от друга. Проблема заключается в объединении результатов работы нескольких исследователей и вычислительных агентов в одну общую работоспособную версию с воспроизводимыми результатами.

Данная проблема и ранее возникала в разных сферах. Например, в области разработки программного обеспечения она решается внедрением систем контроля версий и управления конфигурацией, таких как git или svn. Однако в сфере машинного обучения применение стандартных подходов недостаточно, поскольку объектом совместной работы является не набор строк исходного кода или версий документов, а совокупность постановки задачи, исследовательских гипотез, процедур проведения, критериев успешности, измеряемых параметров, собранных и исходных данных, результатов анализа, отчетов и исходного кода - всего, что входит в описание эксперимента [1].

Описание эксперимента представимо в виде семантической сети или графа знаний [2], в то время как исполняемая процедура проведения эксперимента может быть представлена в виде графа потока работ (pipeline), в котором узлы сопоставлены исполняемым наборам программ или процессов, а ребра - обменам данными между ними. Рассмотрим пример вычислительного эксперимента и процедуры его выполнения в виде графа потока работ на рисунке 1. После старта вершина *Prepare train data* готовит обучающую выборку и передаёт её двум другим вершинам *Train Lin. Model Sgd* и *Train Lin. Model analyt.* Они параллельно обучают модели линейной регрессии: первая ищет решение с использованием стохастического градиентного спуска, а вторая аналитически. Дальше они передают обученные модели в следующую вершину *Compare models on test data*. Эта вершина также принимает на вход тестовый датасет из *Prepare test data*, а затем сравнивает качество моделей на этих данных.

Таким образом, задача организации совместной работы над экспериментом заключается в поддержании согласованного формального описания эксперимента у всех участников, а также в обеспечении возможности распределенного проведения исполняемых процедур вычислительного эксперимента.

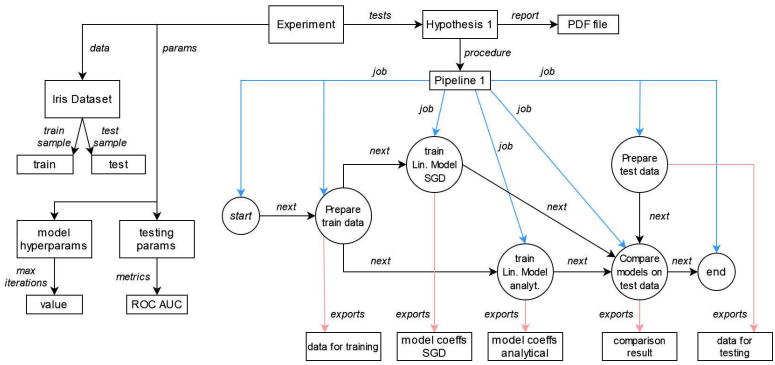


Рис. 1. Пример графа эксперимента, который сравнивает две линейные модели

Для совместной работы над графами, то есть совместным внесением изменений в ребра и вершины, уже разработан ряд техник. Например, в [4] предлагается неблокирующая и линеаризуемая структура данных, позволяющая решать задачи поиска в ширину, подсчёта срединной центральности и поиска кратчайшего расстояния из заданной вершины до всех остальных. А в [5] авторы рассматривают неблокирующий и линеаризуемый алгоритм проверки наличия пути между заданными вершинами. Тем не менее, эти техники не подходят для решения нашей задачи, поскольку нам важны не только возможность проведения расчетов свойств хранимого графа эксперимента, но и внесение изменений в его структуру.

В данной работе мы предлагаем использовать распределенные структуры данных, Conflict-Free Replicated Data Type (CRDT) [3]. Такие структуры данных не требуют от вносящих изменения агентов синхронизации их действий и в то же время гарантируют корректность. Более того, они обладают свойством сильной согласованности в конечном итоге (Strong Eventual Consistency). Согласно ему, после после объединения изменений с течением времени (в конечном итоге) все локальные версии будут иметь одинаковые версии обновленной структуры данных. При этом, если две локальные версии совпадали, то и после внесения изменений они будут совпадать.

Для построения такой распределенной структуры данных мы выбрали операционный подход (operation-based), поскольку альтернативный подход, основанный на обмене копиями данных (state-based), требует передачи избыточной для наших целей информации, и в этом случае затраты на передачу всего графа эксперимента могут оказаться слишком велики.

Для решения задачи объединения изменений в графе эксперимента G в работе мы рассмотрели три типа изменений (операций) $O = \{add, del, edit\}$ над вершинами и ребрами графа, которые агенты могут вносить:

1. **Add** – Добавление новой вершины в граф эксперимента и указанного набора инцидентных ребер;
2. **Del** – Удаление существующей вершины из графа и всех инцидентных ребер;
3. **Edit** – Изменение заданной вершины.

Мы наложили на операции ограничения, обусловленные определением CRDT и, проанализировав все 6 возможных комбинаций, показали, что в таких условиях конкурирующие изменений попарно коммутируют, то есть результат их применения не зависит от порядка. Следствием этого стало следующее.

Утверждение 1. *Построенная структура данных* $\langle G, O \rangle$ *действительно является CRDT и обладает свойством сильной согласованности в конечном итоге.*

Таким образом, в данной работе мы показали, что наш способ представления графа эксперимента и способ обмена изменениями допускает реализацию в виде CRDT. Утверждение 1 теоретически обосновывает возможность применения построенной структуры данных для решения исходной проблемы. Исследователи и вычислительные агенты могут хранить у себя локальные копии графа эксперимента и производить над ним манипуляции, а в конце внесения изменений они могут запустить алгоритм обмена данными. По его окончании каждый агент получит обновлённую версию графа эксперимента, которая содержит как изменения, внесённые другими агентами, так и модификации, созданные им самим.

[1] *Gundersen, Odd Erik & Gil, Yolanda & Aha, David.* (2018). On Reproducible AI: Towards Reproducible Research, Open Science, and Digital Scholarship in AI Publications. // AI Magazine. 39. 56-68. 10.1609/aimag.v39i3.2816.

[2] *Khritankov, A., Pershin, N., Ukhov, N., & Ukhov, A.* (2022). MLDev: Data Science Experiment Automation and Reproducibility Software. // International Conference on Data Analytics and Management in Data Intensive Domains (pp. 3–18).

[3] *Marc Shapiro, Nuno Preguiça, Carlos Baquero, Marek Zawirski.* Conflict-free Replicated Data Types. // SSS 2011 - 13th International Symposium Stabilization, Safety, and Security of Distributed Systems, Grenoble, France. pp.386–400.

[4] *Chatterjee, Bapi and Peri, Sathya and Sa, Muktikanta and Manogna, Komma.* Non-Blocking Dynamic Unbounded Graphs with Worst-Case Amortized Bounds. // 25th International Conference on Principles of Distributed Systems (OPODIS 2021), Dagstuhl, Germany. pp.20:1-20:25.

[5] *Chatterjee, Bapi and Peri, Sathya and Sa, Muktikanta and Singhal, Nandini.* A Simple and Practical Concurrent Non-Blocking Unbounded Graph with Linearizable Reachability Queries. // Proceedings of the 20th International Conference on Distributed Computing and Networking, Bangalore, India. pp.168–177.

Ensuring synchronization in systems of collaborative ML-development

Reshetkov Andrew[1]⋆ reshetkov.aeh@phystech.edu
Khritankov Anton[1] anton.khritankov@phystech.edu
[1]Dolgoprudny, Moscow Institute of Physics and Technology

Experiments reproducibility is one of the key requirements for modern research in the field of machine learning. In conditions when design, debugging and experiments are carried out by large teams of engineers and computing agents, there is a need for methods of organizing collaboration, software tools for distributed development and research. Thanks to them, engineers and data researchers could do their tasks in parallel and independently of each other. The problem is to combine the results of the work of several researchers and computational agents into one common workable version with reproducible results.

This problem has previously arisen in various fields. For example, in the field of software development, it is solved by implementing version control and configuration management systems, such as git or svn. However, in the field of machine learning, the use of standard approaches is not enough, since the object of collaboration is not a set of source code lines or document versions, but a set of problem statements, research hypotheses, procedures, success criteria, measured parameters, collected and source data, analysis results, reports and source code - everything that is included in the description experiment [1].

The description of the experiment is represented in the form of a semantic network or a knowledge graph [2], while the executable procedure of the experiment can be represented in the form of a workflow graph (pipeline), in which nodes are mapped to executable sets of programs or processes, and edges are data exchanges between them.

Let's consider an example of a computational experiment and the procedure for its execution in the form of a workflow graph in Figure 1. After the start, the node *Prepare train data* prepares a training sample and passes it to two other nodes *Train Lin. Model SGD* and *TrainLine. Model analyt.* They train linear regression models in parallel: the first one searches for a solution using stochastic gradient descent, and the second one analytically. Then they pass the trained models to the next node *Compare models on test data.* This node also accepts a test dataset from *Prepare test data* as input, and then compares the quality of models on this data.

Thus, the task of organizing joint work on the experiment is to maintain a consistent formal description of the experiment for all participants, and to ensure the possibility of distributed execution for all executable procedures of a computational experiment.

A number of techniques have already been developed for joint work on graphs, i.e. making changes to edges and nodes together. For example, [4] offers a non-blocking and linearizable data structure that allows solving the problems of breadth-first

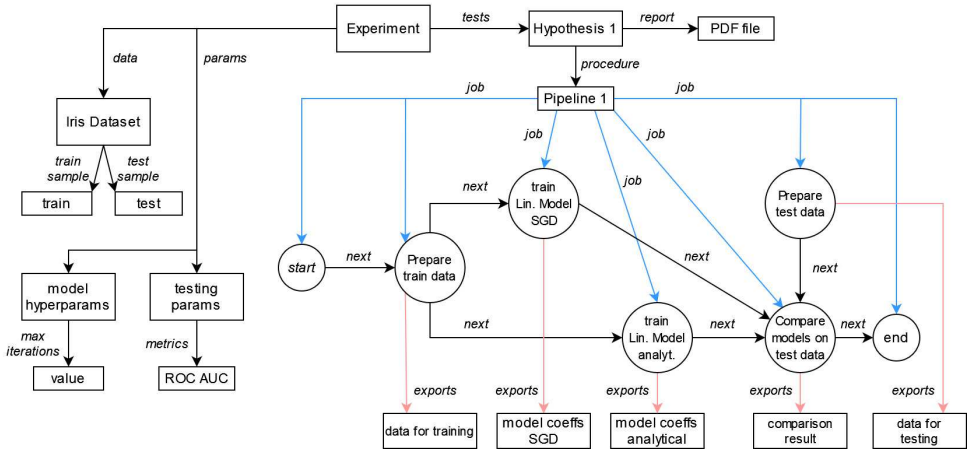


Fig. 1. Example of a graph of the experiment which compares two linear models

search, calculating median centrality, and finding the shortest distance from a given vertex to all others. In [5], the authors consider a non-blocking and linearizable algorithm for checking the presence of a path between specified vertices. However, these techniques are not suitable for solving our problem, since it is essential for us not only to be able to calculate the properties of the stored graph of the experiment, but also to make changes to its structure.

In this paper, we propose to use distributed data structures, Conflict-Free Replicated Data Type (CRDT) [3]. Such data structures do not require the modifying agents to synchronize their actions and at the same time guarantee correctness. Moreover, they have the property of Strong Eventual Consistency. According to it, after applying changes all local versions will eventually have the same states of the updated data structure. At the same time, if two local versions coincided, then after applying changes they will also coincide.

To build such a distributed data structure, we chose an operation-based approach, since an alternative approach based on the exchange of copies of data (state-based) requires the transfer of redundant information for our purposes, and in this case the costs of transferring the entire graph of the experiment may be too high.

To solve the problem of combining changes in the graph of the experiment G we considered three types of modifications (operations) $O = \{add, del, edit\}$ over the nodes and edges of the graph that agents can make:

1. **Add** – Add a new node to the experiment graph and the specified set of incident edges;
2. **Del** – Delete an existing node from the graph and all incident edges;
3. **Edit** – Change the specified node.

Accordingly, the result of training the model at the node *Train Lin. Model SGD*, is added to the experiment graph by adding a new node. Changing the hyperparameter of the learning algorithm is done by removing and adding a new node. Replacing the implementation of the learning algorithm is a changing of the corresponding node in the experiment execution procedure.

We imposed restrictions on the operations due to the definition of CRDT and, after analyzing all 6 possible combinations, showed that in such conditions concurrent changes commute pairwise, that is, the result of their application does not depend on the order. The consequence of this was the following.

Statement 1. *The constructed data structure* $\langle G, O \rangle$ *is indeed a CRDT and has the property of Strong Eventual Consistency.*

Thus, in this paper we have shown that our way of representing the graph of the experiment and the way of exchanging modifications allows for implementation in the form of CRDT. Statement 1 theoretically justifies the possibility of using the constructed data structure to solve the original problem. Researchers and computational agents can store local copies of the experiment graph and manipulate it, and after making changes, they can run the data exchange algorithm. Upon its completion, each agent will receive an updated version of the experiment graph, which contains both changes made by other agents and modifications created by himself.

[1] *Gundersen, Odd Erik & Gil, Yolanda & Aha, David.* (2018). On Reproducible AI: Towards Reproducible Research, Open Science, and Digital Scholarship in AI Publications. // AI Magazine. 39. 56-68. 10.1609/aimag.v39i3.2816.

[2] *Khritankov, A., Pershin, N., Ukhov, N., & Ukhov, A.* (2022). MLDev: Data Science Experiment Automation and Reproducibility Software. // International Conference on Data Analytics and Management in Data Intensive Domains (pp. 3-18). Springer.

[3] *Marc Shapiro, Nuno Preguiça, Carlos Baquero, Marek Zawirski.* Conflict-free Replicated Data Types. // SSS 2011 - 13th International Symposium Stabilization, Safety, and Security of Distributed Systems, Grenoble, France. pp.386-400.

[4] *Chatterjee, Bapi and Peri, Sathya and Sa, Muktikanta and Manogna, Komma.* Non-Blocking Dynamic Unbounded Graphs with Worst-Case Amortized Bounds. // 25th International Conference on Principles of Distributed Systems (OPODIS 2021), Dagstuhl, Germany. pp.20:1-20:25.

[5] *Chatterjee, Bapi and Peri, Sathya and Sa, Muktikanta and Singhal, Nandini.* A Simple and Practical Concurrent Non-Blocking Unbounded Graph with Linearizable Reachability Queries. // Proceedings of the 20th International Conference on Distributed Computing and Networking, Bangalore, India. pp.168–177.

Ускорение расчета термодинамического равновесия методами машинного обучения

Ашинов Бислан Рамазанович[1]⋆ ashinov.bislan1@yandex.ru
Майсурадзе Арчил Ивериевич[1] maysuradze@cs.msu.ru
[1]Москва, МГУ имени М. В. Ломоносова

Моделирование термодинамических процессов обычно проводится путем решения уравнений математической физики с использованием трудоемких численных методов, в частности итерационных алгоритмов. Но бывают случаи, когда трудоемкость расчетов большая, а требования к точности не такие высокие. Ситуация многократно усугубляется в гидродинамических симуляторах, когда указанный расчет надо провести в каждой ячейке пространства в каждый момент модельного времени. Целью исследования было использовать методы машинного обучения (МО) для ускорения расчетов в специальном классе задач термодинамики, получив малое время расчетов и сохранив хорошую точность.

В работе предлагается методика ускорения массовых термодинамических расчетов, когда для индивидуальной физической задачи из рассматриваемого класса строится выборка прецедентов, по ней обучается быстродействующая модель МО определенной архитектуры и далее многократные запуски симулятора проводятся в ускоренном режиме. Такая методика может быть практически полностью автоматизирована, т. е. для каждой следующей индивидуальной физической задачи практически автоматически будет сгенерирован симулятор, включающий ускоренные расчеты.

Подробнее опишем рассматриваемый класс термодинамических задач. Ищется термодинамическое равновесие в ячейке, в которой находится смесь нескольких известных веществ, причем возможна смесь жидкого и газового агрегатных состояний. На вход подаются температура и давление в ячейке, а также общий состав смеси. В первую очередь необходимо найти агрегатное состояние вещества (жидкость, газ или смесь). В случае жидкости надо рассчитать ее плотность и вязкость. В случае газа тоже надо рассчитать его плотность и вязкость. В случае смеси для каждого агрегатного состояния надо рассчитать все указанное раннее, а также состав газа, состав жидкости, и долю жидкости и долю газа в общем веществе. При традиционном физическом моделировании указанное равновесие ищется как решение сложной системы уравнений методом Ньютона.

Особенностью рассматриваемого класса термодинамических задач является то, что как входное, так и выходное описания содержат данные о составе смеси. Данные такого вида называют составными или композиционными (англ. compositional data) [1]. С математической точки зрения вектор состава представляет собой точку многомерного симплекса. Стоит отметить, что такие данные возникают не только в задачах физики и химии. Стандартные методы МО не способны непосредственно принимать наборы признаков, образующие состав.

Такие данные в классических постановках задач машинного обучения нельзя назвать независимыми признаками, они требуют отдельного подхода. Некоторые из таких подходов были использованы в работе.

Сначала для индивидуальной физической задачи выделяется область входов термодинамической задачи. Часто это одна температура и небольшой диапазон давлений. При помощи традиционной физической модели генерируются прецеденты для обучения и контроля. Далее надо обучить модели МО. Важно, чтобы эти модели быстро работали. Предлагается провести преобразование признакового пространства для составных данных, чтобы на преобразованных данных использовать известные модели МО. В ходе исследования на преобразованных прецедентах были обучены простые модели машинного обучения, а именно, логистическая регрессия, линейные регрессии, двухслойные полносвязные нейронные сети.

Результаты экспериментов на ряде индивидуальных физических задач показывают, что обученные модели дают допустимую точность, преимущество в скорости при этом существенное.

Работа выполнена при поддержке НОШ МГУ «Мозг, когнитивные системы, искусственный интеллект», НИР МГУ 5.1.21, гранта РФФИ No. 20-01-00664. Благодарим А. А. Афанасьева и его учеников за постановку термодинамической задачи и доступ к данным традиционного физического моделирования.

[1] *Aitchison J.* The Statistical Analysis of Compositional Data // Journal of the Royal Statistical Society, 1982.

Acceleration of the Thermodynamic Equilibrium Calculation by Machine Learning Methods

Ashinov Bislan[1]⋆ ashinov.bislan1@yandex.ru
Maysuradze Archil[1] maysuradze@cs.msu.ru
[1]Moscow, Lomonosov Moscow State University

Modeling of thermodynamic processes is usually carried out by solving equations of mathematical physics using time-consuming numerical methods, in particular iterative algorithms. But there are cases when the complexity of calculations is large, and the accuracy requirements are not so high. The situation is repeatedly aggravated in hydrodynamic simulators, when the specified calculation must be carried out in each cell of space at each moment of the model time. The aim of the study was to use machine learning (MO) methods to accelerate calculations in a special class of thermodynamics problems, obtaining a short calculation time and maintaining good accuracy.

We propose a method for accelerating mass thermodynamic calculations, when a sample of precedents is built for an individual physical problem from the class under consideration, a high-speed model of a certain architecture is trained on it, and then multiple runs of the simulator are carried out in accelerated mode. Such a technique can be almost completely automated, i.e. for each subsequent individual physical task, a simulator will be generated almost automatically, including accelerated calculations.

Let us describe in more detail the considered class of thermodynamic problems. A thermodynamic equilibrium is sought in a cell containing a mixture of several known substances, and a mixture of liquid and gas aggregate states is possible. The input is the temperature and pressure in the cell, as well as the total composition of the mixture. First of all, it is necessary to find the state of aggregation of a substance (liquid, gas or mixture). In the case of a liquid, its density and viscosity must be calculated. In the case of a gas, it is also necessary to calculate its density and viscosity. In the case of a mixture, for each state of aggregation, it is necessary to calculate everything indicated earlier, as well as the composition of the gas, the composition of the liquid, and the proportion of liquid and the proportion of gas in the total substance. In traditional physical modeling, this equilibrium is sought as a solution to a complex system of equations by Newton's method.

A feature of the considered class of thermodynamic problems is that both the input and output descriptions contain data on the composition of the mixture. This type of data is called composite or compositional data [**?**]. From a mathematical point of view, the composition vector is a point of a multidimensional simplex. It should be noted that such data arise not only in problems of physics and chemistry. Standard ML methods are not able to directly accept sets of features that form a composition. Such data in the classical formulation of machine learning problems

cannot be called independent features, they require a separate approach. Some of these approaches have been used in this work.

First, for an individual physical problem, the region of inputs of the thermodynamic problem is allocated. Often this is one temperature and a small range of pressures. Using a traditional physical model, use cases are generated for training and control. The next step is to train the MO models. It is important that these models work quickly. It is proposed to transform the feature space for composite data in order to use known ML models on the transformed data. During the study, simple machine learning models were trained on transformed precedents, namely, logistic regression, linear regressions, two-layer fully connected neural networks.

The results of experiments on a number of individual physical problems show that the trained models provide acceptable accuracy, while the advantage in speed is significant.

The research is supported by Scientific and educational school of Moscow State University "¡Brain, cognitive systems, artificial intelligence"¿, research work of Moscow State University 5.1.21, RFBR grant No. 20-01-00664. We thank Andrey Afanasyev and his students for setting the thermodynamic problem and access to the data of traditional physical modeling.

[1] *Aitchison J.* The Statistical Analysis of Compositional Data // Journal of the Royal Statistical Society, 1982.

Улучшение качества реидентинфикации людей посредством self-supervised предобучения

Мамедов Тимур Закирович[1,2]⋆ timur.mamedov@graphics.cs.msu.ru
Купляков Денис Анатольевич[1,2] denis.kuplyakov@graphics.cs.msu.ru
Конушин Антон Сергеевич[1,3] anton.konushin@graphics.cs.msu.ru
[1]Москва, Московский государственный университет им. М.В. Ломоносова
[2]Москва, ООО «Технологии видеоанализа»
[3]Москва, Национальный исследовательский университет «Высшая школа экономики»

В данной работе предлагается метод решения задачи реидентификации людей с использованием нейронных сетей. Суть реидентификации заключается в том, что по двум изображениям необходимо определить запечатлен ли на них один и тот же человек или нет. Рассматриваемая задача является практически важной, а алгоритмы ее решения нашли широкое применение в видеоаналитике (в видеонаблюдении, задачах маркетинга, анализе спортивных мероприятий).

Ввиду того, что реидентификация подразумевает работу с реальными данными, исследователям, решающим данную задачу, приходится сталкиваться со множеством трудностей, связанных с разнообразием ситуаций, возникающих в анализируемых массивах данных. Очевидным путем преодоления упомянутой трудности является обучение алгоритмов реидентификации на весьма крупных выборках, которые потенциально могут покрыть бóльшую часть возникающих сценариев в реальных данных.

Однако приведенный выше подход имеет несколько аспектов, которые делают его применение практически невозможным или слишком затратным. Во-первых, выборка должна быть максимально разнообразной как с точки зрения запечатленных ситуаций, так и визуальной составляющей (освещения, качества, ракурсов и т.д.). Во-вторых, как следствие первого аспекта, данных должно быть очень много, речь идет о нескольких миллионах изображений, которые необходимо размечать вручную, что требует привлечения большого числа разметчиков и приводит к крупным денежным затратам. В-третьих, данные для реидентификации весьма сложно собрать из-за специфики самой задачи.

В силу описанных выше обстоятельств, в данной работе предлагается методика сбора данных и обучения нейросетевых алгоритмов реидентификации людей с использованием self-supervised предобучения, призванного уменьшить потребность в больших объемах размеченных вручную данных и улучшить качество решения рассматриваемой задачи.

Методы реидентификации внедряются в системы видеонаблюдения, где чаще всего приходится иметь дело с большими скоплениями людей, поэтому в данной работе предлагается производить реидентификацию не в полный рост человека, как это принято в классических подходах, а по детекциям верхней

части человеческого тела. Состоятельность такого подхода была подтверждена соответствующими экспериментами в [1].

Self-supervised предобучение требует большой и разнообразной коллекции изображений людей. В данной работе предлагается стратегия для ее сбора. С помощью алгоритма сопровождения [1] (трекинга) строятся траектории движения людей на видеозаписях, собранных из открытых источников. Далее, используя полученные траектории, происходит последующая нарезка изображений (кропов) людей с кадров видеозаписей. После чего кропы людей фильтруются в автоматическом режиме: удаляются ложные детекции, не двигающиеся в течение всего видео люди и т.д.

Предложенный выше подход позволяет собирать данные для предобучения нейросети из открытых источников в крупных объемах практически без участия человека. Причем за счет использования алгоритма трекинга полученная выборка состоит не из отдельных изображений людей, а из множества примеров для каждого человека, попавшего в датасет, что благоприятно влияет на качество предобучения нейронной сети для задачи реидентификации. Более того, если взять в рассмотрение тот факт, что одна траектория – это один человек, то имеется возможность в полностью автоматическом режиме получить разметку, которая также может быть применена во время self-supervised предобучения.

В рамках данной работы было отобрано 371 видео с камер видеонаблюдения, публично транслируемых в сети Интернет, общей продолжительностью более 2500 часов. В течение 2,5 недель все видео обрабатывались согласно приведеной выше стратегии, и в итоге было получено около 11,5 миллионов кропов порядка 980 тысяч людей. Таким образом, был собран самый большой набор данных из ныне известных для задачи реиндентификации людей.

Все полученные данные были задействованы во время self-supervised предобучения нейросети с использованием методики, предложенной в [2]. После предобучения нейронная сеть дообучалась на открытом наборе данных для реидентификации MSMT17 [3]. Тестирование полученной сети проходило на датасете DukeMTMC-reID [4]. То есть было проведено кросс-доменное тестирование, что является наиболее репрезентативным для алгоритмов, которые в будущем планируется использовать на практике.

Как итог, предложенный подход позволяет улучшить качество реидентификации с точки зрения общепринятых в данной задаче метрик $Rank_1$ и mAP и ускорить время дообучения нейросети под конечную задачу по сравнению с дообучением весов, полученных в supervised-режиме на датасете ImageNet [5] для классификации, как это подразумевает большинство других методов.

[1] *T. Mamedov, D. Kuplyakov, A. Konushin* Queue Waiting Time Estimation Using Person Re-identification by Upper Body // Proceedings of the 31th International Conference on Computer Graphics and Machine Vision, Nizhny Novgorod: CEUR Workshop Proceedings, 2021. — pp. 464–474.

[2] *X. Chen, H. Fan, R. Girshick, X. Chen* Improved Baselines with Momentum Contrastive Learning // CoRR, 2020.

[3] *L. Wei, S. Zhang, W. Gao, Q. Tian* Improving person re-identification by attribute and identity learning // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach: CVPR, 2019. — pp. 2138–2147.

[4] *R. Ristani, F. Solera, R. S. Zou, R. Cucchiara, C. Tomasi,* Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking // CoRR, 2016.

[5] *J. Deng, W. Dong, R. Socher, R. -J. Li, Kai Li, Li Fei-Fei,* ImageNet: A large-scale hierarchical image database // IEEE Conference on Computer Vision and Pattern Recognition, Miami: CVPR, 2020. — pp. 248–255.

Improving the quality of person re-identification by self-supervised pre-training

Mamedov Timur[1,2]⋆ timur.mamedov@graphics.cs.msu.ru
Kuplyakov Denis[1,2] denis.kuplyakov@graphics.cs.msu.ru
Konushin Anton[1,3] anton.konushin@graphics.cs.msu.ru
[1]Moscow, Moscow State University
[2]Moscow, Video Analysis Technologies LLC
[3]Moscow, National Research University Higher School of Economics

In this paper, we propose a method for solving the problem of person re-identification using neural networks. The essence of re-identification is that it is necessary to determine from two images whether the same person is depicted on them or not. The problem under consideration is practically important, and algorithms for its solution have found wide application in video analytics (in video surveillance, marketing tasks, analysis of sports events).

Due to the fact that re-identification involves working with real data, researchers solving this problem have to face many difficulties associated with a variety of situations arising in the analyzed data. The obvious way to overcome this difficulty is to train re-identification algorithms on large dataset that can potentially cover the majority of emerging scenarios in real data.

However, the above approach has several aspects that make its application practically impossible or too costly. Firstly, the dataset should be as diverse as possible both in terms of captured situations and the visual components (lighting, quality, angles, etc.). Secondly, as a consequence of the first aspect, there should be a lot of data, we are talking about several million images that need to be marked up manually, which requires the involvement of a large number of markers and leads to large monetary costs. Thirdly, re-identification data is very difficult to collect due to the specifics of the task itself.

Due to the circumstances described above, this paper proposes a methodology for data collection and training of neural network algorithms for person re-identification using self-supervised pre-training designed to reduce the need for large amounts of manually marked data and improve the quality of solving the problem under consideration.

Methods of re-identification are being introduced into video surveillance systems, where large crowds of people most often have to deal with. Therefore, in this paper it is proposed to perform re-identification not by a full-body detections, as is done in classical approaches, but by the upper part of the human body detections. The validity of this approach has been confirmed by relevant experiments in [1].

Self-supervised pre-training requires a large and diverse collection of images of people. In this paper, a strategy for collecting it is proposed. With the help of the tracking algorithm [1], the tracks of the movement of people are built on video recordings collected from open sources. Next, using the obtained tracks, images of

people are cropped from video frames of video recordings. After that, the crops of people are filtered automatically: false detections and cases when people not moving during the whole video are removed, etc.

The approach proposed above makes it possible to collect data for the pre-training of a neural network from open sources in large volumes practically without human participation. Moreover, due to the use of the tracking algorithm, the resulting dataset doesn't consist of individual images of people, but of many examples for each person who got into the dataset, which favorably affects the quality of neural network pre-training for the task of person re-identification. Additionally, if we take into consideration the fact that one track is one person, then it is possible to obtain markup in a fully automatic mode, which can also be applied during self-supervised pre-training.

As part of this work, 371 videos from surveillance cameras publicly broadcast on the Internet were selected, with a total duration of more than 2500 hours. For 2.5 weeks, all videos were processed according to the strategy above, and as a result, about 11.5 million crops of about 980 thousand people were obtained. Thus, the largest dataset currently known for the task of person re-indentification was collected.

All the data obtained were used during self-supervised neural network pre-training using the methodology proposed in [2]. After pre-training, the neural network was fine-tuned on an open dataset for re-identification MSMT17 [3]. The resulting neural network was tested on the DukeMTMC-reID [4] dataset, i.e. cross-domain testing was carried out, which is the most representative for algorithms that are planned to be used in practice in the future.

As a result, the proposed approach makes it possible to improve the quality of re-identification from the point of view of the $Rank_1$ and mAP metrics, which generally accepted in this task, and to speed up the time of fine-tuning the neural network for the final task compared with fine-tuning the weights obtained in supervised mode on the ImageNet [5] dataset for classification, as most other methods imply.

[1] *T. Mamedov, D. Kuplyakov, A. Konushin* Queue Waiting Time Estimation Using Person Re-identification by Upper Body // Proceedings of the 31th International Conference on Computer Graphics and Machine Vision, Nizhny Novgorod: CEUR Workshop Proceedings, 2021. — pp. 464–474.

[2] *X. Chen, H. Fan, R. Girshick, X. Chen* Improved Baselines with Momentum Contrastive Learning // CoRR, 2020.

[3] *L. Wei, S. Zhang, W. Gao, Q. Tian,* Improving person re-identification by attribute and identity learning // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach: CVPR, 2019. — pp. 2138–2147.

[4] *R. Ristani, F. Solera, R. S. Zou, R. Cucchiara, C. Tomasi,* Performance Measures and a Data Set for Multi-Target, Multi-Camera Tracking // CoRR, 2016.

[5] *J. Deng, W. Dong, R. Socher, R. -J. Li, Kai Li, Li Fei-Fei,* ImageNet: A large-scale hierarchical image database // IEEE Conference on Computer Vision and Pattern Recognition, Miami: CVPR, 2020. — pp. 248–255.

Улучшение кросс-корреляционного анализа с помощью преобразования пиксельной текстуры изображений методами глубокого обучения и визуализация крупно-волновой фибрилляции на открытом сердце

Мангилева Дарья Владимировна[1] daria.mangileva@urfu.ru

[1]Екатеринбург, Уральский федеральный университет имени первого президента России Б. Н. Ельцина (УрФУ)

Кросс-корреляционный анализ изображений является очень важной составляющей во многих естественнонаучных областях. Однако, для получения качественной информации о полях смещения текстура изображений должна быть достаточно разнородная. Именно поэтому исследователи стараются подготовить эксперименты таким образом, чтобы получить кадры с соответствующим видом, применяя различного рода маркеры. Однако, это не всегда является возможным. Так, например, в кардиологии при исследовании поверхности открытого бьющегося сердца, химическое или механическое вмешательство на мягкие ткани может повлиять на течение естественных электро-механических процессов. В данной работе была разработана новая схема глубокой нейронной сети, обучающая без учителя, которая способна изменить текстуру изображения для получения более достоверной информации с помощью кросс-корреляционного анализа. С помощью данного метода удалось зафиксировать механическую спиральную волну на эпикарде в течение нескольких миллисекунд.

В основу работы, разработанной в данном исследовании нейронной сети, заключены технологии позиционного кодирования, которые с недавних времен начали применяться для многослойных перцептронов (МП) [1]. Схема глубокой нейронной сети изображена на рисунке 1. Она состоит из двух основных компонентов, являющихся МП – генератора сеток (ГС) и генератора изображений (ГИ) [3]. В ГС поступает недеформированное поле координат (x,y) (прямая сетка) и выходит поле с необходимой деформацией (деформированная сетка), а ГИ, на основе данных полей, генерирует изображение. Изображения I_n и I_{n+1} это исходные изображения. Сперва ГИ обучается генерировать I_n, а ГС - недеформированное поле. Этот этап важен в первую очередь для инициализации весов. Далее генератор сеток стремится подать такую деформированную сетку, чтобы генератор изображений выдал следующее изображение I_{n+1}. После этого ГИ немного переобучается, сопоставляя прямую сетку уже с изображением похожим на гауссов шум (I_t). Так получается первое трансформированное изображение I'_n. Завершающим этапом является подача деформированной сетки в ГИ, который выдает I'_{n+1}, имеющий вид гауссова шума. I'_n и I'_{n+1} являются итоговым продуктом данной глубокой нейронной сети, которые имеют схожую деформацию с исходными изображениями, но, иную, более разнородную пиксельную текстуру, способствующую более качественному кросс-корреляционному анализу.

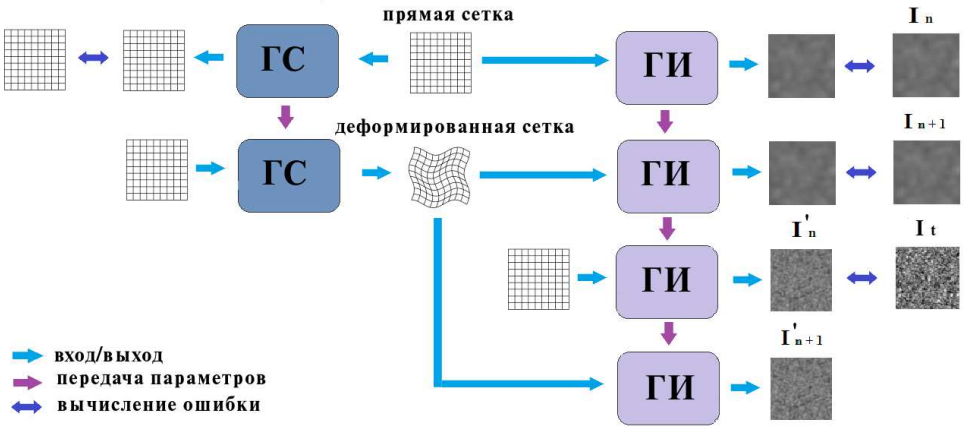


Рис. 1. Схема глубокой нейронной сети, разработанная в данном исследовании

Тестирование новой схемы нейронной сети производилось с помощью синтетических изображений с разной степенью разнородности и частотой полей смещений, полученных с помощью открытых библиотек на Python: muDIC [4] и GSTools [5]. Для доказательства эффективности предложенного метода исходные изображения обрабатывались двумя кросс-корреляционными методами – PIV[6] и DIC[7], а затем та же процедура проводилась и для, сгенерированных нейронной сетью, изображений. Сравнение проводилось с помощью среднеквадратической ошибки (RMSE). Проведенные испытания показали, что применение нейронной сети дает выигрыш в ошибке при определении полей смещения во всех случаях за исключением, когда исходная текстура изображения уже достаточно разнородная.

После проведенного тестирования, созданная в данной работе нейронная сеть, была применена на реальные изображения открытого бьющегося сердца во время фибрилляции с низко разнородной текстурой. Рисунок 2 показывает поля смещения в виде стримлайновых графиков для случая с использованием предложенного метода (a) и без (b) (PIV использовался для кросс-корреляционного анализа). При применении нейронной сети можно разглядеть крупно-волновую механическую спиральную волну, которая возникает в начале фибрилляции [7], зарегистрированной на ЭКГ, в отличие от случая без предобработки изображения данным методом.

Таким образом, в текущем исследовании, созданная схема глубокой нейронной сети, оказалась эффективной для улучшения качества кросс-корреляционного анализа. С помощью нее предположительно удалось зафик-

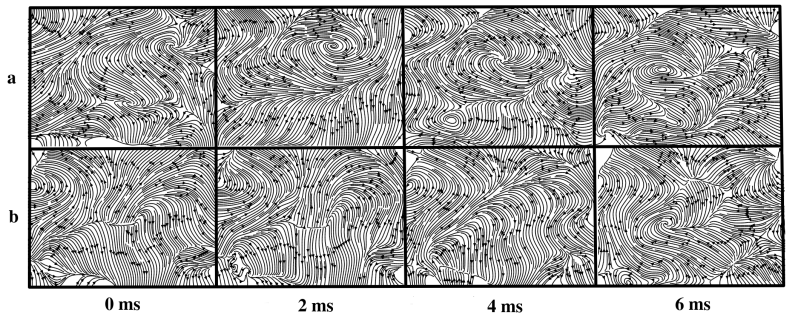


Рис. 2. Стримлайновые графики полей смещение мягких тканей открытого сердца

сировать механическую спиральную волну в течение нескольких миллисекунд, что может помочь более углубленному пониманию природы фибрилляций.

Исследование выполнено при финансовой поддержке Министерства науки и высшего образования Российской Федерации в рамках Программы развития Уральского федерального университета имени первого Президента России Б.Н. Ельцина в соответствии с программой стратегического академического лидерства "Приоритет-2030".

Для расчетов использовался суперкомпьютер "УРАН"в ИММ УрО РАН.

[1] *Tancik M.* Fourier features let networks learn high frequency functions in low dimensional domains // Advances in Neural Information Processing Systems, 2020, Т. 33. – С. 7537-7547.

[2] *Li N.* Unsupervised non-rigid image distortion removal via grid deformation // Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. С. 2522-2532.

[3] *Olufsen S.* muDIC: An open-source toolkit for digital image correlation // SoftwareX, 2020.

[4] *Müller S.* GSTools v1. 3: a toolbox for geostatistical modelling in Python // Geoscientific Model Development, 2022. – Т. 15. – №. 7. – С. 3161-3182.

[5] *Thielicke W.* PIVlab–towards user-friendly, affordable and accurate digital particle image velocimetry in MATLAB // Journal of open research software, 2014. – Т. 2. – №. 1.

[6] *Belloni V.* py2DIC: A new free and open source software for displacement and strain measurements in the field of experimental mechanic // Sensors, 2019. – Т. 19. – №. 18. – С. 3832.

[7] *Moe J.* A computer model of atrial fibrillation // American Heart Journal, 1964. – Т. 67. – №. 2. – С. 200-220.

Improving cross-correlation analysis using deep learning pixel texture transformation and visualization of large wave fibrillation in the open heart

Mangileva Daria[1] daria.mangileva@urfu.ru

[1]Ekaterinburg, Ural Federal University named after the First President of Russia B. N. Yeltsin

Cross-correlation analysis of images is a very important component in many natural sciences. However, to obtain high-quality information about the displacement fields, the texture of the images must be sufficiently heterogeneous. Thus, researchers usually seek to prepare experiments in such a way as to get frames with the appropriate heterogenous texture, using various markers. However, it is not always possible. For example, in cardiology, when examining the surface of an open beating heart, chemical or mechanical intervention on soft tissues can affect the course of natural electro-mechanical processes. In this work, a new unsupervised deep neural network scheme was developed that enable to change the texture of an image to obtain more reliable information using cross-correlation analysis. Using this method, it was possible to catch a mechanical spiral wave in epicardium within a few milliseconds.

The unsupervised deep neural network that developed in this study is based on the technology of positional coding for multilayer perceptrons (MLP) [1]. The

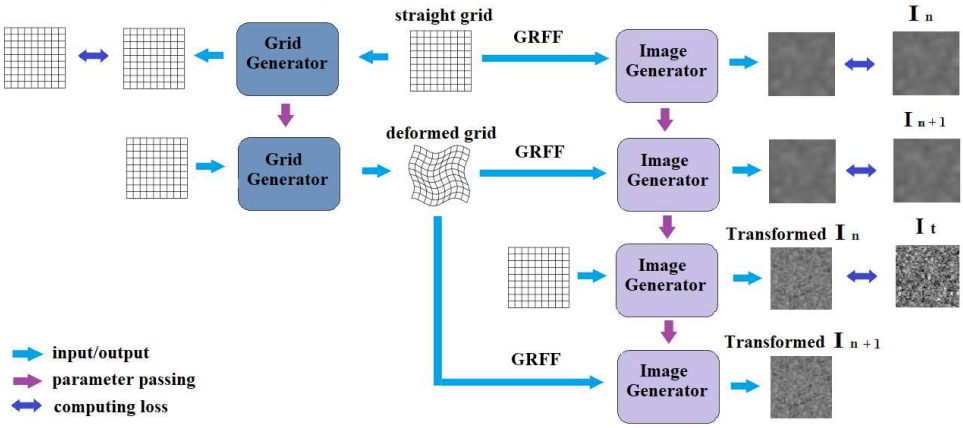


Fig. 1. Scheme of the deep neural network developed in this study

scheme of the deep neural network is shown in Figure 1. It consists of two main MLP components a grid generator (GG) and an image generator (IG) [3]. An non-deformed field of coordinates (x,y) (straight grid) is the input of the GG and a field

with the necessary deformation (deformed grid) is the output of GG. Thus, the GI, based on these fields, generates an image. The images I_n and I_{n+1} are the initial images. First, the IG is trained to generate I_n, and the GG is trained to generate an straight grid field. This step is important first of all for the initialization of the scales. Next, the grid generator seek to simulate such a deformed grid that the image generator produces the following image I_{n+1}. After that, the IG retraines a little in such a way that the straight grid compares with the image similar to Gaussian noise (I_t). As a result, transformed image I'_n is obtained. The final step is to feed the deformed grid into the IG and it produces I'_{n+1} which looks like Gaussian noise. I'_n and I'_{n+1} are the final product of this unsupervised deep neural network. They have a similar deformation with the initial images, but more heterogeneous pixel texture that contributes to better quality of cross-correlation analysis.

Testing of the novel neural network scheme was carried out using synthetic images with varying degrees of pixel heterogeneity and frequency of displacement fields, obtained using open libraries in Python: muDIC [4] and GSTools [5]. To prove the effectiveness of the proposed method, the original images were processed by two cross-correlation methods - PIV[6] and DIC[7], and then the same procedure was carried out for the images generated with the unsupervised deep neural network. Comparison was performed using root mean square error (RMSE). The tests have shown that the applying of a neural network gives a gain in the error for determining the displacement fields in all cases, except when the initial image pixel texture is already quite heterogeneous.

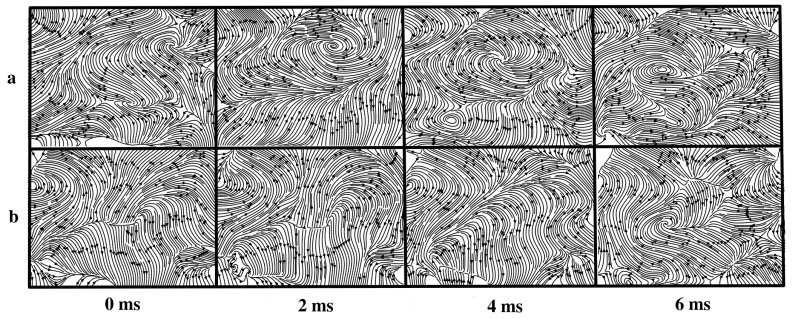


Fig. 2. Streamline plots of open heart soft tissue displacement fields

After testing, the neural network created in this work was applied to real images of an open beating heart during fibrillation with a low heterogeneous texture. Figure 2 shows the displacement fields as streamline plots for the case where proposed method was used (a) and without it (b) (PIV was used as cross-correlation method). In the first case, one can see a large mechanical spiral wave that occurs at the

beginning of fibrillation [7] recorded on the ECG, in contrast to the case without image preprocessing with proposed the neural network.

Thus, in the current study, the created unsupervised deep neural network scheme proved to be effective in improving the quality of cross-correlation analysis. With the help of it, it has become possible to fix a mechanical spiral wave within a few milliseconds that can help to better understand the nature of fibrillation.

The research funding from the Ministry of Science and Higher Education of the Russian Federation (Ural Federal University project within the Priority-2030 Program) is gratefully acknowledged.

Supercomputer URAN of IMM UrB RAS was used for calculations.

[1] *Tancik M.* Fourier features let networks learn high frequency functions in low dimensional domains // Advances in Neural Information Processing Systems, 2020. Vol. 33. Pp. 7537-7547.

[2] *Li N.* Unsupervised non-rigid image distortion removal via grid deformation // Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. Pp. 2522–2532.

[3] *Olufsen S.* muDIC: An open-source toolkit for digital image correlation // SoftwareX, 2020.

[4] *Müller S.* GSTools v1. 3: a toolbox for geostatistical modelling in Python // Geoscientific Model Development, 2022. – Vol. 15. – №. 7. – С. 3161-3182.

[5] *Thielicke W.* PIVlab–towards user-friendly, affordable and accurate digital particle image velocimetry in MATLAB // Journal of open research software, 2014. – Т. 2. – №. 1.

[6] *Belloni V.* py2DIC: A new free and open source software for displacement and strain measurements in the field of experimental mechanic // Sensors, 2019. – Т. 19. – №. 18. – С. 3832.

[7] *Moe J.* A computer model of atrial fibrillation // American Heart Journal, 1964. – Т. 67. – №. 2. – С. 200-220.

Прогнозирование влияния пандемии COVID-19 на человеческий капитал региона с помощью алгоритмов глубокого обучения

Каширина Ирина Леонидовна* kash.irina@mail.ru
Бондаренко Юлия Валентиновна bond.julia@mail.ru
Воронеж, Воронежский государственный университет

Во время пандемии COVID-2019 на рынке труда произошли глобальные и существенные изменения, что привело к переоценке человеческого капитала. Пандемия COVID-2019, которая охватила весь мир, трансформировала функционирование многих сфер деятельности, существенно увеличив долю дистанционно выполняемых работ. В условиях пандемии значительно возросли требования к обоснованности стратегических решений, принимаемых в области управления человеческим капиталом. Принимаемые в период пандемии решения должны учитывать достоверные прогнозы развития эпидемиологической ситуации в отдельных регионах и стране в целом. При этом большой популярностью при моделировании эпидемиологических процессов в последнее время пользуются алгоритмы машинного обучения, которые сами способны находить закономерности и аппроксимировать зависимости, опираясь на имеющуюся базу наблюдений [1, 2, 3, 4, 5]. Для обучения таких моделей и разработки точных прогнозных инструментов необходимо применение деперсонифицированных баз медицинских данных, а также выявление дополнительных факторов, которые оказывают влияние на развитие эпидемиологического процесса [3]. Таким образом, в настоящее время существует необходимость в разработке интеллектуальных подходов к исследованию различных аспектов распространения эпидемии COVID-19 в разрезе их влияния на человеческий капитал на региональном уровне.

Процесс разработки подобных подходов можно разбить на следующие этапы.

1. Разведочный анализ данных, включающий выявление и исследование признаков, оказывающих влияния на распространение COVID-19.
2. Разработка инструментов краткосрочного и долгосрочного прогнозирования динамики волн распространения эпидемии с применением гибридных нейросетевых технологий моделирования нелинейных зависимостей.
3. Разработка инструментов предсказания регионального уровня массовости и тяжести протекания эпидемиологического процесса с учетом анализа траектории динамики развития заболевания в России в целом.

В качестве исходных данных для построения моделей машинного обучения в данном исследовании использовался набор деперсонифицированных данных, предоставленных Воронежским областным клиническим консультативно-диагностическим центром (ВОККДЦ), включающий данные обо всех ПЦР тестах на COVID-19, которые были проведены в Воронежской области в период

с марта 2020 года по сентябрь 2022 года. Датасет содержит следующие показатели: уникальный id пациента; пол; возраст; дата забора теста; результат теста (положительный или отрицательный); район Воронежской области, в котором проживает пациент; медицинская организация, которая проводила тестирование; тест сдан амбулаторно или в стационаре; был ли тест сдан в одном из стационаров, в которые направляются преимущественно пациенты с тяжелыми случаями заболевания; зарегистрированы ли у данного пациента осложнения после перенесённого заболевания COVID-19, состоит ли пациент на диспансерном учете для реабилитации после COVID-19, был ли пациент привит более 2 недель назад. База данных постоянно пополняется. На 1 сентября 2022 года она содержала более 2.3 миллиона записей, включающих сведения о результатах ПЦР тестирования 998 тысяч уникальных пациентов.

Согласно значениям в столбце “Пол”, исходный датасет содержит 60 процентов женщин и 40% мужчин, то есть женщины в Воронежской области сдают ПЦР тесты в полтора раза чаще, чем мужчины. Если рассмотреть только случаи положительных тестов, то среди них пропорции такие же — 60 процентов женщин и 40 процентов мужчин. Исходя из этого можно сделать предположение, что заболеваемость COVID-19 не зависит от пола.

Средний возраст пациента в исходном датасете составляет 44 года, среди пациентов с ПЦР+ 47 лет. Распределение заболеваемости по возрасту являются бимодальным – первый пик приходится на возраст $35 - 40$ лет, второй – $60 - 65$ лет.

Были исследованы случаи повторных заболеваний, чтобы понять, как их учитывать в моделях прогнозирования динамики заболеваемости. Среднее значение периода между повторными заболеваниям составляет 282 дня со средним квадратическим отклонением в 129 дней.

Для выявления дополнительных признаков, оказывающих влияние на распространение пандемии, был проведен корреляционный анализ взаимосвязи между случаями COVID-19 и некоторыми внешними факторами. Для оценки корреляции были рассмотрены такие характеристики, число новых случаев ПЦР+ в определенный день в Воронежской области, скользящие средние для температуры, количества осадков и новых случаев ПЦР+ за предшествующую неделю, число активных случаев в предыдущий день, а также количество запросов по поисковым словам “Лечение коронавируса”, “Covid” и “Вызвать скорую” за предшествующую неделю.

Статистика запросов была взята с сайта `https://wordstat.yandex.ru`. Во время пандемии COVID-19 было опубликовано несколько исследований с использованием веб-данных, которые показали, что данные выдачи поисковых систем могут быть полезны для прогнозирования дальнейшего развития эпидемии [3]. При анализе связи заболеваемости с числом поисковых запросов был обнаружен интересный факт- пики в запросах, связанных с COVID происхо-

дили за $3-5$ дней до того, как достигался пик по новым заболевшим в базе ПЦР-тестов.

Метеорологические данные анализе были использованы потому, что в предыдущих исследованиях было отмечено, что они тоже могут оказывать влияние на заболеваемость [2].

Самая сильная корреляция числа заболевших в определенный день наблюдается с числом заболевших ранее и числом активных случаев. Однако все перечисленные выше показатели оказались значимы, включая температуру и количество осадков. Для построения модели прогнозирования динамики распространения COVID-19 в Воронежском регионе использовалась гибридная модель глубокого обучения, которая включает рекуррентные сети долгой краткосрочной памяти (LSTM), позволяющие на входе модели использовать несколько временных рядов, при этом признаки из временных рядов автоматически извлекаются с помощью сверточной нейронной сети (CNN).

Средняя абсолютная ошибка, рассчитанная для прогноза динамики на 15 дней вперед, составила $= 40.8$. Средняя процентная ошибка для этого же периода составила $P = 5\%$, то есть прогнозное значение в среднем на 5% отклонялось от реального.

Построенная модель может быть использована для принятия своевременных управляющих решений, направленных на снижение негативных последствий пандемии на человеческий капитал. Практическое значение моделей прогнозирования случаев инфицирования COVID-19 для управления развитием человеческого капитала включает в себя оценку характеристик пандемии для данного прогнозируемого периода и местности, сценарное планирование в секторе управления регионом, оптимизацию моделей и оценку корреляции различных факторов с динамикой заболеваемости.

Исследование выполнено при поддержке РФФИ, проект No. 19-29-07400.

[1] *Kashirina I., Bondarenko Y., Azarnova T.* Analysis and forecasting of the market of educational services of the region. Proceedings - 2021 // 1st International Conference on Technology Enhanced Learning in Higher Education, TELE 2021, Lipetsk, 2021. — p. 30–34.

[2] *Venkatesh U., Gandhi P. A.* Prediction of COVID-19 outbreaks using google trends in India: A retrospective analysis // Healthcare informatics research, 2020. —26(3). — p. 175–184.

[3] *Mavragani A., Gkillas K.* COVID-19 predictability in the United States using Google Trends time series // Scientific reports, 2020. —10(1). —p. 1–12.

[4] *Фирюлина М. А., Каширина И. Л.* Прогнозирование развития инфаркта миокардана основании сезонных и метеорологических факторов // Воронеж: Вестник Воронежского института высоких технологий, 2021. —No. 2. —С. 19–24.

[5] *Ketu S., Mishra P. K.* India perspective: CNN-LSTM hybrid deep learning model-based COVID-19 prediction and current status of medical resource availability // Soft Computing, 2022. —26(2) —p. 645–664.

Predicting the impact of the COVID-19 pandemic on the human capital of the region using deep learning algorithms

Kashirina Irina ⋆ kash.irina@mail.ru
Bondarenko Julia bond.julia@mail.ru
Voronezh, Voronezh State University

During the COVID-2019 pandemic, global and significant changes took place in the labor market, which led to a revaluation of human capital. The COVID-2019 pandemic, which has swept the whole world, has transformed the functioning of many areas of activity, significantly increasing the share of remote work. In the context of the pandemic, the requirements for the validity of strategic decisions taken in the field of human capital management have increased significantly. Decisions made during a pandemic should take into account reliable forecasts for the development of the epidemiological situation in individual regions and the country as a whole. At the same time, machine learning algorithms, which themselves are able to find patterns and approximate dependencies based on the existing observational base, have recently become very popular in modeling epidemiological processes [1, 2, 3, 4, 5]. To train such models and develop accurate predictive tools, it is necessary to use depersonalized medical databases, as well as to identify additional factors that influence the development of the epidemiological process [3]. Thus, there is currently a need to develop intelligent approaches to the study of various aspects of the spread of the COVID-19 epidemic in terms of their impact on human capital at the regional level.

The process of developing such approaches can be divided into the following stages.

1. Exploratory data analysis, including the identification and study of signs that affect the spread of COVID-19.
2. Development of tools for short-term and long-term forecasting of the dynamics of epidemic spread waves using hybrid neural network technologies for modeling non-linear dependencies.
3. Development of tools for predicting the regional level of mass and severity of the epidemiological process, taking into account the analysis of the trajectory of the dynamics of the development of the disease in Russia as a whole.

As input data for building machine learning models in this study, we used a set of depersonalized data provided by the Voronezh Regional Clinical Consultative and Diagnostic Center (VOKKDC), which includes data on all PCR tests for COVID-19 that were carried out in the Voronezh region since March 2020 to September 2022. The dataset contains the following parameters: unique patient id ; floor; age; date of sampling; test result (positive or negative); the district of the Voronezh region where the patient lives; the medical organization that conducted the testing; the test was passed on an outpatient basis or in a hospital; whether the test was passed in one of

the hospitals, which are referred mainly to patients with severe cases of the disease; whether this patient has complications after suffering COVID-19 disease, whether the patient is registered for rehabilitation after COVID-19, whether the patient was vaccinated more than 2 weeks ago. The database is constantly updated. As of September 1, 2022, it contained more than 2.3 million records, including information about the results of PCR testing of 998,000 unique patients.

According to the values in the "Sex" column, the original dataset contains 60 percent of women and 40% of men, that is, women in the Voronezh region take PCR tests one and a half times more often than men. If we consider only cases of positive tests, then among them the proportions are the same - 60 percent of women and 40 percent of men. Based on this, it can be assumed that the incidence of COVID-19 does not depend on gender.

Average age of the patient in the original dataset is 44 years, among patients with PCR+ 47 years. The distribution of incidence by age is bimodal - the first peak occurs at the age of $35 - 40$ years, the second - $60 - 65$ years.

Cases of recurrent illnesses were investigated to understand how to account for them in models for predicting the dynamics of morbidity. The mean relapse interval is 282 days with a standard deviation of 129 days.

To identify additional features that affect the spread of the pandemic, a correlation analysis of the relationship between cases of COVID-19 and some external factors was carried out. To assess the correlation, the following characteristics were considered: the number of new PCR+ cases on a certain day in the Voronezh region, moving averages for temperature, precipitation and new PCR+ cases for the previous week, the number of active cases on the previous day, as well as the number of requests for search words "Coronavirus treatment", "Covid" and "Call an ambulance" for the previous week.

Query statistics were taken from the site `https://wordstat.yandex.ru`. During the COVID-19 pandemic, several studies using web data were published that showed that search engine results data can be useful in predicting the future development of the epidemic [3]. When analyzing the relationship between the incidence and the number of search queries, an interesting fact was found - peaks in queries related to COVID occurred 3-5 days before the peak was reached for new cases in the PCR test database.

Meteorological data were used in the analysis because previous studies had noted that they too could have an impact on morbidity[2].

The strongest correlation of the number of cases on a given day is observed with the number of cases earlier and the number of active cases. However, all of the above indicators were significant, including temperature and precipitation. To build a model for predicting the dynamics of the spread of COVID-19 in the Voronezh region, a hybrid deep learning model was used, which includes recurrent networks of long short-term memory (LSTM), which allow using several time series at the

input of the model, while features from time series are automatically extracted using convolutional neural networks (CNN).

The average absolute error calculated for the dynamics forecast for 15 days ahead was $MAE = 40.8$. The average percentage error for the same period was $MAPE = 5\%$, that is, the predicted value deviated from the real one by an average of 5%.

The constructed model can be used to make timely management decisions aimed at reducing the negative effects of the pandemic on human capital. The practical significance of COVID-19 infection forecasting models for managing human capital development includes assessing the characteristics of the pandemic for a given forecast period and locality, scenario planning in the management sector of the region, optimizing models, and evaluating the correlation of various factors with incidence dynamics.

The study was supported by the Russian Foundation for Basic Research, project 19-29-07400.

[1] *Kashirina I., Bondarenko Y., Azarnova T.* Analysis and forecasting of the market of educational services of the region. Proceedings - 2021 // 1st International Conference on Technology Enhanced Learning in Higher Education, TELE 2021, Lipetsk, 2021. — p. 30–34.

[2] *Venkatesh U., Gandhi P. A.* Prediction of COVID-19 outbreaks using google trends in India: A retrospective analysis // Healthcare informatics research, 2020. —26(3). — p. 175–184.

[3] *Mavragani A., Gkillas K.* COVID-19 predictability in the United States using Google Trends time series // Scientific reports, 2020. —10(1). —p. 1–12.

[4] *Firulina , M. A., Kashirina I. L.* Prediction of the development of myocardial infarction based on seasonal and meteorological factors // Voronezh: Bulletin of the Voronezh Institute of High Technologies, 2021. —2(37). —p. 19-24

[5] *Ketu S., Mishra P. K.* India perspective: CNN-LSTM hybrid deep learning model-based COVID-19 prediction and current status of medical resource availability // Soft Computing, 2022. —26(2) —p. 645–664.

Блочная реализация внимания в Трансформере для сквозного распознавания речи

Чучупал Владимир Яковлевич[1] v.chuchupal@mail.ru

[1]Москва, Федеральный исследовательский центр «Информатика и управление» Российской академии наук

Успешное применение нейросетевых моделей типа кодер-декодер с вниманием связано с высокой вычислительной нагрузкой [1]. В докладе предложены новые методы численной реализации само–внимания и внимания для распознавания речи, которые используют свойства речевого сигнала. Это позволяет существенно понизить вычислительную требования по объёму используемой памяти и по числу операций.

В [2] был предложен, основанный на свойстве кратковременности речи, экономный по памяти и объёму вычислений пошаговый алгоритм вычисления само–внимания в кодере Трансформера, который аппроксимировал результат для всей последовательности данных оценками, полученными на коротких блоках. Алгоритм имеет вычислительную сложность порядка $O(n)$ как по памяти так и по числу операций, по сравнению с $O(n^2)$ для общепринятого.

Помимо само–внимания в кодере Трансформер содержит также подслои внимания и само–внимания в декодере, вычислительные требования которых также могут понижены за счёт учёта физических свойств речи.

Само–внимание в декодере выполняет функцию модели языка, поэтому можно использовать кратковременный характер зависимости между символами. В частности, использовать ограничение по длине для контекстов из предыдущих символов. Такие ограничения можно реализовать с помощью использования маскировочных матрицы ленточного вида, в отличие от нижних треугольных матриц в общепринятом методе. Если придерживаться аналогии с n-граммными языковыми моделями, само–внимание для каждого символа в этом случае будет иметь один и тот же порядок.

Более формально идея экономного алгоритма вычисления само-внимания декодера имеет вид:

```
Вход -- последовательность Q = q[0],...,q[|Q|-1] символов,
   n -- размер блока (порядок модели языка)
Положим: K = V = Q
Вычислим:
    Маскировочную матрицу: M[i,j] = 1 if 0<= i-j <= n else 0, где 0<=i,j<|Q|
    Счета внимания:        S  = b * matmul(tr(Q),K) * M
    Веса внимания:         A  = softmax (S)
    Внимание:              Y  = matmul(A, V)
    Выходные значения      OUT  = ReLu(Y)
```

Здесь matmul обозначает перемножение матриц, b = 1/sqrt(d), где d – размерность модели, ReLu - полулинейная функция активации.

Реализация вычисления внимания в декодере аналогично методу [2] неочевидна из–за десинхронизации символов декодера и выходов кодера. Кодер работает поблочно, поэтому для поиска соответствия между символами декодера и блоками кодера используем средние от данных на блоке. Относим символ к тому блоку, в котором суммарная величина счетов внимания символа максимальна. После того, как для блока вычислены все его символы, внимание вычисляется только на этом блоке и для этих символов.

Более формально идея алгоритма внимания декодера:

```
Вход:
    1. q[i], i=0,...,|Q|-1 -- последовательность символов,
    в виде матрицы Q[i,j], Q[i,*] -- векторный код символа i.
    2. выход кодера K, как последовательность из N блоков K[i]:
         K = K[0],K[1]|, ... ,K[N-1]
Вычислим:
   Матрицу счетов внимания
       S = matmul(tr(Q),K),  где S[i,j] -- счёт q[i] на блоке K[j]
   Для i=0,...,|Q|-1:
       Найдём номер j блока символа q[i]: j = argmax_k S[i,k], 0<=k<N.
   Для j=0, ... ,N-1:
       Соберем символьную матрицу Q[j] из строк Q для символов K[j].
       Вычислим на блоке K[j]:
            Счета внимания: S = matmul(tr(Q[j]),K[j])
            Веса внимания:  A = softmax(S)
            Внимание:       Y[j] = matmul(tr(Q[j]),A)
Вычислим выход как конкатенацию: OUTPUT = [Y[0]|Y[1]|....|Y[N-1]]
```

Поскольку используется фиксированный размер блоков, сложность по памяти и объёму вычислений (при последовательном выполнении) всех методов – $O(n)$.

Численные эксперименты выполнялись на цифровой части корпуса TeCoRus [3]. Для обучения использовалось 16357 последовательностей, произнесённых 155 дикторами. Тестирование проводилось на 186 фразах от новых дикторов. Использовалось стандартная параметризация речи на основе мел-спектральных признаков (24 мел–спектральных коэффициента). Язык состоял из букв и буквосочетаний, символов пунктуации. Регулярные слова на выходе декодера получаются конкатенацией декодированных символов до появления символа–признака конца слова.

Сквозная система распознавания речи реализована как Трансформер с двумя слоями в кодере и декодере. Функция потерь при обучении – кросс–энтропия. Алгоритм оптимизации - Adam. Использовалось линейное понижения коэффи-

циента скорости обучения (начиная с 0.003) с коэффициентом 0.96 каждые 300 шагов.

При выборе размера пакетов от 150 предложений и более функция потерь убывала почти монотонно, потери практически обнулялись за 300 эпох.

В таблице 1 показана зависимость пословной точности распознавания от выбора размеров блоков. Очевидно, что использование блочных алгоритмов не привело к ухудшению точности. Значения точности для разных размеров блоков и контекстов различаются несущественно (в пределах 1-2 процентов). Для само–внимания в кодере максимальная точность наблюдается при размере блока в 1 сек. возможно потому, что размер контекста примерно соответствует максимальной длительности звуков. Для само–внимания в декодере оптимальная длительность левого контекста отличается от максимально возможной и составляет 5 символов, что можно объяснить размером обучающего датасета.

Таблица 1. Зависимость точности распознавания от размеров блоков контекстов внимания и само–внимания. L,C,R обозначают длины левого, центрального и правого контекстов, в секундах или символах. Обозначение inf соответствует максимальному по длине контексту в общепринятом методе.

Кодер			Декодер			Точность
С–Вним.(сек.)			С–Вним.(симв)		Вним.(сек.)	
C	L	R	C	L	C	
inf	inf	inf	inf	inf	inf	0.942
2.0	1.0	1.0	inf	inf	inf	0.956
1.0	0.5	0.5	inf	inf	inf	0.966
1.0	0.	0.	15	15	inf	0.939
1.0	0.	0.	7	7	inf	0.942
1.0	0.	0.	5	5	inf	0.976
1.0	0.	0.	5	5	8.	0.971
1.0	0.	0.	15	7	4.	0.972
1.0	0.	0.	15	7	2.	0.941

[1] *Чучупал В. Я.* Акустическое и языковое моделирование в сквозных системах распознавания речи. // Цифровая обработка сигналов, Москва: Российское НТО радиотехники, электроники и связи им. А.С.Попова, 2020. —No.4, — С. 34–43.

[2] *Чучупал В. Я.* Экономная модель трансформера для акустического моделирования речи // Тезисы докладов 20-й Всесоюзной конференции Математические методы распознавания образов, Москва: Российская академия наук, 2021. — С. 239–244.

[3] *Чучупал В. Я.* Речевой корпус данных TeKoPyc. // Свидетельство о регистрации базы данных No 2005620205, Москва: Роспатент, 2005.

Block implementation of attention in Transformer for end-to-end speech recognition

Chuchupal Vladimir[1] v.chuchupal@mail.ru

[1]Moscow, Federal Research Center "Computer Science and Control" of Russian Academy of Sciences

The successful application of encoder-decoder models with attention is associated with a high computational load [1]. The report proposes new methods for the implementation of self-attention and attention for speech recognition applications, based on the properties of a speech signal. This allows significantly reduce the computational requirements in terms both of the amount of memory used and the number of operations.

In [2], an algorithm for calculation of self-attention in the Transformer encoder was proposed. The algorithm is based on the short-time property of speech signals and is economical in terms of memory and amount of calculations. The idea is approximate the result for the entire sequence of data by estimates obtained on short blocks. The algorithm has a computational complexity of the order of $O(n)$ both in memory and in the number of operations, compared to $O(n^2)$ for the conventional one.

In addition to self-attention in the encoder, the Transformer network also contains sublayers of attention and self-attention in the decoder, the computational requirements of which can also be reduced by an order of magnitude by taking into account the physical properties of speech.

Self-attention in the decoder acts like a language model, so one can use the short-term nature of the dependence between characters. In particular, one can put a length limit on context from the previous characters. It is natural to implement such limitation by using masking matrices of the band type, in contrast to the lower triangular matrices in the conventional approach. If one will follow the analogy with n-gram language models, the self-attention for each character in this case will have the same order.

More formally, the idea of a decoder self-attention computation is:

```
Input: sequence Q = q[0],...,q[|Q|-1] of language symbols,
n -- the block size (the language model order)
Put: K = V = Q
Calculate:
    Mask matrix:      M[i,j]= 1 if 0<= i-j <= n else 0, 0<= i,j <|Q|
    Attention scores: S  = b * matmul(tr(Q),K) * M
    Attention weights: A  = softmax (S)
    Attention:        Y  = matmul(A, V)
    Output:           OUT  = ReLu(Y)
```

Here matmul denotes the matrix multiplication operation, b = 1/sqrt(d), d – the size of a symbol model, ReLu – rectification activation.

The implementation of the attention computation in the decoder, similar to [2] is not obvious due to the lack of synchronization between decoder's symbols and the encoder's outputs. Encoder works step by step, so one can establish a correspondence between the symbols and the encoder blocks using the means of the block data. We attribute the symbol to that block in which the total value of the attention scores for the symbol is maximum. After a block has all of its symbols found, the attention is computed only on data of that block and for the selected symbols.

More formally, the idea of the algorithm is:

```
Input:
    q[i], i=0,...,|Q|-1 -- a sequence of symbols,
    as a matrix Q, where row Q[i,*] is embedding of symbol i.
    matrix K, an encoder output, as a sequence of N blocks K[i]:
    K = K[0],K[1]|, ... ,K[N-1]
Compute:
    Attention score matrix S:
    S = matmul(tr(Q),K),  S[i,j] is score of q[i] at block K[j]
For each i=0,...,|Q|-1:
    Find the number j of the block to which the symbol q[i] belongs:
    j = argmax_k S[i,k], 0 <= k < N.
For each j=0, ... ,N-1:
    Assemble the matrix Q[j] from Q[i,*]: q[i] belonging to K[j].
     Calculate using data of block K[j]:
        Attention scores:   S = matmul(tr(Q[j]),K[j])
            Attention weights:  A = softmax(S)
        Attention:          Y[j] = matmul(tr(Q[j]),A)
Compute the output as concatenation:
    Output = [Y[0]|Y[1]|....|Y[N-1]]
```

Since all blocks have the same fixed size, the computational complexity in terms of memory and the amount of calculations (for the sequential execution) of all above described methods is $O(n)$.

Computational experiments was performed on the TeCoRus [3] speech corpus, the part consisted of digital sequences. For training, 16357 utterances were used, spoken by 155 speakers. Testing was carried out on 186 utterances from new speakers. We used the standard parametrization of speech based on mel-spectral features. The language consisted of letters and letter combinations, punctuation symbols. Regular words at the output of the decoder are obtained by concatenating the decoded symbols until the word feature appears.

The end-to-end speech recognition system is implemented as a Transformer with a two-layer encoder and decoder. Cross-entropy was used as the loss function. Opti-

mization method – Adam. A linear decrease in the learning rate factor (from initial 0.003) with a factor of 0.96 every 300 steps was used. When choosing a packet size of 150 sentences or more, the loss function decreased almost monotonously, the loss almost vanished in 300 epochs.

The table 1 shows the word recognition acciracy vs. sizes of the context blocks.

Obviously, the use of the block algorithm did not lead to a deterioration in the recognition accuracy. The accuracy values for different block sizes and contexts differ insignificantly (within 1-2 percent). For self-attention in encoder, the maximum accuracy is observed at a block size of one second, accordingly to the assumption that the size of the context should correspond to the duration of recognized sounds. For the self–attention in decoder, the best duration of the left context of the language model, appears to be different from the maximum possible and in this case is 5 characters. It is explained by the size of the training dataset.

Table 1. Dependence of the accuracy of recognition on the size of blocks of contexts of attention and self-attention. L,C,R denote the lengths of the left, center and right context blocks, in seconds or symbols. The notation inf corresponds to the maximum context (generally accepted method) length.

Encoder			DeCoder			Accuracy
S–Atten.(sec.)			S–Atten.(sym.)		Atten.(sec.)	
C	L	R	C	L	C	
inf	inf	inf	inf	inf	inf	0.942
2.0	1.0	1.0	inf	inf	inf	0.956
1.0	0.5	0.5	inf	inf	inf	0.966
1.0	0.	0.	15	15	inf	0.939
1.0	0.	0.	7	7	inf	0.942
1.0	0.	0.	5	5	inf	0.976
1.0	0.	0.	5	5	8.	0.971
1.0	0.	0.	15	7	4.	0.972
1.0	0.	0.	15	7	2.	0.941

[1] *Chuchupal V. J.* Acoustic and language modeling in end-to-end speech recognition systems. // Digital Signal Processing, Moscow: Popov A.S Russian Science and Technical Society on Radio, Electronic and Communitations, 2020. —No 4, — Pp. 34–43.

[2] *Chuchupal V. J.* A computationally-effective transformer model for acoustic speech modeling // Abstracts of 20th All–Russian conference on Mathematical Methods of Pattern Recognition, Moscow: Russian academy of sciences, 2021. — Pp. 239–244.

[3] *Chuchupal V., Makovkin K., Chichagov A A., Kouznetsov V., Ogaryshev V.* Speech data corpus TeCoRus. // Moscow: RosPatent, registration certificate 200562020, 2005.

Декомпозиции обучения в пространстве признаков для задачи распознавания лиц на изображениях

Таранов Сергей Константинович[1]⋆ taranov.sk@gmail.com
Гнеушев Александр Николаевич[1,2] gneushev@ccas.ru
[1]Москва, Московский физико-технический институт (НИУ)
[2]Москва, Федеральный исследовательский центр "Информатика и управление"РАН

Рассматривается задача распознавания изображений лиц, востребованная и активно исследуемая область компьютерного зрения. Идентификация лиц относится к задаче многоклассовой классификации, при этом можно выделить две подзадачи: выделение признаков на изображении и классификация на их основе. Современным подходом является использование нейросетевых моделей для совместного извлечение признаков и последующей классификации.

Распознавание лиц является задачей принятия решения о принадлежности классам, неизвестным на этапе обучения. В процессе обучения оптимизируется функция извлечения признаков, а функция классификации строится на основе данных, используемых в конкретном практическом приложении. Цель обучения — получение обобщенного пространства признаков, пригодного для разделения векторов признаков изображений лиц людей, не участвовавших в обучении. Основная проблема — одновременная разделимость межклассовых векторов и компактность внутриклассового представления.

Используемые функции потерь можно разделить на два типа: классификационные (multi-class classification loss) [1] и на основе попарного сравнения (pair-based loss) [2]. Первые максимизируют принадлежности изображений к своему классу и минимизируют принадлежности к остальным, используются для обучения полной модели состоящей из функции извлечения признаков и классификатора. Во втором случае оптимизируется только функция извлечения признаков, для чего используется критерий максимизации схожести между векторами признаков, соответствующих одному классу, и минимизации схожести для векторов разных классов. В работе предлагается метод декомпозиции сверточной нейросетевой модели и ее поэтапного обучения с функциями потерь обоих типов для задачи распознавания изображений.

Пусть даны изображения лиц $I = \{I_k\}_{k=1..N}$, N – размер выборки, и $Y = = \{y_k\}_{k=1..N}$, $y \in 1,..M$ – разметка, номера классов, к которым принадлежат изображения $\{I_k\}$ лиц из обучающей выборки, M - общее число классов, задающих различных людей. Определим классифицирующую нейросетевую модель с помощью функции $Y = F(I, \mathbf{W})$, W параметры модели. Тогда $F(I, \mathbf{W}) = = (\Phi \circ \Psi)(I, \mathbf{W}) = \Phi(\Psi(I, \theta), \mathbf{w})$, где $\mathbf{W} = \{\mathbf{w}, \theta\}$ - совокупность параметров модели, $\Psi(I, \theta)$ – функция выделения признаков с параметрами $\theta \subset \mathbb{R}^{D_\theta}$, $\Phi(X, \mathbf{w})$ – функция классификации признаков с параметрами $\mathbf{w} \subset \mathbb{R}^{D_\mathbf{w}}$. $X = \Psi(I, \theta), X = = \{x_k\}_{k=1,..,N}, x_k \in \mathbf{X} \subset \mathbb{R}^{D_x}$ – множество векторов признаков изображений, $\mathbf{X}$ – пространство векторов признаков; $D_\theta, D_\mathbf{w}, D_x$ – размерности элементов

множеств θ, w, x соответственно. В качестве функции Φ будем рассматривать линейный классификатор с векторами весов $w = \{w_m\}_{m=1,..,M}$.

Предлагается обучать нейросетевую модель двумя этапами.На первом этапе производится обучение функции F со случайной инициализации весов с помощью классификационной функции потерь CosFace Loss [1], вида:

$$L = -\frac{1}{N}\sum_{k=1}^{N}\log\frac{\exp\left(\cos\left(\mu_1\tau_{k,y_k}+m_a\right)\right)}{\exp\left(\cos\left(\mu_1\tau_{k,y_k}+m_a\right)\right)+\sum_{j=1,j\neq y_k}^{M}\exp\left(\cos\left(\tau_{k,j}\right)\right)}, \qquad (1)$$

где τ_{k,y_m} – угол между нормированными векторами $\hat{w}_m = \frac{\mathbf{w_m}}{\|\mathbf{w_m}\|}$, $\hat{x}_k = \frac{\mathbf{x_k}}{\|\mathbf{x_k}\|}$; μ_1, m_a – гиперпараметры метода. Благодаря виду функции потерь структура пространства $\mathbf{X}$ формируется на каждой итерации глобально сразу для всех классов. Модель, полученная после первого этапа, способна решать задачу распознавания.

На втором этапе рассмотрим декомпозиции $\Psi = (\psi_K \circ ... \circ \psi_1)$, которую разобьем на две группы $\{\psi_j\}_{j=1..k}, \{\psi_j\}_{j=k+1..K}$. Параметры первой фиксируются, параметры второй обучаются с помощью используя функцию потерь на основе попарных сравнений MultiSimilarity Loss [2] вида:

$$L_{MS} = \frac{1}{m^{'}}\sum_{k=1}^{m^{'}}\frac{1}{\alpha}\log[1+\sum_{r\in P_k}\exp(-\alpha S_{\mathbf{x_k},\mathbf{x_r}})] + \frac{1}{\beta}\log[1+\sum_{r\in N_k}\exp(\beta S_{\mathbf{x_k},\mathbf{x_r}})], \quad (2)$$

где $m^{'}$ – количество векторов в одной итерации обучения; $S_{\mathbf{x_k},\mathbf{x_r}} = \langle x_k, x_r\rangle - \lambda$ – мера схожести векторов $\hat{x}_k, \hat{x}_r$; α, β, λ – гиперпараметры метода; P_k, N_k – множества векторов совпадающих и отличающихся по классовой принадлежности для вектора x_k соответственно. Составление P_k, N_k – сложная задача, от них сильно зависит итоговое качество. Эксперименты показали, что наилучшим решением является построение множеств с наиболее сложными примерами, то есть пар векторов с максимальным внутриклассовым расстоянием и минимальным межклассовым. Существует два основных подхода: предварительный расчёт этих множеств по всей базе и итерационный перерасчёт во время обучения на основе мини-батча, могут применятся и смешанные решения Первый – существенно точнее и быстрее, но быстро теряет актуальность, так как структура пространства изменяется и требует перерасчёт. Второй менее точен и обеспечивает более медленное обучение, но актуален во время всего обучения. Предлагаемая декомпозиция и, как следствие, обучение части модели, сохранет общую структуру пространства признаков, что позволяет использовать только предварительный расчёт, дающий результаты лучше, чем метод, предложенный в [2].

В работе используется модель на основе глубокой свёрточной сети Inception-Resnet-v2. Исходя из блочной структуры Inception-Resnet-v2, предлагается рассматривать в качестве элементарных функций $\{\psi_j\}_{j=1..k}$ отдельные блоки как на схеме (Рис. 1).

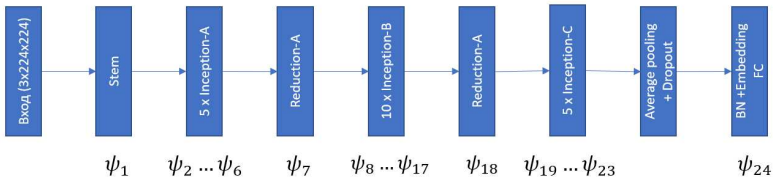


Рис. 1. Схема декомпозиции функции выделения признаков на части

В процессе обучения гиперпараметры базовых методов (1,2) не изменялись и соответсвуют предложенным их авторами [1,2]. Обучающая выборка — открытый датасет Trillion-Pairs DeepGlint представленный в соревновании Trillionpairs, содержащий 180855 классов и 6753545 изображений и состоящий из датасетов баз MS-Celeb-1M-v1c и Asian-Celeb. Для тестирования использовалась отдельная база, состоящая из приблизительно 110000 классов, из которой примерно 450 классов, содержащих в среднем по 25 изображений, использовались для теста идентификации, по сценарию сравнения $1 : k$. Результаты теста идентификации для декомпозиции $\Psi_{1,k}, k = \{7, 18, 23, 24\}$ приведены в таблице ниже.

Зафиксированная часть модели	True Positive Rate		
	FMR=10^{-8}	FMR=10^{-7}	FMR=10^{-6}
$\Psi_{1,24}$ (базовая модель)	0.370	0.485	0.603
$\Psi_{1,7}$	0.316	0.451	0.581
$\Psi_{1,18}$	0.373	0.502	0.629
$\Psi_{1,23}$	**0.426**	**0.557**	**0.669**

Эксперименты показали, что предложенный метод декомпозиции и поэтапного обучения оптимизирует локальную структуру пространства признаков без разрушения глобальных связей, что приводит к увеличению точности распознавания изображений лиц.

Работа поддержана грантом РФФИ No.21-51-53019

[1] *Wang H., Wang Y., Zhou Z. et al.* Cosface: Large margin cosine loss for deep face recognition// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition — 2018. — June. — Pp. 5265–5274.

[2] *Wang X., Han X., Huang W. et al.* Multi-similarity loss with general pair weighting for deep metric learning// 019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). — 2019. — Pp. 5017–5025

Face recognition feature space learning decomposition

Taranov Sergei[1]⋆ taranov.sk@phystech.edu
Gneushev Alexander[1,2] gneushev@ccas.ru
[1]Moscow, Moscow Institute of Physics and Technology
[2]Moscow, Federal Research Center "Computer Science and Control" of RAS

The problem of face image recognition, a highly demanded and actively researched area of computer vision, is considered. Face identification is multiclass classification task, and two subtasks can be distinguished: image features extraction and features classification. A modern approach is to use neural network models for joint feature extraction and subsequent classification.

Face recognition is open-set recognition task. During training process the features extraction function is optimized and the features classification function is built based on the data used in a particular practical application. The training goal is obtaining a generalized feature vector space suitable for separating feature vectors of images of faces of people who did not participate in training. The main problem is the simultaneous separability of interclass vectors and the compactness of the intraclass representation.

The used loss functions can be divided into two types: classification (multiclass classification loss) [1] and based on pairwise comparison (pair-based loss) [2]. Classification losses maximize the classification score of images to their class and minimize classification scores to the other classes. Those losses are used for complete model training consisting of a feature extraction function and a classifier. In the second case, only the feature extraction function is optimized, for which the criterion of maximizing the similarity between feature vectors corresponding to the same class and minimizing similarity for vectors of different classes is used. The paper proposes a method for decomposition of a convolutional neural network model and its stage-by-stage training with loss functions of both types for the problem of image recognition.

Let's consider face images $I = \{I_k\}_{k=1..N}$, N is the sample size, and $Y = \{y_k\}_{k=1..N}$, $y \in 1,..M$ is labeling, numbers of classes to which images $\{I_k\}$ of faces from the training sample belong, M is total number of classes defining different people.

Let's define a classifying neural network model using the function $Y = F(I, \mathbf{W})$, W model parameters. Then $F(I, \mathbf{W}) = (\Phi \circ \Psi)(I, \mathbf{W}) = \Phi(\Psi(I, \theta), \mathbf{w})$, where $\mathbf{W} = \{\mathbf{w}, \theta\}$ is the set of model parameters, $\Psi(I, \theta)$ is feature extraction function with parameters $\theta \subset \mathbb{R}^{D_\theta}$, $\Phi(X, \mathbf{w})$ is feature classification function with parameters $\mathbf{w} \subset \mathbb{R}^{D_\mathbf{w}}$. $X = \Psi(I, \theta), X = \{x_k\}_{k=1,..,N}, x_k \in \mathbf{X} \subset \mathbb{R}^{D_x}$ is set of image feature vectors, $\mathbf{X}$ is space of feature vectors; $D_\theta, D_\mathbf{w}, D_x$ are the dimensions of the elements of the sets θ, w, x,respectively. As a function Φ let's consider a linear classifier with weight vectors $w = \{w_m\}_{m=1,..,M}$.

It is proposed to train the neural network model in two stages. At the first stage, the function F is trained with random initialization of weights using the classification loss function CosFace Loss [1], of the form:

$$L = -\frac{1}{N}\sum_{k=1}^{N}\log\frac{\exp\left(\cos\left(\mu_1\tau_{k,y_k}+m_a\right)\right)}{\exp\left(\cos\left(\mu_1\tau_{k,y_k}+m_a\right)\right)+\sum_{j=1,j\neq y_k}^{M}\exp\left(\cos\left(\tau_{k,j}\right)\right)}, \quad (1)$$

where τ_{k,y_m} is the angle between normalized vectors $\hat{w}_m = \frac{\mathbf{w_m}}{\|\mathbf{w_m}\|}$, $\hat{x}_k = \frac{\mathbf{x_k}}{\|\mathbf{x_k}\|}$; μ_1, m_a – method hyperparameters. Due to the form of the loss function, the feature space structure $\mathbf{X}$ is formed at each iteration globally for all classes at once. The model obtained after the first stage can be used for face recognition.

At the second stage, we consider the decomposition $\Psi = (\psi_K \circ ... \circ \psi_1)$, which we divide into two groups $\{\psi_j\}_{j=1..k}, \{\psi_j\}_{j=k+1..K}$. The parameters of the first group are fixed, the parameters of the second group are trained using the loss function based on pairwise comparisons MultiSimilarity Loss [2] of the form:

$$L_{MS} = \frac{1}{m^{'}}\sum_{k=1}^{m^{'}}\frac{1}{\alpha}\log[1+\sum_{r\in P_k}\exp(-\alpha S_{\mathbf{x_k},\mathbf{x_r}})] + \frac{1}{\beta}\log[1+\sum_{r\in N_k}\exp(\beta S_{\mathbf{x_k},\mathbf{x_r}})], \quad (2)$$

where $m^{'}$ is the number of vectors in one learning iteration; $S_{\mathbf{x_k},\mathbf{x_r}} = \langle x_k, x_r\rangle - \lambda$ is similarity of vectors $\hat{x}_k, \hat{x}_r$; α, β, λ are method hyperparameters; P_k, N_k are sets of vectors that match and differ in class for the vector x_k, respectively. Selection of P_k, N_k is a difficult task, it strongly effect the model quality. It is demonstrated by different experiments that the best solution is to build sets with the most difficult examples, that are pairs of vectors with the maximum intraclass distance or the minimum interclass distance. There are two main approaches of P_k, N_k selection: pre-calculation of these sets over the entire database and iterative recalculation during training based on a mini-batch, mixed solutions can also be applied. The first one is much more accurate and faster, but quickly loses its relevance, since the structure of space changes and requires recalculation. The second is less accurate and provides slower learning convergence, but is relevant during the entire learning. The proposed decomposition and, as a result, training of a part of the model will preserve the overall structure of the feature space, which allows using only a preliminary calculation that gives better results than the method proposed in [2].

In this paper a used model isbased on the deep convolutional network Inception-Resnet-v2. Based on the block structure of Inception-Resnet-v2, it is proposed to consider individual blocks as elementary functions $\{\psi_j\}_{j=1..k}$ as in the diagram (Fig. 1).

During the learning process, the hyperparameters of the basic methods (1,2) did not change and correspond to those proposed by their authors [1,2]. The training

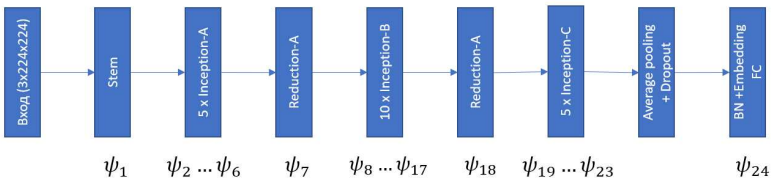


Fig. 1. Feature extraction function decomposition scheme

dataset is an public available Trillion-Pairs DeepGlint dataset presented in the Trillionpairs competition, containing 180855 classes and 6753545 images and consisting of MS-Celeb-1M-v1c and Asian-Celeb datasets. For testing, an another database was used, consisting of approximately 110,000 classes, of which approximately 450 classes containing an average of 25 images were used for the identification test, according to the $1 : k$ comparison scenario. The results of the identification test for the decomposition $\Psi_{1,k}, k = \{7, 18, 23, 24\}$ are shown in the table below.

Fixed model parts	True Positive Rate		
	FMR=10^{-8}	FMR=10^{-7}	FMR=10^{-6}
$\Psi_{1,24}$ (base model)	0.370	0.485	0.603
$\Psi_{1,7}$	0.316	0.451	0.581
$\Psi_{1,18}$	0.373	0.502	0.629
$\Psi_{1,23}$	**0.426**	**0.557**	**0.669**

Experiments have demonstrate that the proposed decomposition 2 stage learning method optimizes the local structure of the feature space without global structure transformation, which result in accuracy increase for face image recognition task.

This research is funded by RFBR, grant 21-51-53019

[1] *Wang H., Wang Y., Zhou Z. et al.* Cosface: Large margin cosine loss for deep face recognition// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition — 2018. — June. — Pp. 5265–5274.

[2] *Wang X., Han X., Huang W. et al.* Multi-similarity loss with general pair weighting for deep metric learning// 019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). — 2019. — Pp. 5017–5025

Patch2Vec: простой и эффективный алгоритм свёртки для мобильных нейронных сетей

Брыкин Глеб Сергеевич[1]⋆ glebbrykin@colorfulsoft.ru

[1]Москва, МГТУ им. Н.Э.Баумана

Развитие технологий машинного обучения сделало возможной автоматизацию решения множества задач из самых разных областей. Одним из наиболее перспективных направлений являются искусственные нейронные сети, нашедшие применение во множестве прикладных задач. Расширяющийся круг применений ИНС, а также связанная с этим потребность в переносе соответствующего ПО на ПК, мобильные и встраиваемые устройства, требуют разработки новых эффективных вычислительных алгоритмов. Операция свёртки является основой современных нейронных сетей, применяемых в задачах обработки звука, изображений и видео; причём вычислительная сложность архитектур свёрточных нейронных сетей определяется в основном свёрточными слоями. Целью данной работы является разработка простого универсального алгоритма свёртки, обеспечивающего оптимальное соотношение эффективности по времени и памяти.

Предлагаемый в данной работе алгоритм (Patch2Vec) разработан с учётом общих особенностей аппаратной составляющей современных ЭВМ, за счёт чего достигается прирост скорости выполнения по сравнению с не оптимизированным алгоритмом при сопоставимой вычислительной сложности. Ключевым отличием Patch2Vec от аналогичных алгоритмов, использующих схожие подходы к оптимизации, является существенно сниженный размер буфера временных данных и отсутствие связи между размером буфера и размером входного и выходного изображения. В отличие от некоторых наиболее быстрых алгоритмов (быстрой свёртки методом Ш. Винограда [1] и др.), Patch2Vec является универсальным, т.е. не требует специфичных реализаций функции свёртки для различных гиперпараметров. Patch2Vec выполняет свёртку с любыми допустимыми значениями гиперпараметров.

Основной идеей Patch2Vec является минимизация числа непоследовательных обращений к адресам оперативной памяти за счёт перегруппировки значений входного изображения таким образом, чтобы все значения, необходимые для применения ядер свёртки, были расположены в памяти последовательно, что позволит задействовать возможности механизма опережающего считывания процессора и снизить влияние ограниченной пропускной способности оперативной памяти на выполнение операции свёртки. Аналогичный подход используется в алгоритмах, основанных на преобразованиях Im2Col и Im2Row [2]. Im2Col и Im2Row формируют матрицу, каждый столбец или строка которой содержит все необходимые для выполнения свёртки в данной позиции ядра значения входного изображения (патчи). Затем данная матрица умножается на матрицу ядер свёртки, в результате образуя матрицу выходного изображения.

Можно заметить, что эти алгоритмы требуют память для хранения $H_{out} \cdot W_{out} \cdot \cdot C_{in} \cdot H_k \cdot W_k$ значений временной матрицы, где H_{out} – высота выходного изображения; W_{out} – ширина выходного изображения; C_{in} – количество каналов входного изображения; H_k – высота ядра свёртки; W_k – ширина ядра свёртки. Patch2Vec не сохраняет все патчи входного изображения в матрицу: вместо этого на каждой позиции ядра свёртки из входного изображения извлекается патч и записывается в память в виде вектора, который затем умножается на матрицу весов, формируя вектор значений каналов соответствующей точки выходного изображения, компоненты которого сразу же записываются в соответствующие адреса памяти. Такой подход позволяет отвязать размер буфера (вектора) от размеров изображений, таким образом, для выполнения операции методом Patch2Vec требуется память для хранения $C_{in} \cdot H_k \cdot W_k$ значений, что много меньше, чем в случае Im2Col и Im2Row. Patch2Vec, ровно как и Im2Col может быть эффективно оптимизирован с использованием многопоточности и векторных инструкций процессора.

В ходе сравнения производительности были получены данные, однозначно указывающие на эффективность предлагаемого метода. Некоторые результаты представлены в таблице:

Условия	"Наивный" алгоритм		Im2Col		Patch2Vec	
f64k3s1 1b64c256h256w	100%	0 б	**29%**	144 Мб	36%	~ **2 Кб**
f64k5s1 1b64c256h256w	100%	0 б	**37%**	394 Мб	42%	~ **6 Кб**
f64k7s1 1b64c256h256w	100%	0 б	**40%**	760 Мб	45%	~ **12 Кб**
f64k9s1 1b64c256h256w	100%	0 б	**39%**	1.2 Гб	49%	~ **20 Кб**
f128k3s1 1b64c1024h1024w	100%	0 б	32%	2.25 Гб	**27%**	~ **2 Кб**
f128k5s1 1b64c1024h1024w	100%	0 б	–	>4 Гб, OoM	**41%**	~ **6 Кб**

В результате проделанной работы был разработан новый эффективный алгоритм свёртки, превосходящий многие известные аналоги. Предлагаемый алгоритм был успешно применён на практике в приложении DeOldify.NET, сократив время обработки изображения на 7% по сравнению со свёрткой на базе Im2Col.

[1] *Lavin, Scott Gray.* Fast Algorithms for Convolutional Neural Networks // arXiv preprint arXiv:1509.09308, 2015.

[2] *Anton V. Trusov, Elena E. Limonova, Dmitry P. Nikolaev and Vladimir V. Arlazarov.* p-im2col: Simple Yet Efficient Convolution Algorithm With Flexibly Controlled Memory Overhead // IEEE Access, vol. 9, 2021. — P. 168162-168184.

Patch2Vec: a simple and efficient convolution algorithm for mobile neural networks

Brykin Gleb[1]⋆ glebbrykin@colorfulsoft.ru

[1] Moscow, Bauman Moscow State Technical University

The development of machine learning technologies has made it possible to automate the solution of many tasks from a variety of fields. One of the most promising areas is artificial neural networks, which have found application in a variety of applied tasks. The expanding range of applications of the ANN, as well as the related need to transfer the corresponding software to PCs, mobile and embedded devices, require the development of new efficient computational algorithms. The convolution operation is the basis of modern neural networks used in audio, image and video processing tasks; moreover, the computational complexity of convolutional neural network architectures is determined mainly by convolutional layers. The purpose of this work is to develop a simple universal convolution algorithm that provides an optimal ratio of efficiency in time and memory.

The algorithm proposed in this paper (Patch2Vec) is developed taking into account the general features of the hardware component of modern computers, due to which an increase in execution speed is achieved compared to an un-optimized algorithm with comparable computational complexity. The key difference between Patch2Vec and similar algorithms using similar optimization approaches is the significantly reduced size of the temporary data buffer and the absence of a connection between the buffer size and the size of the input and output images. Unlike some of the fastest algorithms (fast convolution by Sh. Vinograd [1], etc.), Patch2Vec is universal, i.e. it does not require specific implementations of the convolution function for various hyperparameters. Patch2Vec performs convolution with any valid hyperparameter values.

The main idea of Patch2Vec is to minimize the number of inconsistent accesses to RAM addresses by rearranging the values of the input image so that all the values used by convolution kernels are sequentially located in memory, which will enable the use of the processor's advanced readout mechanism and reduce the impact of limited RAM bandwidth on the execution of the convolution operation. A similar approach is used in algorithms based on Im2Col and Im2Row [2] transformations. Im2Col and Im2Row form a matrix, each column or row of which contains all the values of the input image (patches) necessary to perform convolution at a given kernel position. Then this matrix is multiplied by the matrix of convolution kernels, resulting in the matrix of the output image. It can be noted that these algorithms require memory to store $H_{out} \cdot W_{out} \cdot C_{in} \cdot H_k \cdot W_k$ values of the temporary matrix, where H_{out} is the height of the output image, W_{out} is the width of the output image, C_{in} is the number of channels of the input image, H_k is the height of the convolution kernel, W_k is the width of the convolution kernel. Patch2Vec does not save all the patches of the input image to the matrix: instead, at each position of the convolution kernel,

a patch is extracted from the input image and written to memory as a vector, which is then multiplied by a matrix of weights, forming a vector of channel values of the corresponding point of the output image, the components of which are immediately written to the corresponding memory addresses. This approach allows you to untie the size of the buffer (vector) from the size of the images, thus, to perform the Patch2Vec operation, memory is required to store $C_{in} \cdot H_k \cdot W_k$ values, which is much less than in the case of Im2Col and Im2Row. Patch2Vec, just like Im2Col, can be efficiently optimized using multithreading and vector processor instructions.

During the performance comparison, data were obtained that clearly indicate the effectiveness of the proposed method. Some results are presented in the table:

Conditions	"Naive" algorithm		Im2Col		Patch2Vec	
f64k3s1 1b64c256h256w	100%	0 b	**29%**	144 Mb	36%	~ **2 Kb**
f64k5s1 1b64c256h256w	100%	0 b	**37%**	394 Mb	42%	~ **6 Kb**
f64k7s1 1b64c256h256w	100%	0 b	**40%**	760 Mb	45%	~ **12 Kb**
f64k9s1 1b64c256h256w	100%	0 b	**39%**	1.2 Gb	49%	~ **20 Kb**
f128k3s1 1b64c1024h1024w	100%	0 b	32%	2.25 Gb	**27%**	~ **2 Kb**
f128k5s1 1b64c1024h1024w	100%	0 b	–	>4 Gb, OoM	**41%**	~ **6 Kb**

As a result of the work done, a new efficient convolution algorithm was developed that surpasses many well-known analogues. The proposed algorithm has been successfully applied in practice in the DeOldify.NET application, reducing the image processing time by 7% compared to the convolution based on Im2Col.

[1] *Lavin, Scott Gray.* Fast Algorithms for Convolutional Neural Networks // arXiv preprint arXiv:1509.09308, 2015.

[2] *Anton V. Trusov, Elena E. Limonova, Dmitry P. Nikolaev and Vladimir V. Arlazarov.* p-im2col: Simple Yet Efficient Convolution Algorithm With Flexibly Controlled Memory Overhead // IEEE Access, vol. 9, 2021. — P. 168162-168184.

Трансформерная языковая модель ruSciBERT для векторизации и обработки научных текстов на русском языке

Герасименко Николай Александрович[1,2]⋆ nikgerasimenko@gmail.com
Чернявский Александр Сергеевич[2,3] alschernyavskiy@gmail.com
Никифорова Мария Андреевна[2,3] labenzom@gmail.com
Воронцов Константин Вячеславович[1] vokov@forecsys.ru
[1]Москва, ФИЦ ИУ РАН
[2]Москва, ПАО Сбербанк
[3]Москва, НИУ «Высшая школа экономики»

Значительный рост числа научных публикаций и количества научных отчетов делает задачу их обработки и анализа сложной и трудозатратной. Языковые модели, основанные на архитектуре Трансформер и предобученные на больших текстовых коллекциях, позволяют качественно решать множество задач анализа текстовых данных. Для работы с научными текстами на английском языке существуют модели SciBERT [1] и ее модификация SPECTER [2], однако они не поддерживают русский язык в связи с малым количеством текстов в обучающей выборке. Кроме того, способ оценки качества языковых моделей для научных текстов, бенчмарк SciDocs, также поддерживает только английский язык. Предлагаемая модель ruSciBERT позволит решать широкий спектр задач, связанных с анализом научных текстов на русском языке, а прилагаемый к ней бенчмарк ruSciDocs позволит оценивать качество языковых моделей применительно к этим задачам.

SciBERT является языковой моделью, обученной на многодоменном корпусе научных статей, написанных преимущественно на английском языке. Предлагается дообучить базовую модель BERT на задаче предсказания маскированных токенов. Результаты, полученые авторами на нескольких задачах классификации и распознавания именованных сущностей (NER) для научных статей, значительно превосходят результаты базовой модели. Дополнительно обученный токенизатор позволил улучшить качество языковой модели. Мы используем похожие идеи и предлагаем дообучение модели RoBERTa на русскоязычном корпусе научных текстов с собственным токенизатором. Данную модель мы называем RuSciBERT. В качестве базовой модели нами была выбрана RoBERTa в связи с тем, что она обучена на расширенном количестве данных, большем количестве задач и достигла лучших результатов по сравнению с базовым BERT.

SciDocs является бенчмарком для оценки качества семантических векторных представлений, получаемых с помощью языковых моделей. Он включает в себя четыре типа задач: классификация на основе классификаторов MAG и MeSH; предсказание цитирования на основе Semantic Scholar Academic Graph; предсказание активности пользователей Semantic Scholar; рекомендации статей похожих на статью-запрос. Все задачи, кроме задачи классификации, сформулированы как задачи ранжирования.

ruSciBERT планируется обучить на датасете, включающем около 1 млрд токенов. Данные для обучения собраны из датасета Semantic Scholar Academic Graph, аннотаций отчетов с сайта ЕГИСУ НИОКТР, а также работ из системы ИСТИНА МГУ и других открытых источников. Размер словаря токенизатора в нашем случае равен 50265 по аналогии с базовой моделью RoBERTa.

Бенчмарк ruSciDocs планируется составить из задач, аналогичных задачам оригинального SciDocs: классификация текстов по категориям Microsoft Academic Graph и OECD из ЕГИСУ НИОКТР, государственного сайта для учета научно-исследовательских работ; предсказание цитирования на основе данных из Semantic Scholar Academic Graph и других открытых источников.

На данный момент мы обучили модель RuSciBERT на датасете в 300 млн токенов на двух эпохах. Она показывала хорошие результаты при заполнении пробелов в текстовых фразах, а также гораздо более низкий уровень перплексии на отложенной выборке по сравнению с общей языковой моделью ruBERT, обученной на текстах всех тематик. Так, RuSciBERT имеет перплексию 4.81, причем она монотонна снижается на последних шагах обучения, в следствии чего модель можно дообучать дальше. В то же время ruBERT имеет перплексию 9.64.

Примеры работы нашей модели заполнения маскированных токенов показаны ниже. В них маскированные токены обозначены через □, а модель предсказывает три наиболее вероятных варианта токенов для замены.

— «при использовании в усилителе мощности адаптивной измерительной □ появится возможность» → «системы», «аппаратуры», «станции»
— «указанные оппоненты не имеют □ проектов и публикаций с соискателем» → «совместных», «собственных», «аналогичных»
— «новый метод управления □ характеристиками ао фильтров» → «техническими», «технологическими», «функциональными»

RuBERT также показывает неплохие результаты, но некоторые из его вариантов заполнения являются менее удачными. Так, в первом примере среди предсказанного множество токенов есть «технологии» (меньше подходит по смыслу чем остальные варианты), во втором — «своих», а в третьем — «всеми» (возможные варианты, но более общие и поэтому менее качественные).

Основываясь на текущих результатах, можно предположить, что ruSciBERT, обученный на датасете в 1 млрд токенов, покажет наилучшие разультаты на бенчмарке ruSciDocs по сравнению с другими существующими подходами.

[1] *Iz Beltagy and Kyle Lo and Arman Cohan.* SciBERT: Pretrained Language Model for Scientific Text // EMNLP, 2019.

[2] *Arman Cohan and Sergey Feldman and Iz Beltagy and Doug Downey and Daniel S. Weld.* SPECTER: Document-level Representation Learning using Citation-informed Transformers // ACL, 2020.

Transformer-based Language Model ruSciBERT for Russian Scientific Document Representations and Processing

Gerasimenko Nikolai[1,2]⋆ nikgerasimenko@gmail.com
Chernyavskiy Alexander[2,3] alschernyavskiy@gmail.com
Nikiforova Maria[2,3] labenzom@gmail.com
Vorontsov Konstantin[1] vokov@forecsys.ru
[1]Moscow, FRC CSC RAS
[2]Moscow, Sberbank
[3]Moscow, Higher School of Economics

The significant increase in the number of scientific publications and scientific reports makes their processing and analysis difficult and time-consuming. Transformer-based language models pretrained on large text collections allow solving various NLP tasks. For instance, there are SciBERT [1] and SPECTER [2] models for working with scientific texts in English, but they do not support Russian due to the small number of such texts in the training sample. In addition, the SciDocs benchmark for scientific text representations also supports only English. The proposed ruSciBERT model will allow solving a wide range of tasks related to the analysis of scientific texts in Russian. Despite the model, we propose the benchmark ruSciDocs to evaluate the quality of language models on these tasks.

SciBERT is a language model pretrained on a multidomain corpus of scientific articles written primarily in English. Here, authors suggested to train the base BERT model for the problem of predicting masked tokens for this corpus. The results obtained on several classification and named entities recognition (NER) tasks for scientific articles significantly outperform the results of the base BERT model. The tokenizer, additionally trained on the corpus, improved the quality of the language model. Similarly, we suggest training of the RoBERTa model on the Russian-language corpus of scientific texts with its own tokenizer. This model is called RuSciBERT. We selected RoBERTa as the base model because it was trained on the extended dataset and achieved better results compared to BERT.

SciDocs is a benchmark for assessing the quality of representations obtained with language models. It includes four types of tasks: classification based on MAG and MeSH classifiers; citation prediction based on the Semantic Scholar Academic Graph; prediction of Semantic Scholar activity; article-query recommendations. All tasks except the classification task are formulated as ranking problems.

RuSciBERT is planned to be trained on a dataset that includes about 1 billion tokens. The data for training is collected from the Semantic Scholar Academic Graph, abstracts of reports from the `rosrid.ru` website, `istina.msu.ru`, and other open sources. The dictionary size of the tokenizer in our case is 50265, similar to the basic RoBERTa model.

The ruSciDocs benchmark is planned to be composed of tasks similar to those of the original SciDocs: classification of texts by categories Microsoft Academic Graph

and OECD from `rosrid.ru`, the public site for research work accounting; Prediction of citation based on data from the Semantic Scholar Academic Graph and other open sources.

As yet we have trained the RuSciBERT model on a 300 million-token dataset in two epochs. It showed good results in filling gaps in phrases, as well as a much lower level of perplexity on the validation sample compared to the general language model ruBERT, trained on texts of all subjects. RuSciBERT has 4.81 perplexity, and it decreases monotonously at the last steps of training. As a result, the model can be improved further. At the same time, ruBERT has 9.64 perplexity.

Examples of how ruSciBERT works on the "fill-mask" task are shown below. In these, masked tokens are marked with square □, and the model predicts the three most likely token replacements.

— «при использовании в усилителе мощности адаптивной измерительной □ появится возможность» → «системы», «аппаратуры», «станции»
— «указанные оппоненты не имеют □ проектов и публикаций с соискателем» → «совместных», «собственных», «аналогичных»
— «новый метод управления □ характеристиками ао фильтров» → «техническими», «технологическими», «функциональными»

RuBERT also shows good results, but some of them are less successful. For instance, in the first example among the predicted set of tokens there are «технологии» (less suitable than the other options), in the second - «всех» and in the third - «всеми» (possible options, but more general and therefore less qualitative).

Based on current results, ruSciBERT, trained on a 1 billion token dataset, will show the best results on the ruSciDocs benchmark compared to other existing approaches.

[1] *Iz Beltagy and Kyle Lo and Arman Cohan.* SciBERT: Pretrained Language Model for Scientific Text // EMNLP, 2019.

[2] *Arman Cohan and Sergey Feldman and Iz Beltagy and Doug Downey and Daniel S. Weld.* SPECTER: Document-level Representation Learning using Citation-informed Transformers // ACL, 2020.

Контролируемая генерация графов

Бишук Антон Юрьевич[1]⋆ bishuk.ayu@phystech.edu
Зухба Анастасия Викторовна[1] azukhba@gmail.com
[1]Москва, Московский физико-технический институт

В последнее время набирает популярность представление данных в виде графов, которые используют не только признаковые описания объектов-вершин, но и взаимосвязь между ними. При разработке алгоритмов машинного обучения на данных такой структуры проблема разнообразия датасетов стоит еще более остро, поэтому важной является задача генерации синтетических графов с заданными свойствами.

В существующих работах на тему генерации графов хорошо себя показали модели, основанные на вариационном автокодировщике (GraphVAE). Генерация, происходит из скрытого вектора, который включает, как интересующую нас информацию (сложные зависимости в графах), так и легко рассчитываемую статистику[1].

Мы предлагаем вычислить на предварительном этапе характеристики, называемые нами простыми признаками: статистики графа, которые можно рассчитать за линейное время от входа, и учесть их отдельно. Это позволит выделить из скрытого вектора составляющие характеризующие сложные зависимости.

В работе было предложено несколько вариантов архитектур на основе GraphVAE и DSSM. GraphVAE используется для создание скрытого представления и генерации, а DSSM для разделения информации внутри скрытого представления. Матрица смежности и признаки вершин используются в качестве входа в GraphVAE. Энкодер GraphVAE преобразовывает исходный граф в скрытый вектор. Декодер трансформирует вектор в новую матрицу смежности. После применения энкодера, мы приводим скрытый вектор и вектор простых признаков в одно пространство при помощи структуры на основе DSSM. Мы строим приближение скрытого вектора на основе вектора простых признаков и вычитаем одно из другого. Полученную разность назовем сложными признаками. Они отвечают за микроструктуру графа.

Исследуется две модификации модели, которые отличаются этапом добавления простых признаков перед пропусканием скрытого вектора через декодер.

[1] *Chamberlain, B., Rowbottom, J., Gorinova, M. I., Bronstein, M., Webb, S., & Rossi, E.* Grand: Graph neural diffusion // International Conference on Machine Learning, PMLR, 2021. — С. 1407–1418.

Controlled Graph Generation

Bishuk Anton[1]⋆ bishuk.ayu@phystech.edu
Zukhba Anastasia[1] azukhba@gmail.com
[1]Moscow, Moscow Institute of Physics and Technology

Recently, the representation of data in the form of graphs is gaining popularity, which allow using not only indicative descriptions of vertex objects, but also the relationship between them. When developing machine learning algorithms on data of such a structure, the problem of dataset diversity is even more acute, so the task of generating synthetic graphs with given properties is important.

In existing works on the topic of graph generation, models based on the variational autoencoder (GraphVAE) have shown themselves well. Generation comes from a hidden vector, which includes both the information of interest to us (complex dependencies in graphs) and easily calculated statistics[1].

We propose to calculate at the preliminary stage the characteristics that we call simple features: graph statistics that can be calculated in linear time from the input, and take them into account separately. This will make it possible to extract from the hidden vector the components that characterize complex dependencies.

Several variants of architectures based on GraphVAE and DSSM were proposed in the work. GraphVAE is used to create a hidden view and generate, and DSSM is used to separate information within a hidden view. The adjacency matrix and vertex features are used as input to GraphVAE. The GraphVAE encoder transforms the original graph into a hidden vector. The decoder transforms the vector into a new adjacency matrix. After applying the encoder, we bring the hidden vector and the vector of simple features into one space using a structure based on DSSM. We build a hidden vector approximation based on a vector of simple features and subtract one from the other. The resulting difference is called complex features. They are responsible for the microstructure of the graph.

At the moment, we are investigating two modifications of the model, which differ in the step of adding simple features before passing the latent vector through the decoder.

[1] *Chamberlain, B., Rowbottom, J., Gorinova, M. I., Bronstein, M., Webb, S., & Rossi, E.* Grand: Graph neural diffusion // International Conference on Machine Learning, PMLR, 2021. — С. 1407–1418.

Модификации градиентных методов с экономичным одномерным поиском

Аникин Антон Сергеевич anikin@icc.ru
Иркутск, ИДСТУ СО РАН

В настоящее время в теории и практике оптимизации активное развитие получили т.н. методы стохастической оптимизации, что обусловлено вызывным ростом интереса к задачам машинного обучения. К одному из наиболее ярких их примеров стоит отнести задачи «обучения» искусственных нейронных сетей, фактически являющиеся задачами минимизации существенно невыпуклых функций. Применение для решения таких задач именно стохастических методов уже практически стало мейнстримом в широкой практике, что обусловлено как «неприятными» свойствами минимизируемых функций (большое число экстремумов), так и чисто техническими проблемами – многие актуальные нейронные сети просто не могут быть обработаны на полном датасете в рамках современных вычислительных устройств.

Тем не менее, «классические» методы минимизации («полноградиентные») также рано сбрасывать со счетов, что подтверждается их наличием в составе современных фреймворков для создания нейронных сетей, таких как, например, PyTorch. Вычислительный опыт автора позволяет утверждать что при наличии вычислительных возможностей (хватает памяти GPU, например) такие методы зачастую показывают гораздо более лучшие результаты, чем популярные варианты стох-методов. Дополнительным плюсом продвинутых вариантов полноградиентных методов является их адаптивность, т.е. способность самостоятельно «настраиваться» на свойства решаемой задачи.

В работе предлагаются модификации ряда полноградиентных методов, направленные на повышение их вычислительной эффективности. Основной идеей подхода является отказ от сложных процедур одномерной минимизации, широко представленных в литературе. Автором продвигается идея «простых» итераций, когда традиционные подходы линейного поиска, надёжные, но громоздкие и вычислительно затратные, заменяются на существенно более простые алгоритмы, в удачных случаях требующие лишь 2-3, а иногда и вовсе 1 вычисление функции за итерацию. Для получения таких результатов в алгоритмы методов внесены модификации, направленные на проведение «правильного» масштабирования направления одномерного поиска.

Результаты вычислительных экспериментов для ряда модельных и прикладных постановок, включая задачи обучения многослойных нейросетей, продемонстрировали эффективность представленных оптимизационных алгоритмов, и позволяют надеться на дальнейшее развитие предложенных подходов, а также адаптацию ряда похожих идей, см. например [1].

[1] *Andrei N.* A double parameter self-scaling memoryless BFGS method for unconstrained optimization // Computational and Applied Mathematics, 39, 159, 2020.

Modifications of gradient methods with economical one-dimensional search

Anikin Anton anikin@icc.ru
Irkutsk, ISDCT SB RAS

Currently, in the theory and practice of optimization, the stochastic optimization methods have been actively developed, which is due to the rapid growth of interest in machine learning problems. One of the most striking examples of them is the artificial neural networks "learning", which are actually problems of essentially non-convex functions minimization. The use of stochastic methods for solving such "learning" problems has practically become mainstream, which is due to both the "unpleasant" properties of minimized functions (a large number of extremes) and purely technical problems – many actual neural networks simply cannot be processed with a full dataset on modern computing devices, such as GPU.

However, "classical" minimization methods (so-called "full-gradient") are also too early to be discounted, which is confirmed by their presence as part of modern neural networks frameworks, such as PyTorch [1]. The author's numerical experience allows us to assert that in the presence of computational capabilities (GPU memory, for example) such methods often show much better results than the popular variants of stochastic methods, like SGD or Adam. An additional benefit of advanced variants of full-gradient methods is their adaptability, i.e. the ability to independently "auto-tune" to the properties of the problem being solved.

The paper proposes modifications of a number of full-gradient methods aimed at improving their numerical efficiency. The main idea of the approach is the rejection of complex one-dimensional minimization procedures that are widely presented in the specialized literature. The author promotes the idea of "simple iterations", when traditional linear search procedures, reliable, but rather complex and computationally expensive, are replaced by significantly simpler algorithms, in successful cases requiring only 2-3, and sometimes even 1 calculation of a function per iteration. To obtain such results, modifications have been made to the proposed optimization algorithms aimed at "proper" scaling of the one-dimensional search direction.

The results of numerical experiments for a number of test and applied problems, including the problems of multilayer neural networks learning, have demonstrated the effectiveness of the presented optimization algorithms, and allow us to hope for further development of the proposed approaches, as well as the adaptation of a number of similar ideas, see for example [2].

[1] `https://pytorch.org/docs/stable/generated/torch.optim.LBFGS.html`

[2] *Andrei N.* A double parameter self-scaling memoryless BFGS method for unconstrained optimization // Computational and Applied Mathematics, 39, 159, 2020.

Технология обучения нейронных сетей на основе метода отжига

Краснопрошин Виктор Владимирович[1,2] krasnoproshin@bsu.by
Мацкевич Вадим Владимирович[2]⋆ matskevich1997@gmail.com
[1]Минск, Белорусский государственный университет
[2]Минск, Белорусский государственный университет

В условиях цифровой трансформации общества возрастает роль информационных технологий, направленных на извлечения полезной информации из массивов цифровых данных. Одной из таких технологий, в частности, является машинное обучение. При этом в качестве алгоритмического инструментария в них часто используются нейронные сети. Последние являются гибкой математической моделью, с помощью которой решается широкий круг прикладных задач. Для настройки нейронной сети на конкретную предметную область производят ее обучение, которая является типичной оптимизационной задачей.

На начальном этапе развития нейронных сетей для их обучения применялся, как правило, градиентный подход. Данный подход получил широкое распространение за счет высокой скорости сходимости реализующих его методов. Однако с развитием вычислительной техники ситуация кардинально изменилась и данный фактор перестал быть определяющим. Это дало возможность развивать альтернативные подходы к обучению.

В работе рассматривается подход, основанный на идеологии случайного поиска. В частности, исследуется одна из простых модификаций метода отжига – Больцмановский отжиг [1]. В отличие от других модификаций он прост в реализации и при решении прикладных задач не требует сложной настройки параметров.

Данный вариант метода отжига определяется двумя основными этапами: – заданием последовательности температур $T_0, T_1, T_2, \ldots$, элементы которой связаны между собой соотношением:

$$T_k = T_0/ln(k+2)$$

С помощью данных значений определяется вероятность перехода из текущего решения x в новое решение y. Вероятность определяется по формуле:

$$P(y|x) = min\{1, (F(x) - F(y))/T_k\}$$

– и построением процедуры генерации случайных решений. Предлагается оригинальный алгоритм, реализующий Больцмановский отжиг.

Пусть, например, нейронная сеть определяется M группой параметров. Тогда алгоритм можно описать следующим образом.

Предварительный этап. Инициализация параметров.

Шаг 0. Задание начальных значений параметров задачи $x = (x_1, x_2, \dots x_n)$ и температуры T_0.

Общая k-ая итерация.

Шаг 1. Генерируются M случайных величин n_i по формуле

$$n_i = \lfloor R[0; m_i] \rfloor$$

где $R[a; b]$ – реализация равномерно распределенной случайной величины на отрезке $[a; b] \subset R$, m_i–количество параметров в каждой группе. Значения n_i, - количество изменяемых параметров.

Шаг 2. Генерируются случайные перестановки Q_i, длинной m_i, $i = \overline{1, M}$. Первые n_i элементов задают индексы изменяемых параметров в каждой из групп параметров.

Пусть, например, $J_i = \{x_{qi1}, x_{qi2}, \dots, x_{qini}\}$, – множества изменяемых параметров.

Шаг 3. Решение $y = (y_1, y_2, \dots y_n)$ генерируется по формуле:

$$y_k = \begin{cases} y_k, y_k \notin J_i, i = \overline{1, M} \\ y_k + a_{ik}, y_k \in J_i, i = \overline{1, M} \\ a_{ik} = R[-l_i; l_i], i = \overline{1, M}, \end{cases}$$

Шаг 4. Вычисляется значение целевой функции $F(y)$.

Шаг 5. Новое решение выбирается в соответствии значением вероятности (2).

Шаг 6. Принцип останова.

Если время, на обучение истекло, то алгоритм завершает работу, в противном случае значения k увеличивается на единицу и происходит переход на Шаг 1.

Нетрудно видеть, что одной из главных проблем данного класса алгоритмов является его сходимость. Было доказано [2], что данный алгоритм гарантированно сходится по вероятности к оптимальному решению из любого начального приближения. Также было установлено теоретическое соотношение значений параметров алгоритма и скорости сходимости метода. На основе результатов теоретических исследований была разработана процедура настройки параметров алгоритма для достижения максимальной скорости сходимости к оптимальному решению. Кроме того было показано, что описанный выше алгоритм может легко быть адаптирован к обучению нейронных сетей другой архитектуры.

По результатам теоретических исследований разработана программная технология (в виде фреймворка) с использованием нейронных сетей, покрывающая все этапы решения прикладных задач (от создания сети до ее применения). Данная технология реализована в виде совокупности отдельных программных

модулей, соответствующих всем этапам решения задачи. В частности, в рамках алгоритмического модуля реализована библиотека, включающая предложенный алгоритм и ряд модификаций известных градиентных алгоритмов (метод простого градиента, метод моментов, метод адаптивного момента и др.). Настройки параметров алгоритмов обучения хранятся в отдельных файлах и могут изменяться.

Необходимо отметить, что при реализации описанного выше алгоритма использовались различные техники параллельных вычислений. В частности, была разработана специальная процедура распараллеливания для эффективного использования кэшей процессора и видеокарты, а также для обеспечения загрузки ядер и минимизации издержек синхронизации. Кроме того, в рамках фреймворка разработана процедура для автоматического выбора эффективного сценария распараллеливания с учетом оценки мощности вычислительных устройств компьютера.

Фреймворк поддерживает возможность прерывания процесса обучения. При запуске обучения указывается время обучения и при его завершении сохраняются все необходимые данные для возобновления процесса обучения (при последующих запусках).

Описанный фреймворк успешно применялся для решения прикладных задач, связанных с анализом и сжатием цветных растровых изображений. Также был проведен ряд экспериментов по проверке эффективности, описанного выше алгоритма обучения. На известных выборках проводилось его сравнение с популярными в литературе алгоритмами градиентного спуска. Показано, что предложенный подход, реализующий идею случайного поиска, при использовании специальных процедур распараллеливания не уступает по скорости обучения градиентным методам. Более того, предложенный алгоритм более чем вдвое превосходит по их качеству.

[1] *Kirckpatrick S.* Optimization by simulated annealing // Science, USA: American Association for the Advancement of Science, 1983. — С. 671–680.

[2] *Krasnoproshin V.* Random search in neural networks training // Proceedings of the 13-th International Conference “Computer Data Analysis and Modeling” – CDAM’2022, Minsk: Belarusian state university, 2022. — С. 96–99.

Technology for training neural networks based on the annealing method

Krasnoproshin Victor[1,2] krasnoproshin@bsu.by
Matskevich Vadim[2]⋆ matskevich1997@gmail.com
[1]Belarus, Minsk, Belarusian State University
[2]Belarus, Minsk, Belarusian State University

In the context of the digital transformation of society, the role of information technologies is increasing, aimed at extracting useful information from digital data arrays. One such technology, in particular, is machine learning. At the same time, neural networks are often used as algorithmic tools. The latter are a flexible mathematical model that can be used to solve a wide range of applied problems. To tune the neural network to a specific subject area, it is trained, which is a typical optimization problem.

At the initial stage of the development of neural networks, as a rule, a gradient approach was used for their training. This approach has become widespread due to the high convergence rate of the methods that implement it. However, with the development of computer technology, the situation has changed dramatically and this factor has ceased to be decisive. This made it possible to develop alternative approaches to training.

The paper considers an approach based on the ideology of random search. In particular, one of the simple modifications of the annealing method, Boltzmann annealing [1], is studied. Unlike other modifications, it is easy to implement and does not require complex parameter settings when solving applied problems.

This version of the annealing method is determined by two main stages: – defining temperature sequence $T_0, T_1, T_2, \ldots$, the elements of which are interconnected by the relation:

$$T_k = T_0/ln(k+2)$$

Using these values, the transition probability from the current solution x into new solution y is determined. The probability is determined by the formula:

$$P(y|x) = min\{1, (F(x) - F(y))/T_k\}$$

– and by constructing generating random solutions procedure. An original algorithm is proposed that implements Boltzmann annealing.

Let, for example, the neural network is defined by M group of parameters. Then the algorithm can be described as follows.

Preliminary stage. Parameters initialization.

Step 0. Setting initial values for task parameters: $x = (x_1, x_2, \ldots x_n)$ and temperature T_0.

General k-th iteration.

Step 1. M random variables n_i are generated by formula

$$n_i = \lfloor R[0; m_i] \rfloor$$

where $R[a; b]$ – is realization of uniform distributed random variable on a segment $[a; b] \subset R$, m_i–amount of parameters in each group. Values n_i, is amount of changing parameters.

Step 2. Random permutations Q_i are generated, with a length of m_i, $i = \overline{1, M}$. First n_i elements determine changing parameters indexes in each group of parameters.

Let, for example, $J_i = \{x_{qi1}, x_{qi2}, \ldots, x_{qini}\}$, are sets of changing parameters.

Step 3. Solution $y = (y_1, y_2, \ldots y_n)$ is generating according to the formula:

$$y_k = \begin{cases} y_k, y_k \notin J_i, i = \overline{1, M} \\ y_k + a_{ik}, y_k \in J_i, i = \overline{1, M} \\ a_{ik} = R[-l_i; l_i], i = \overline{1, M}, \end{cases}$$

Step 4. The objective function $F(y)$ value is calculated.

Step 5. A new solution is chosen according to the probability value (2).

Step 6. Stop principle.

If the time for training has expired, then the algorithm terminates, otherwise the value k is increasing by one and going to Step 1.

It is easy to see that one of the main problems of this class of algorithms is its convergence. It was proved [2] that this algorithm is guaranteed to converge in probability to the optimal solution from any initial approximation. A theoretical relationship between the values of the algorithm parameters and the rate of convergence of the method was also established. Based on the results of theoretical studies, a procedure for adjusting the algorithm parameters was developed to achieve the maximum rate of convergence to the optimal solution. In addition, it was shown that the algorithm described above can be easily adapted to training neural networks of a different architecture.

Based on the results of theoretical studies, a software technology (in the form of a framework) was developed using neural networks, covering all stages of solving applied problems (from creating a network to its application). This technology is implemented as a set of individual software modules corresponding to all stages of solving the problem. In particular, within the framework of the algorithmic module, a library is implemented that includes the proposed algorithm and a number of modifications of known gradient algorithms (simple gradient method, moment method, adaptive moment method, etc.). The settings of training algorithms parameters are stored in separate files and can be changed.

It should be noted that various parallel computing techniques were used in the implementation of the algorithm described above. In particular, a special parallelization procedure was developed to efficiently use the processor and video card caches, as well as to ensure core loading and minimize synchronization overhead. In

addition, within the framework, a procedure has been developed for automatically selecting an efficient parallelization scenario, taking into account the assessment of the power of computer computing devices.

The framework supports the ability to interrupt the training process. When training is started, the training time is indicated and when it is completed, all the necessary data is saved to resume the training process (at subsequent launches).

The described framework has been successfully used to solve applied problems related to the analysis and compression of color raster images. A number of experiments were also carried out to test the effectiveness of the training algorithm described above. On known samples, it was compared with gradient descent algorithms popular in the literature. It is shown that the proposed approach, which implements the idea of random search, when using special parallelization procedures, is not inferior in terms of training speed to gradient methods. Moreover, the proposed algorithm is more than double their quality.

[1] *Kirckpatrick S.* Optimization by simulated annealing // Science, USA: American Association for the Advancement of Science, 1983. — p. 671–680.

[2] *Krasnoproshin V.* Random search in neural networks training // Proceedings of the 13-th International Conference “Computer Data Analysis and Modeling” – CDAM’2022, Minsk: Belarusian state university, 2022. — p. 96–99.

Вычислительные технологии оптимизации атомно-молекулярных кластеров Саттона-Чена размерностей от 101 до 130 атомов

Сороковиков Павел Сергеевич[1]⋆ pavel@sorokovikov.ru
Горнов Александр Юрьевич[1] gornov@icc.ru
[1]Иркутск, Институт динамики систем и теории управления имени В.М. Матросова СО РАН

Возрастающая с каждым днем производительность вычислительных систем дает возможность численного исследования различных содержательных задач глобальной оптимизации больших размерностей. Одной из таких задач является поиск низкопотенциальных состояний атомно-молекулярных кластеров, сводящийся к минимизации многоэкстремальных потенциальных функций. Особенность указанного класса задач состоит в чрезвычайно быстром увеличении числа локальных оптимумов с ростом размерности. Поэтому для исследования задач оптимизации атомно-молекулярных кластеров необходимо применение специализированных подходов, в которых учитывается нетривиальность рассматриваемых структур.

В работе предложены вычислительные технологии, построенные на основе специализированных алгоритмов глобальной и локальной оптимизации. Коллекция алгоритмов нелокального поиска состоит из модификаций алгоритмов MSBH («Monotonic Sequence Basin-Hopping»), «экспертного поиска», Пауэлла, Гергеля, «случайных покрытий», поиска с запретами, Лууса–Яаколы, туннельного поиска, Растригина, Розенброка и ряда биоинспирированных алгоритмов (гармонического поиска, биогеографии, роя частиц, генетического поиска, дифференциальной эволюции, оптимизации по принципу «учитель-ученик», опыления цветков и прочих). Для локальной оптимизации применяются модификации метода многомерного дихотомического поиска, метода Поляка 1969 г., декомпозиционного градиентного метода, алгоритмы L-BFGS, BFGS, сопряжённых градиентов и другие.

С использованием предложенных вычислительных технологий выполнено численное исследование кластеров Саттона-Чена [1, 2] со сверхбольшим числом атомов. Постановка задачи имеет следующий вид:

$$f(x) = \varepsilon \sum_{i=1}^{N} \left[\frac{1}{2} \sum_{j \neq i} \left(\frac{a}{r_{ij}} \right)^p - c \sqrt{\sum_{j \neq i} \left(\frac{a}{r_{ij}} \right)^m} \right] \to \min,$$

$$\varepsilon = 1.0,\ c = 144.41,\ a = 1.0,\ p = 12.0,\ m = 6.0,$$

$$r_{ij} = \sqrt{\sum_{k=1}^{3} \left(x_{3(i-1)+k} - x_{3(j-1)+k} \right)^2}.$$

Здесь N – количество атомов; ε, c, a, p, m – специальные параметры; r_{ij} – расстояние между атомами i и j. К настоящему времени опубликованы результаты расчетов для кластеров Саттона-Чена из 3–80 (в базе данных [1]), 81–100 атомов (в статье авторов [3]). В данной работе проведены системные численные расчеты нахождения низкопотенциальных состояний кластеров из 101–130 частиц с шагом 1. В Таблице 1 приведены наилучшие найденные значения целевой функции при указанных размерностях. Сравнительный анализ результатов численных экспериментов не выявил резких отклонений от наблюдаемой закономерности между найденными значениями потенциальной функции и числом атомов. Авторам неизвестно о других попытках численного решения задач оптимизации атомно-молекулярных кластеров Саттона-Чена для рассматриваемых размерностей.

Таблица 1. Найденные значения потенциальной функции

N	Значение	N	Значение	N	Значение
101	-100559.4244145	111	-110675.4711413	121	-121197.8264439
102	-101494.6104804	112	-111804.7537754	122	-122239.0745511
103	-102339.0141058	113	-112899.2666217	123	-123161.0703135
104	-103361.5083377	114	-113983.8650900	124	-124338.9615661
105	-104635.8467008	115	-115064.4225977	125	-125544.1311698
106	-105528.5077622	116	-115946.4155691	126	-126394.8683179
107	-106534.7214208	117	-117002.2997003	127	-127376.7933742
108	-107584.0958892	118	-118188.8766323	128	-128348.7347942
109	-108639.6763096	119	-119356.7138626	129	-129626.4128232
110	-109681.3715923	120	-120085.4988394	130	-130585.2199066

Результаты получены в рамках госзадания Минобрнауки России по проекту «Теория и методы исследования эволюционных уравнений и управляемых систем с приложениями» (No. гос. регистрации: 121041300060-4).

[1] *Wales D. J., Doye J. P. K.* The Cambridge Energy Landscape Database, URL http://www-wales.ch.cam.ac.uk/CCD.html.

[2] *Todd B. D., Lynden-Bell R. M.* Surface and bulk properties of metals modelled with Sutton-Chen potentials // Surface Science, Vol. 281, No 1-2, 1993. — Pp. 191–206.

[3] *Sorokovikov P., Gornov A.* Modifications of flower pollination, teacher-learner and firefly algorithms for solving multiextremal optimization problems // Algorithms, Vol. 15, No 10, 2022. — P. 359.

Computational technologies for optimizing atomic-molecular Sutton-Chen clusters with dimensions from 101 to 130 atoms

Sorokovikov Pavel[1]⋆ pavel@sorokovikov.ru
Gornov Alexander[1] gornov@icc.ru
[1]Irkutsk, Matrosov Institute for System Dynamics and Control Theory of SB RAS

The increasing productivity of computing systems every day provides the possibility of numerical research of various substantive global optimization problems of large dimensions. One of these problems is the search for low-potential states of atomic-molecular clusters, which reduces to the minimization of multi-extremal potential functions. A feature of this class of problems is the extremely rapid increase in the number of local optima with increasing dimension. Therefore, to investigate optimization problems for atomic-molecular clusters, it is necessary to use specialized approaches that take into account the non-triviality of the structures under consideration.

The paper proposes computational technologies built on the basis of specialized algorithms for global and local optimization. The collection of non-local search algorithms consists of modifications of the MSBH ("Monotonic Sequence Basin-Hopping"), "expert search", Powell, Gergel, "random coverings", taboo search, Luus–Jaakola, tunnel search, Rastrigin, Rosenbrock algorithms and a number of bioinspired algorithms (harmonic search, biogeography, particle swarm, genetic search, differential evolution, teacher-learner optimization, flower pollination, etc.). For local optimization, modifications of the multidimensional dichotomous search method, Polyak's method of 1969, the decomposition gradient method, L-BFGS, BFGS, conjugate gradient algorithms, and others are used.

A numerical study of Sutton-Chen clusters [1, 2] with an extremely large number of atoms has been carried out using the proposed computational technologies. The problem statement has the following form:

$$f(x) = \varepsilon \sum_{i=1}^{N} \left[\frac{1}{2} \sum_{j \neq i} \left(\frac{a}{r_{ij}} \right)^p - c \sqrt{\sum_{j \neq i} \left(\frac{a}{r_{ij}} \right)^m} \right] \to \min,$$

$$\varepsilon = 1.0,\ c = 144.41,\ a = 1.0,\ p = 12.0,\ m = 6.0,$$

$$r_{ij} = \sqrt{\sum_{k=1}^{3} \left(x_{3(i-1)+k} - x_{3(j-1)+k} \right)^2}.$$

Here N is the number of atoms; ε, c, a, p, m are special parameters; r_{ij} is the distance between atoms i and j. To date, the results of computations have been published for Sutton-Chen clusters with 3–80 (in the database [1]), 81–100 atoms (in the paper of the authors [3]). In this work, we performed systemic numerical computations of finding the low-potential states of clusters of 101–130 atoms with

a step of 1. Table 1 shows the best-found values of the objective function for the specified dimensions. A comparative analysis of the results of numerical experiments did not disclose any sharp deflections from the observed regularity between the found values of the potential function and the number of atoms. The authors aren't aware of other efforts to numerically solve the problems of optimizing the Sutton-Chen atomic-molecular clusters for the dimensions under consideration.

Table 1. The found values of the potential function

N	Value	N	Value	N	Value
101	-100559.4244145	111	-110675.4711413	121	-121197.8264439
102	-101494.6104804	112	-111804.7537754	122	-122239.0745511
103	-102339.0141058	113	-112899.2666217	123	-123161.0703135
104	-103361.5083377	114	-113983.8650900	124	-124338.9615661
105	-104635.8467008	115	-115064.4225977	125	-125544.1311698
106	-105528.5077622	116	-115946.4155691	126	-126394.8683179
107	-106534.7214208	117	-117002.2997003	127	-127376.7933742
108	-107584.0958892	118	-118188.8766323	128	-128348.7347942
109	-108639.6763096	119	-119356.7138626	129	-129626.4128232
110	-109681.3715923	120	-120085.4988394	130	-130585.2199066

This research is funded by the grant from the Ministry of Education and Science of Russia within the framework of the project "Theory and methods of research of evolutionary equations and controlled systems with their applications" (state registration number 121041300060-4).

[1] *Wales D. J., Doye J. P. K.* The Cambridge Energy Landscape Database, URL http://www-wales.ch.cam.ac.uk/CCD.html.

[2] *Todd B. D., Lynden-Bell R. M.* Surface and bulk properties of metals modelled with Sutton-Chen potentials // Surface Science, Vol. 281, No 1-2, 1993. — Pp. 191–206.

[3] *Sorokovikov P., Gornov A.* Modifications of flower pollination, teacher-learner and firefly algorithms for solving multiextremal optimization problems // Algorithms, Vol. 15, No 10, 2022. — P. 359.

Использование локально-оптимальных решений при построении кратчайшего незамкнутого пути между объектами

Сурков Егор Эдуардович[1]⋆ eg-su@mail.ru
Середин Олег Сергеевич[1] oseredin@yandex.ru
Копылов Андрей Валериевич[1] and.kopylov@gmail.com
[1]Тула, Тульский государственный университет

В предыдущей работе [1] описана задача поиска кратчайшего незамкнутого пути (КНП) и предложены 9 алгоритмов для ее решения. Самый очевидный способ найти оптимальный КНП – это полный перебор всех всех возможных комбинаций точек, представляющих объекты некоторого набора данных, и выбор пути с минимальной оценкой заданного критерия (алгоритм A0). Такой алгоритм применим только к данным с небольшим количеством объектов и в работе предложены различные модификации жадного алгоритма поиска КНП или квази-КНП (A1-A4R) [1] . Эксперименты показали, что алгоритмы A0-A4R обеспечивают оптимальное решение задачи поиска КНП на небольших данных, а также находят устойчивое решение для данных большого объема [1]. В предыдущей работе отмечается, что с эффективностью модификаций жадных алгоритмов растет и их вычислительная сложность. По этой причине, например, алгоритм A4R не удалось применить на наборе данных Iris Data Set [2] и Abalone Dataset [3]. Более того, алгоритм A0 в работе [1] не удалось применить на данных с количеством объектов больше 15.

Основной идеей новой работы является улучшение решений, а именно уменьшение длины пути, полученных в предыдущей работе за счет поиска локально-оптимальных решений. Так как вычислительная сложность предложенных в работе [1] алгоритмов растет с увеличением количества объектов, предлагается улучшать путь не целиком, а на его отдельных участках (отрезках). Идея заключается в том, чтобы выбрать отрезок пути из N точек, упорядоченный одним из алгоритмов, предложенных ранее [1], зафиксировать две крайние точки и применить алгоритм A0 на объектах между ними. Важно отметить, что алгоритм A0 применяется на всех точках кроме зафиксированных, однако при оценке длины отрезка их необходимо учитывать. Такая процедура гарантировано не увеличит длину, но может ее уменьшить. Предлагается два варианта применения процедуры:

1) A0P. Последовательный проход вдоль пути. Необходимо задать количество объектов участка пути (длину окна) n и шаг смещения s. Затем перемещаться вдоль пути с заданным шагом и применять процедуру для выделенного отрезка. Последовательный проход можно выполнить M раз. Количество обращений к пути за одну итерацию эквивалентно: $\left\lfloor \frac{N-n}{s} + 1 \right\rfloor$.

2) A0PR*. Случайный выбор отрезков. Алгоритм случайным образом выбирает участок пути, к которому будет применена процедура. Параметрами алгоритма являются длина окна n и количество итераций применения алгоритма M. Количество обращений к пути за одну итерацию эквивалентно: $N - n + 1$.

Результатом работы является улучшение решений, найденных алгоритмами, описанными в предыдущей работы на тех же данных. Для алгоритма A0P были установлены параметры $s = 1$, $n = 10$, $M = 10$, а для алгоритма A0PR* параметры $n = 10$, $M = 10$. Для набора Iris Data Set [2] удалось уменьшить длину КНП алгоритмом A0P с 51.627 до 50.130 за время работы алгоритма 295 мс, для набора данных из работы по обнаружению активностей людей на основе базисной совокупности объектов [4] удалось уменьшить длину КНП с 5.95 до 5.902 за 249 мс, а для данных Abalone Dataset [1] с 173.392 до 171.330 за 6 сек. Результаты применения алгоритма A0PR* для данных [2] и [4] такой же, но время потраченное на вычисления выросло в 3-4 раза. Для данных Abalone Dataset [3] длину пути удалось уменьшить до 171.3 за 24 сек.

Алгоритм A0P работает стабильно, однако его решение, за счет последовательного обхода, может остановиться в локальном экстремуме и к глобальному решению прийти не получится. А алгоритм A0PR* за счет случайного выбора отрезков для может обойти локальное решение. Таким образом, рекомендуется запускать алгоритм A0PR* несколько раз, чтобы исключить попадание в локальный экстремум.

Работа выполнена при финансовой поддержке Министерства науки и высшего образования РФ в рамках государственного задания FEWG-2021-0012.

[1] *Seredin O., Surkov E., Kopylov A., Dvoenko S.* Multidimensional Data Visualization Based on the Shortest Unclosed Path Search // Artificial Intelligence in Data and Big Data Processing. ICABDE 2021, Ho Chi Minh City: Springer, 2021. — p. 279–299.

[2] *Fisher R.A.* The Use of Multiple Measurements in Taxonomic Problems // Annals of Eugenics (Vol. 7), 1936. — p. 179–188.

[3] *Warwick J., Tracy L., Simon R., Andrew J., Cawthorn B.-Y.., Kot A.C.,* The Population Biology of Abalone (Haliotis species) in Tasmania. I. Blacklip Abalone (H. rubra) from the North Coast and Islands of Bass Strait // Sea Fisheries Division, No. 48

[4] *Сурков Е.Э.* Исследование базисной совокупности скелетных представлений в задаче детектирования падений // Ломоносов-2021: Сборник тезисов XXVIII Международной научной конференции студентов, аспирантов и молодых ученых, Москва, 12–23 апреля 2021 года, Москва: Издательский отдел факультета ВМК МГУ, ООО "МАКС Пресс 2021. – С. 82-83.

The shortest unclosed path search between objects using locally optimal solutions

Surkov Egor[1]⋆ eg-su@mail.ru
Seredin Oleg[1] oseredin@yandex.ru
Kopylov Andrei[1] and.kopylov@gmail.com
[1]Tula, Tula State University

The previous work [1] defines the shortest unclosed path (SUP) search problem and propose nine algorithms to solve it. The most obvious approach to search the optimal SUP is a brute force of connecting points which represent objects from a certain dataset and choose the shortest path with minimal sum of distances between the points (A0 algorithm). This algorithm could be only applied to small datasets and the work also propose several modifications of greedy algorithm to search the SUP and quasi-SUP (A1-A4R) [1]. Experiments show that A0-A4R algorithms provide the optimal solution on the small dataset and also find a steady solution for a large volume data [1]. The previous work notes that the more effectiveness of greedy algorithm modification, the more computational complexity. For example, A4R algorithm could not be applied to Iris Data Set [2] and Abalone Dataset [3]. Moreover, A0 algorithm couldn't be applied to datasets with volume larger then 15 [1].

The main idea of this work is an improvement of previously obtained solution, namely path length, by the locally optimal solutions. In the previous work computational complexity of algorithms increases due to number of objects in the dataset and we propose to improve the length not on the whole path but on small parts. We propose to choose an N-points segment of the ordered by some previous algorithm path [1] then fix two extreme points and apply the A0 algorithm on the objects between them. That's important to note that A0 algorithm applies on all segment points except of fixed points but the length estimation have to count them all. Such a procedure is guaranteed not to increase the length but it could reduce that. We propose two variants of this procedure applying:

1) A0P. Sequential passage along the path. It is necessary to set a number if segment objects (window length) n, step s and move along a path by step and apply the procedure to selected segment. Sequential passage could be applied M times. The number of path access per iteration is $\left\lfloor \frac{N-n}{s} + 1 \right\rfloor$.

2) A0PR*. The random segment choice. The algorithm randomly choose the part of path for the procedure applying. The algorithm parameters are window length n and iterations number of the algorithm application M. The number of path access per iteration is $N - n + 1$.

Result of algorithms computation is the improvement of solutions obtained by algorithms from previous work on the same data [1]. The parameters $s = 1$, $n = 10$, $M = 10$ were set for A0P algorithm and parameters $n = 10$, $M = 10$ for A0PR*

algorithm. We decrease the path length from 51.627 to 50.130 for 295 ms on the Iris Data Set [2], from 5.95 to 5.902 for 249 ms on the basic assembly data from the human activity recognition work [4] and from 173.392 to 171.330 for 6 sec on the Abalone Dataset [3]. The result of A0PR* application the same on datasets [2] and [4], but the computational time is increases by 3-4 order of magnitude. Also, the path length was decreased to 171.3 for 24 sec on the dataset [3].

A0P algorithm works stably, however, its solution due to sequential traversal could stuck at local extremum and it would not be possible to reach a global one. On the contrary, an A0PR* algorithm could skip a local solution because of randomly chosen segments. Thus, it is recommended to run A0PR* algorithm several times to exclude the getting stuck into a local extremum.

This research is funded by the Ministry of Science and Higher Education of the Russian Federation within the framework of the state task FEWG-2021-0012.

[1] *Seredin O., Surkov E., Kopylov A., Dvoenko S.* Multidimensional Data Visualization Based on the Shortest Unclosed Path Search // Artificial Intelligence in Data and Big Data Processing. ICABDE 2021, Ho Chi Minh City: Springer, 2021. — p. 279–299.

[2] *Fisher R.A.* The Use of Multiple Measurements in Taxonomic Problems // Annals of Eugenics (Vol. 7), 1936. — p. 179–188.

[3] *Warwick J., Tracy L., Simon R., Andrew J., Cawthorn B.-Y.., Kot A.C.,* The Population Biology of Abalone (Haliotis species) in Tasmania. I. Blacklip Abalone (H. rubra) from the North Coast and Islands of Bass Strait // Sea Fisheries Division, No. 48

[4] *Surkov E.* The study about basic assembly of skeletal representations in the fall detection task // Lomonosov-2021: Collection of abstracts XXVIII International scientific conference of students, postgraduates and young scientists, Moscow: 2021, Publishing department of the faculty of computational mathematics and cybernetics, P. 82–83.

Гибридный алгоритм для поиска оптимальных логических правил в данных путем совмещения эвристических и регулярных алгоритмов комбинаторной оптимизации

Масич Игорь Сергеевич i-masich@yandex.ru
Красноярск, Сибирский федеральный университет

Рассматривается задача выявления оптимальных логический правил в данных, которые характеризуют объекты некоторого класса и максимальны по покрытию таких объектов. Такие правила лежат в основе построения "прозрачных"классификаторов. В методологии логического анализа данных такие правила называются паттернами и представляют собой конъюнкцию литералов (значений бинарных признаков и их отрицаний). Наибольший интерес представляют паттерны с высокой обобщающей способностью, которая численно может оцениваться как покрытие объектов обучающей выборки. Эта задача может быть формализована как задача оптимизации: поиск паттерна с наибольшим покрытием объектов некоторого класса при недопустимости покрытия объектов других классов [1]:

$$\sum_{\beta\in K^+}\prod_{\substack{i=1\\ \beta_i\neq\alpha_i}}^{n}(1-y_i)\to\max_{Y},\quad \sum_{\substack{i=1\\ \gamma_i\neq\alpha_i}}^{n} y_i\geqslant 1 \text{ для всех } \gamma\in K^-.$$

где $y_i=\begin{cases}1, & \text{если } i\text{-ый признак фиксирован в } P^\alpha,\\ 0, & \text{иначе.}\end{cases}$

Нахождение паттернов с максимальным покрытием является задачей условной псевдобулевой оптимизации со следующими особенностями:

— нелинейность;
— строгая булевость;
— наличие ограничений;
— потоковость (необходимость решения серии однотипных задач).

Точность получаемого решения задачи оптимизации оказывает влияние на качество обработки данных [2]: получение паттернов с большей обобщающей способностью; качество классификации.

Для решения такой задачи оптимизации могут применяться следующие подходы:

— линеаризация: сведение к задачи целочисленного линейного программирования;
— биоинспирированные алгоритмы (генетический алгоритм);
— точные алгоритмы комбинаторной оптимизации (метод ветвей и границ);
— жадные алгоритмы в различных вариантах (поиск максимальных первичных паттернов, поиск максимальных сильных паттернов).

Гибридный алгоритм оптимизации основывается на жадной эвристике и методе ветвей и границ [3] для логического анализа данных:

— Жадный алгоритм используется для нахождения первого допустимого решения.
— На основе свойств оптимизационной модели исключаются неперспективные подкубы.
— Предлагаемый метод ветвления разбивает оставшиеся обрасти на подкубы, которые являются ветвями для дальнейшего поиска.
— В получаемых ветвях производится отбор и оценка верхней границы.
— Жадный алгоритм используется для нахождения нового допустимого решения в выбранном перспективном подкубе, и т.д.

Использование гибридного алгоритма повышает качество генерируемых паттернов по сравнению с жадным алгоритмом. При этом сложность остается значительно ниже, чем сложность поиска точного решения с помощью целочисленного или смешанного линейного программирования. Предложенная схема гибридизации позволяет контролировать компромисс между качеством генерируемых паттернов и сложностью поиска паттернов в логическом анализе данных.

Работа поддержана Министерством науки и высшего образования Российской Федерации (грант 075-15-2022-1121).

[1] *Bonates T. O., Hammer P. L., Kogan A.* Maximum patterns in datasets // Discrete Appl. Math. 156, 2008. — P. 846–861.

[2] *Lancia G., Serafini P.* Computational Complexity and ILP Models for Pattern Problems in the Logical Analysis of Data // Algorithms 2021, 14, 235. https://doi.org/10.3390/a14080235

[3] *Kazakovtsev L. A., Masich I. S.* A branch-and-bound algorithm for a pseudo-boolean optimization problem with black-box functions // Facta Universitatis, Series Mathematics and Informatics. Vol. 33. No 2. 2018. — P. 337–360.

Hybrid algorithm for finding optimal logical rules in data by combining heuristic and regular combinatorial optimization algorithms

Masich Igor i-masich@yandex.ru
Krasnoyarsk, Siberian Federal University

The problem of identifying optimal logical rules in data that characterize objects of a certain class and are maximal in terms of coverage of such objects is considered. Such rules underlie the construction of "transparent" classifiers. In the methodology of logical analysis of data, such rules are called patterns and are conjunctions of literals (values of binary features and their negations). Of greatest interest are patterns with a high generalizing ability, which can be numerically estimated as a coverage of training objects. This problem can be formalized as an optimization problem: finding a pattern with the largest coverage of objects of a certain class, while covering objects of other classes is unacceptable [1]:

$$\sum_{\beta\in K^+} \prod_{\substack{i=1\\ \beta_i\neq\alpha_i}}^{n} (1-y_i) \to \max_Y, \quad \sum_{\substack{i=1\\ \gamma_i\neq\alpha_i}}^{n} y_i \geqslant 1 \text{ for all } \gamma\in K^-.$$

where $y_i = \begin{cases} 1, \text{ if } i^{\text{th}} \text{ attribute is fixed in } P^\alpha, \\ 0, \text{ otherwise.} \end{cases}$

Finding patterns with maximum coverage is a conditional pseudo-Boolean optimization problem with the following features:

— non-linearity;
— strong Boolean;
— presence of constraints;
— streaming (the need to solve a series of similar problems).

The accuracy of the resulting solution to the optimization problem affects the quality of data processing [2]: obtaining patterns with a greater generalizing ability; classification quality.

To solve such an optimization problem, the following approaches can be used:

— linearization: reduction to integer linear programming problems;
— bioinspired algorithms (genetic algorithm);
— exact combinatorial optimization algorithms (branch-and-bound method);
— greedy algorithms in various variants (search for maximal prime patterns, search for maximal strong patterns).

The hybrid optimization algorithm is based on greedy heuristics and branch-and-bound method [3] for logical analysis of data:

— The greedy algorithm is used to find the first feasible solution.
— Based on the properties of the optimization model, unpromising subcubes are excluded.
— The proposed branching method splits the remaining regions into subcubes, which are branches for further search.
— In the resulting branches, the selection and evaluation of the upper bound is performed.
— A greedy algorithm is used to find a new feasible solution in a given perspective subcube, and so on.

The use of the hybrid algorithm improves the quality of generated patterns compared to greedy heuristics. At the same time, the complexity remains significantly lower than the complexity of finding the exact solution using integer or mixed linear programming. The proposed hybridization scheme allows us to control the trade-off between the quality of generated patterns and the complexity of searching for patterns in logical analysis of data.

This work was supported by the Ministry of Science and Higher Education of the Russian Federation (Grant No.075-15-2022-1121).

[1] *Bonates T. O., Hammer P. L., Kogan A.* Maximum patterns in datasets // Discrete Appl. Math. 156, 2008. — P. 846–861.

[2] *Lancia G., Serafini P.* Computational Complexity and ILP Models for Pattern Problems in the Logical Analysis of Data // Algorithms 2021, 14, 235. https://doi.org/10.3390/a14080235

[3] *Kazakovtsev L. A., Masich I. S.* A branch-and-bound algorithm for a pseudo-boolean optimization problem with black-box functions // Facta Universitatis, Series Mathematics and Informatics. Vol. 33. No 2. 2018. — P. 337–360.

Гендерный генетический алгоритм и его сравнение с обычным генетическим алгоритмом

Куприянов Гавриил[1]⋆ gavriil101@yandex.ru
Исаев Игорь Викторович[2,3] isaev_igor@mail.ru
Доленко Сергей Анатольевич[2] dolenko@srd.sinp.msu.ru
[1]Физический факультет Московского государственного университета имени М.В. Ломоносова, Москва, Россия
[2]Институт ядерной физики Московского государственного университета имени М.В. Ломоносова, Москва, Россия
[2]Институт радиотехники и электроники им. Котельникова Российской академии наук, Москва, Россия

Методология генетических алгоритмов базируется на адаптации Дарвиновского учения о естественном отборе и принципах генетики к решению математических оптимизационных задач. Основная идея генетических алгоритмов (ГА) заключается в том, чтобы путем целенаправленного отбора и объединения хороших решений задачи осуществлять квази-стохастический поиск в пространстве решений. ГА успешно используется для решения широкого набора оптимизационных задач различных областей, таких как управление операциями, обработка данных спектроскопии, автоматическая обработка изображений, обучение нейронных сетей и многих других. Одним из недостатков обычного ГА является преждевременная сходимость к локальному экстремуму. Это связано с быстрым сокращением генетического разнообразия в процессе отбора. Чтобы уменьшить влияние этого фактора, вводится оператор мутации, который случайным образом изменяет часть особи с небольшой вероятностью. Однако слишком высокая вероятность оператора мутации может повлиять на сходимость алгоритма. Чтобы найти компромисс между сохранением популяционного разнообразия и конвергенцией алгоритма, были предложены гендерно-генетические алгоритмы (ГГА). Основной идеей ГГА является разделение особей популяции на две группы, одна из которых направлена на сохранение разнообразия благоприобретенных признаков (особи женского пола), а вторая - на поиск новых решений (особи мужского пола). Желаемый эффект часто достигается путем установки различных значений вероятности мутации для разных полов (вероятность мутации мужского пола намного выше). Другая группа широко распространенных модификаций ГГА связана с изменением оператора отбора. Например, существуют исследования операторов отбора, которые учитывают возраст особи; взаимоотношения между особями; показатели приспособленности родителей особи; близость скрещивающихся особей. В данном исследовании авторы предлагают следующую модификацию GGA: вводится ограничение на то, сколько раз особь женского пола может быть выбрана в течение одного поколения. Целью данного исследования является сравнительный анализ эффективности обычного ГА и предложенной авторами модификации ГГА с ограничением на

скрещивание женской особи, а также изучение влияния данного ограничительного параметра качество решения. Результаты этой работы можно сформулировать следующим образом:

1. ГГА продемонстрировал свое превосходство над стандартным ГА при решении многоэкстремальных задач с ярковыраженным глобальным экстремумом.
2. С простыми задачами одинаково хорошо работают как стандартные ГА, так и ГГА.
3. ГГА сходится немного медленнее, чем ГА, из-за его более высокого разнообразия.

[1] *Kupriyanov, G., Isaev, I., Dolenko, S.*. A Gender Genetic Algorithm and Its Comparison with Conventional Genetic Algorithm(2023). // Advances in Neural Computation, Machine Learning, and Cognitive Research VI. NEUROINFORMATICS 2022. Studies in Computational Intelligence, vol 1064

A Gender Genetic Algorithm and its Comparison with Conventional Genetic Algorithm

Kupriyanov Gavriil[1]⋆ gavriil101@yandex.ru
Isaev Igor[2,3] isaev_igor@mail.ru
Dolenko Sergey[2] author_email@site.ru
[1]Faculty of Physics, M.V. Lomonosov Moscow State University, Moscow, Russia
[2]Institute of Nuclear Physics, M.V. Lomonosov Moscow State University, Moscow, Russia
[3]Kotelnikov Institute of Radio Engineering and Electronics, Russian Academy of Sciences, Moscow, Russia

The methodology of genetic algorithms (GA) is based on the adaptation of Darwin's doctrine of natural selection and the principles of genetics to the solution of mathematical optimization problems. The main idea of GA is to carry out a quasi-stochastic search in the solution space by purposefully selecting and combining good solutions to a problem. GA is successfully used to solve a wide range of optimization problems from various fields, such as operations management, spectroscopy data processing, automatic image processing, training neural networks and many others. One of the disadvantages of conventional GA is the premature convergence of the process to a local extremum. This is due to the rapid reduction in genetic diversity through the selection process. To reduce the influence of this factor, the mutation operator is introduced, which randomly changes a part of an individual with a small probability. However, a too high probability of the mutation operator can affect the convergence of the algorithm. In order to find a compromise between the conservation of population diversity and the convergence of the algorithm, biologically inspired gender genetic algorithms (GGA) have been proposed. The main idea of the GGA is the division of individuals of the population into two groups, one of which is aimed at preserving the diversity of acquired traits (female individuals), and the second at search for new solutions (male individuals). The desired effect is often achieved by setting different values of the probability of mutation for different genders (much higher for the male gender). Another group of widespread GGA modifications is associated with modifying the selection operator. For example, there are studies of selection operators that consider the age of an individual; the relationship between individuals; the fitness values of the parents of the individual; the proximity of the crossing individuals. In this study, the authors propose such a modification of the GGA as the restriction on how many times a female individual may be selected within one generation. The purpose of this study is the comparative analysis of the efficiency of the conventional GA and the modification of the GGA proposed by the authors with a restriction on crossover of female individual, as well as the study of influence of the crossover restriction parameter value on the solution quality. Results of this work could be formulated as:

1. GGA showed its superiority over the standard GA in solving problems with a set of local extremums and a global extremum with much different fitness function value.
2. With simple tasks, both standard GA and GGA work equally well.
3. GGA converges slightly slower than GA due to its higher diversity.

[1] *Kupriyanov, G., Isaev, I., Dolenko, S..* A Gender Genetic Algorithm and Its Comparison with Conventional Genetic Algorithm(2023). // Advances in Neural Computation, Machine Learning, and Cognitive Research VI. NEUROINFORMATICS 2022. Studies in Computational Intelligence, vol 1064,

О задачах оптимизации, возникающих при применении топологического анализа данных к поиску алгоритмов прогнозирования с фиксированными с корректорами в рамках алгебраического подхода Ю.И.Журавлёва

Торшин Иван Юрьевич[1]⋆ tiy135@yahoo.com
[1]Москва, ФИЦ ИУ РАН

Корректирующие операции (корректоры) в мультиалгоритмических конструкциях алгебраического подхода могут строиться на основе известных физических моделей и/или многоуровневых описаний физических объектов. При этом, в рамках топологического подхода к анализу плохо-формализованных задач, поиск включаемых в корректор алгоритмов может рассматриваться как задача комбинаторной оптимизации либо как задача минимизации некой функции потерь. Исследование окрестностей цепей в решётке подмножеств объектов позволило получить ряд критериев ранговой оптимизации, перспективных для решения задач прогнозирования числовых таргетных переменных. Формализм апробирован на задаче взаимодействия лиганд-рецептор в рамках хемокиномного анализа молекул лекарств (данные ProteomicsDB). Наилучшие результаты прогнозирования констант наблюдались именно при использовании полученных ранговых критериев (коэффициент корреляции на скользящем контроле 0.86 ± 0.20, усреднение по 300 биологическим активностям).

Работа поддержана грантом РФФИ No. 20-07-0053.

[1] *Торшин И. Ю.* О применении топологического подхода к анализу плохо формализуемых задач для построения алгоритмов виртуального скрининга квантово-механических свойств органических молекул. Часть 2. Сопоставление формализма с конструктами квантовой механики и экспериментальная апробация предложенных алгоритмов // Информатика и её применения, Москва, 2022. — 16(2), С. 35–43.

On optimization problems arising from the application of topological data analysis to the search for forecasting algorithms with fixed correctors in the framework of Yu.I. Zhuravlev's algebraic approach

Torshin Ivan[1]⋆ tiy135@yahoo.com

[1]Moscow, FIC IU RAS

Corrective operations (correctors) in multialgorithmic constructions of the algebraic approach can be based on known physical models and/or multilevel descriptions of physical objects. At the same time, within the framework of the topological approach to the analysis of poorly formalized problems, the search for algorithms included in the corrector can be considered as a combinatorial optimization problem or as a problem of minimizing a certain loss function. The study of the neighborhoods of chains in the lattice of subsets of objects made it possible to obtain a number of rank optimization criteria that are promising for solving problems of predicting numerical target variables. The formalism was tested on the problem of ligand-receptor interaction in the framework of the chemokinomic analysis of drug molecules (data from ProteomicsDB). The best results of predicting constants were observed precisely when using the obtained rank criteria (correlation coefficient on a sliding control 0.86 ± 0.20, averaging over 300 biological activities).

This research is funded by RFBR, grant 20-07-0053.

[1] *Torshin I.* On the application of the topological approach to the analysis of poorly formalized problems for the construction of virtual screening algorithms for the quantum mechanical properties of organic molecules. Part 2. Comparison of the formalism with the constructs of quantum mechanics and experimental testing of the proposed algorithms // Informatics and its applications, Moscow, 2022. — 16(2), p. 35–43.

Асимптотически точный подход к решению задачи поиска нескольких реберно-непересекающихся остовных деревьев минимального суммарного веса с фиксированным диаметром в полном неориентированном графе со случайными весами ребер

Гимади Эдуард Хайрутдинович[1,2] gimadi@math.nsc.ru
Штепа Александр Александрович[2]⋆ shoomath@gmail.com
[1]Новосибирск, Институт Математики им. С. Л. Соболева СО РАН
[2]Новосибирск, Новосибирский Национальный Исследовательский Государственный Университет (НГУ)

Мы рассматриваем труднорешаемую задачу поиска нескольких реберно-непересекающихся остовных деревьев минимального суммарного веса с фиксированным диаметром в полном неориентированном графе со случайными весами ребер из нескольких классов непрерывных распределений: равномерное, смещенное экспоненциальное, усеченно-нормальное. В данной статье мы предлагаем приближенный алгоритм решения выше указанной задачи с трудоемкостью $O(n^2)$, где n — это количество вершин в графе, а также приводим условия асимптотической точности для этого алгоритма в случае каждого из рассматриваемых распределений.

Диаметр дерева — это количество ребер в самом длинном простом пути в дереве, соединяющим пару вершин. В качестве обобщения классической задачи поиска минимального остовного дерева можно рассмотреть следующую задачу: дан реберно-взвешенный граф и число d, необходимо найти в этом графе минимальный остов, имеющий диаметр, ограниченный сверху или снизу числом d. Обе задачи в общей постановке являются NP-трудными.

Вариант задачи с ограничением на диаметр сверху является полиномиально разрешимым для значений диаметра два и три, и NP-трудным для любого диаметра между 4 и $(n-1)$, даже для весов ребер, равных 1 или 2 [1, стр. 206]. Задача с ограничением снизу является NP-трудной, потому что она содержит в качестве подзадачи, для $d = n-1$ задачу "гамильтонов путь" [1]. Рассматриваемая нами задача может быть получена из выше описанных таким образом, что диаметр искомого дерева ограничен числом d и сверху, и снизу. Легко понять, что эта задача также NP-трудна.

Основной результат работы — теоретический. С помощью теоремы Петрова [4] и нескольких вспомогательных лемм для каждого из рассматриваемых распределений вероятности доказывается теорема с достаточными условиями асимптотической точности предлагаемого приближенного алгоритма. Вариант задачи с одним остовным деревом встречается в проектировании сетей связи, текстовых информационно-поисковых системах, специальных беспроводных системах [3]. Стоит, однако, отметить, что эти применения для ограниченной задачи поиска минимального остовного дерева, где диаметр дерева ограничен свер-

ху числом d. Но рассматриваемая в текущей работе задача — это специальный случай этой задачи, когда диаметр равен d, поэтому для нее также верны эти приложения.

В работе [2] представлен вероятностный анализ полиномиального алгоритма и предложены условия его асимптотической точности для решения задачи поиска минимального остовного дерева с фиксированным диаметром в случае полного ориентированного графа. К сожалению, анализ этого алгоритма оказывается неприемлемым в случае неориентированного графа, поскольку вероятностные свойства веса одного и того же ребра, рассматриваемого по ходу работы алгоритма, нельзя считать независимыми, в отличие от случая ориентированного графа.

Работа выполнена в рамках государственного задания ИМ СО РАН (проект FWNF-2022-0019).

[1] *Garey M. R., Johnson D. S.* Computers and Intractability, Freeman, San Francisco, 1979, 340 p.

[2] *Gimadi E. Kh., Shevyakov A. S., Shtepa A. A.* A Given Diameter MST on a Random Graph // In: N. Olenev, Y. Evtushenko, M. Khachay, V. Malkova (eds.), Optimization and Applications - 11th International Conference OPTIMA 2020, LNCS **12422**, 2020, P. 110–121. doi: 10.1007/978-3-030-62867-3_9

[3] *Gruber M.* Exact and Heuristic Approaches for Solving the Bounded Diameter Minimum Spanning Tree Problem // Vienna University of Technology, PhD Thesis, 2009.

[4] *Petrov V. V.* Limit Theorems of Probability Theory. Sequences of Independent Random Variables, Clarendon Press, Oxford, 1995, 304 p.

On asymptotically optimal approach for the problem of finding several edge-disjoint spanning trees of minimal total weight with given diameter in a complete undirected graph with random edge weights

Gimadi Edward[1,2] gimadi@math.nsc.ru
Shtepa Alexandr[2]⋆ shoomath@gmail.com
[1]Novosibirsk, Sobolev Institute of Mathematics SB RAS
[2]Novosibirsk, Novosibirsk State University (NSU)

We consider the intractable problem of finding several edge-disjoint spanning trees of minimal total weight with a given diameter in a complete undirected graph with random edge weights from the several classes of continuous probability distributions: uniform distribution, biased exponential distribution, and truncated normal distribution. In this work we give an $O(n^2)$-time approximation algorithm, where n is number of vertices in the graph, with conditions of its asymptotic optimality for the case of each considered probability distribution.

The diameter of a tree is the number of edges in the longest simple path within the tree connecting a pair of vertices. As a generalization of the classic Minimum Spanning Tree (MST) problem, one may consider the following problem: given an edge-weighted undirected graph and a number d, the goal is to find a spanning tree T of minimal total weight having its diameter bounded above to given number d, or from below to given number d in the graph. Both problems are NP-hard in general.

The bounded from above MST problem is polynomially solvable for diameters two or three, and NP-hard for any diameter between 4 and $(n-1)$, even for the edge weights equal to 1 or 2 [1, p. 206]. The bounded from below MST problem is NP-hard, because its particular case for $d = n-1$ is the problem "Hamiltonian Path" [1]. The considered problem can be obtained from described above problems in such a way that the diameter of the desired tree is limited by the number d both from above and below. It is easy to see that this problem is also NP-hard.

The main result of this paper is theoretical. By help of Petrov's theorem [4] and several auxiliary lemmas, three theorems with sufficient asymptotic conditions of proposed approximation algorithm are proved for three considered probability distributions. The variant of the problem with one spanning tree arises in network design problem, information retrieval systems, ad-hoc wireless networks [3]. It must be noted that all theese applications are for the bounded from above MST problem. But considered problem in current work is a special case of this problem, we consider problem with diameter equals d in current work, so all the applications are valid for the problem.

The work [2] gives a probabilistic analysis of an effective algorithm for finding one and several spanning trees of minimal total weight in the case of complete directed graph. Unfortunately, the algorithm analysis, presented in this work becomes unacceptable for a problem on undirected graphs. The appearance of the difficulty

of probabilistic analysis in the case of the undirected graph arises from the need to take into account the possible dependence between different random variables in the course of the algorithm.

This research was carried out within the framework of the state contract of the Sobolev Institute of Mathematics (project FWNF-2022-0019).

[1] *Garey M. R., Johnson D. S.* Computers and Intractability, Freeman, San Francisco, 1979, 340 p.

[2] *Gimadi E. Kh., Shevyakov A. S., Shtepa A. A.* A Given Diameter MST on a Random Graph // In: N. Olenev, Y. Evtushenko, M. Khachay, V. Malkova (eds.), Optimization and Applications - 11th International Conference OPTIMA 2020, LNCS **12422**, 2020, P. 110–121. doi: 10.1007/978-3-030-62867-3_9

[3] *Gruber M.* Exact and Heuristic Approaches for Solving the Bounded Diameter Minimum Spanning Tree Problem // Vienna University of Technology, PhD Thesis, 2009.

[4] *Petrov V. V.* Limit Theorems of Probability Theory. Sequences of Independent Random Variables, Clarendon Press, Oxford, 1995, 304 p.

Минимальное Информационное Пространство как Основа Эффективной Распределенной Обработки Больших Данных

Голубцов Петр Викторович[1] golubtsov@physics.msu.ru

[1]Москва, Московский государственный университет имени М. В. Ломоносова

Данные в современных исследованиях нередко имеют огромный объем, распределены между многочисленными сайтами и постоянно пополняются. В результате собрать все относящиеся к исследованию данные на одном компьютере, как правило, невозможно и непрактично, поскольку один компьютер не сможет обработать их в разумные сроки. Подходящий алгоритм анализа данных должен, параллельно работая на многих компьютерах, извлекать из каждого набора исходных данных некоторую промежуточную компактную «информацию», постепенно объединять ее и, наконец, использовать накопленную информацию для получения результата. По мере поступления новых данных он должен иметь возможность добавлять содержащуюся в них информацию к накопленной и, в конечном итоге, обновлять результат. В работах [1, 2] было показано, что для эффективной обработки распределенных данных ключевую роль играет возможность введения промежуточной формы представления информации, обладающей определенными алгебраическими свойствами.

Данная работа посвящена алгебраической формализации распределенной обработки данных при весьма общих условиях. Определяется понятие информационного пространства и, в частности, минимального информационного пространства, предоставляющего максимально компактную форму накопления информации и позволяющего оптимально распараллелить обработку данных.

Пусть $\mathcal{D}$ — множество возможных значений входных данных, а $\mathcal{R}$ — множество значений результатов обработки. В задачах больших данных на вход процедуры обработки поступают наборы элементов из $\mathcal{D}$, причем эти наборы могут быть разбросаны по многим компьютерам. Для математического представления множества всех таких наборов с операцией их слияния обычно используется свободный моноид $\mathcal{D}^*$ с операцией конкатенации. Однако, поскольку результат обработки обычно не должен зависеть от порядка поступления данных, удобно представлять пространство всевозможных наборов исходных данных *свободным коммутативным моноидом* $\mathcal{D}^+$ с множеством образующих $\mathcal{D}$. Его элементами являются конечные мультимножества на множестве $\mathcal{D}$ (в которых элемент может повторяться несколько раз) с операцией *объединения* мультимножеств (при которой кратности элементов складываются). *Процедура обработки* с наборами данных из множества данных $\mathcal{D}$ и результатами из множества $\mathcal{R}$ определяется как отображение p из свободного коммутативного моноида $\mathcal{D}^+$ в множество $\mathcal{R}$, т.е. $p : \mathcal{D}^+ \to \mathcal{R}$.

Определение 1. *Информационное пространство (ИП)* $(\mathcal{U}, q, r)$ *для процедуры* $p : \mathcal{D}^+ \to \mathcal{R}$ *– это коммутативный моноид* $\mathcal{U}$*, сюръективный гомоморфизм (СГ)* $q : \mathcal{D}^+ \to \mathcal{U}$ *и отображение* $r : \mathcal{U} \to \mathcal{R}$*, такие что* $r \circ q = p$*.*

Фактически, гомоморфизм q «сжимает» исходные данные без потери информации. Его гомоморфность означает, что объединению наборов данных отвечает сумма соответствующих фрагментов информации, а его сюръективность обеспечивает отсутствие в $\mathcal{U}$ «лишних» элементов, которые никогда не могут возникнуть.

Определение 2. *Будем говорить, что ИП* $(\mathcal{U}, q, r)$ *меньше (лучше), чем* $(\mathcal{U}', q', r')$ *и обозначать это как* $(\mathcal{U}, q, r) \ll (\mathcal{U}', q', r')$, *если существует отображение* $h : \mathcal{U}' \to \mathcal{U}$ *такое, что* $h \circ q' = q$.

Поскольку q – СГ, такое *преобразование* информационных пространств h единственно и также является СГ. При этом $r \circ h = r'$, т.е. $(\mathcal{U}, h, r)$ можно рассматривать как ИП для процедуры $r' : \mathcal{U}' \to \mathcal{R}$. Отношение $\ll$ является предпорядком, причем если $\mathcal{U} \ll \mathcal{U}'$ и $\mathcal{U}' \ll \mathcal{U}$, то эти ИП изоморфны. *Минимальное* в смысле этого упорядочения ИП $(\mathcal{U}, q, r)$ обладает тем свойством, что любое ИП $(\mathcal{U}', q', r')$ для p факторизуется через него, т.е. существует (единственный) СГ $h : \mathcal{U}' \to \mathcal{U}$ такой что $h \circ q' = q$ и $r' = r \circ h$.

Для доказательства существования минимального ИП дадим следующее

Определение 3. *Будем говорить, что элементы* x *и* y *из* $\mathcal{U}$ *неразличимы относительно* $r : \mathcal{U} \to \mathcal{R}$ *и обозначать* $x \sim_r y$, *если* $\forall z \in \mathcal{U}\ r(x + z) = r(y + z)$.

Теорема 1 (Существование). *Минимальное ИП для процедуры* $p : \mathcal{D}^+ \to \mathcal{R}$ *существует и с точностью до изоморфизма совпадает с фактормоноидом* $(\mathcal{D}^+/\!\sim_p, q, r)$ *по конгруэнции неразличимости на* $\mathcal{D}^+$ *относительно* p. *При этом* $q : \mathcal{D}^+ \to \mathcal{D}^+/\!\sim_p$ *— соответствующий канонический эпиморфизм,* $q(x) = [x]_{\sim_p}$, $x \in \mathcal{D}^+$, *а отображение* $r : \mathcal{D}^+/\!\sim_p \to \mathcal{R}$ *определяется как* $r([x]_{\sim_p}) = p(x)$ *для* $x \in \mathcal{D}^+$.

Следующее утверждение дает удобный критерий минимальности.

Теорема 2 (Критерий минимальности). *ИП* $(\mathcal{U}, q, r)$ *является минимальным если все его элементы различимы относительно конгруэнции* $\sim_r$.

Алгебраическая структура информационного пространства позволяет естественным образом определить понятие *качества информации*. А именно для элементов x и y ИП $\mathcal{U}$ будем говорить, что x *лучше* (представляет больше информации), чем y и обозначать $x \succcurlyeq y$ если $\exists z \in \mathcal{U}\ x = y + z$. Отношение качества на ИП $\mathcal{U}$ является отношением предпорядка, согласованным с алгебраической структурой, т.е., $x' \succcurlyeq x\ \&\ y' \succcurlyeq y \implies x' + y' \succcurlyeq x + y$ и $x \succcurlyeq 0$. Более того, если $h : \mathcal{U}' \to \mathcal{U}$ – преобразование ИП, то h сохраняет упорядочение качества: $x \succcurlyeq y \implies h(x) \succcurlyeq h(y)$.

Использование минимального ИП позволяет максимально эффективно распараллеливать процесс накопления информации в рамках модели распределенного анализа данных MapReduce [3] и организовать эффективную обработку

без необходимости передачи и накопления самих исходных данных. В контексте этой модели Map преобразует наборы исходных данных в элементы ИП, а Reduce объединяет все эти фрагменты частичной информации в один элемент, представляющий все исходные данные, Рис. 1.

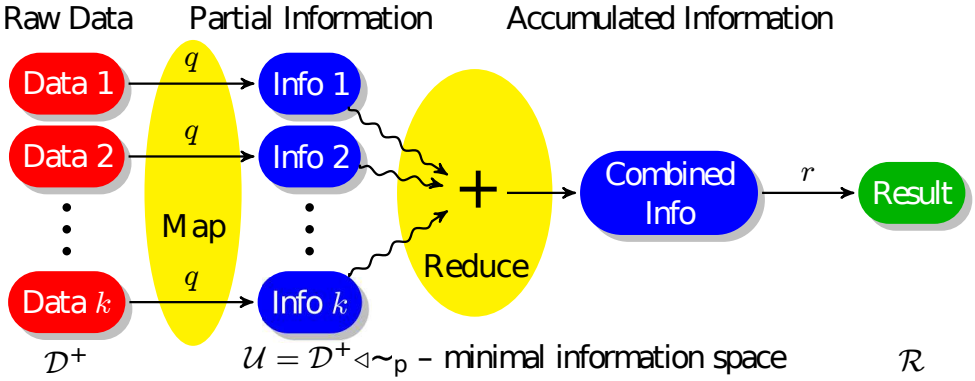


Рис. 1. Параллельная обработка распределенных данных с использованием минимального информационного пространства в модели MapReduce.

При этом минимальное информационное пространство определяет «оптимальную» математическую структуру для представления информации, содержащейся в данных, для модели MapReduce и описывает «теоретический предел» компактности представления информации.

Как показано в этой работе, при весьма общих предположениях, проблема оптимизации распределенной обработки данных приводит к математическому представлению информации, содержащейся в данных, как элементу минимального информационного пространства. При этом в терминах ИП естественным образом выражаются сложение и качество информации.

Понятие информации всегда было предметом преимущественно теоретического интереса. Сейчас проблематика больших данных требуют компактных, эффективных и хорошо организованных форм представления информации. Такие идеальные формы могут отражать самую суть информации, содержащейся в данных. Поэтому изучение таких форм и их свойств может приблизить нас к адекватному математическому описанию самого понятия информации.

Работа выполнена при финансовой поддержке РФФИ, грант № 19-29-09044.

[1] *Голубцов П. В.* Понятие информации в контексте задач обработки больших данных // НТИ Сер. 2. Информационные процессы и системы, 2017. №1. С. 31–36.

[2] *Golubtsov P.* Scalability and Parallelization of Sequential Processing: Big Data Demands and Information Algebras // Advances in Intelligent Systems and Computing, Springer, Cham. 2020. Vol. 1127. Pp. 274–298.

[3] *Dean J. and Ghemawat S.* MapReduce: simplified data processing on large clusters // Communications of the ACM, 2008. Vol. 51. No. 1. Pp. 107–113.

Minimal Information Space as a Background for Efficient Distributed Big Data Processing

Golubtsov Peter[1] golubtsov@physics.msu.ru

[1]Moscow, Lomonosov Moscow State University

Data in modern research are often huge, distributed among numerous sites and constantly updated. As a result, it is generally impossible and impractical to collect all research-relevant data on one computer, as one computer will not be able to process them in a reasonable amount of time. A suitable data analysis algorithm should, working in parallel on many computers, extract some compact intermediate "information" from each set of initial data, gradually merge it and, finally, use the accumulated information to obtain a result. As new data arrives, it must be able to add the information contained in new data to the accumulated information and, ultimately, update the result. In the works [1, 2] it was shown that for the efficient processing of distributed data, the possibility of introducing an appropriate intermediate form of information representation with certain algebraic properties plays a key role.

This work is devoted to the algebraic formalization of distributed data processing under very general conditions. The concept of information space is defined and, in particular, the minimum information space is introduced. It provides the most compact form of information accumulation and allows optimal parallelization of data processing.

Let $\mathcal{D}$ be the set of possible input data values, and $\mathcal{R}$ be the set of processing result values. In big data problems, collections of elements from $\mathcal{D}$ are fed to the input of the processing procedure, and these collections can be scattered over many computers. For the mathematical representation of the set of all such collections with the operation of their merge, the free monoid $\mathcal{D}^*$ with the operation of concatenation is usually used. However, since the result of processing usually should not depend on the order in which the data arrives, it is convenient to represent the space of all possible collections of initial data by the *free commutative monoid* $\mathcal{D}^+$ with the set of generators $\mathcal{D}$. Its elements are finite multisets on the set $\mathcal{D}$ (in which an element can be repeated several times) with the multiset *union* operation (where the multiplicities of the elements are added). *Processing procedure* for data sets with elements from the set $\mathcal{D}$ and results from the set $\mathcal{R}$ is defined as a mapping p from the free commutative monoid $\mathcal{D}^+$ to the set $\mathcal{R}$, i.e., $p : \mathcal{D}^+ \to \mathcal{R}$.

Definition 1. *An information space (IS) $(\mathcal{U}, q, r)$ for the procedure $p : \mathcal{D}^+ \to \mathcal{R}$ is a commutative monoid $\mathcal{U}$, a surjective homomorphism (SH) $q : \mathcal{D}^+ \to \mathcal{U}$ and a mapping $r : \mathcal{U} \to \mathcal{R}$ such that $r \circ q = p$.*

In fact, the homomorphism q "compresses" the original data without loss of information. Homomorphic means that the sum of the corresponding pieces of in-

formation corresponds to the union of data sets, and its surjectivity ensures that $\mathcal{U}$ does not contain "extra" elements that can never appear.

Definition 2. *We will say that the IS* $(\mathcal{U}, q, r)$ *is smaller (better) than* $(\mathcal{U}', q', r')$ *and denote it as* $(\mathcal{U}, q, r) \ll (\mathcal{U}', q', r')$ *if there exists a mapping* $h : \mathcal{U}' \to \mathcal{U}$ *such that* $h \circ q' = q$.

Since q is a SH, such *transformation* h of information spaces is unique and is also a SH. Besides, $r \circ h = r'$, i.e. $(\mathcal{U}, h, r)$ can be thought of as an IS for the procedure $r' : \mathcal{U}' \to \mathcal{R}$. The relation $\ll$ is a preorder, and if $\mathcal{U} \ll \mathcal{U}'$ and $\mathcal{U}' \ll \mathcal{U}$, then these ISs are isomorphic. A *minimal* in the sense of this ordering IS $(\mathcal{U}, q, r)$ has the property that any IS $(\mathcal{U}', q', r')$ for p factorizes through it, i.e., there exists a (unique) SH $h : \mathcal{U}' \to \mathcal{U}$ such that $h \circ q' = q$ and $r' = r \circ h$.

To prove the existence of a minimal IS, we give the following

Definition 3. *We say that elements* x *and* y *from* $\mathcal{U}$ *are indistinguishable with respect to* $r : \mathcal{U} \to \mathcal{R}$ *and denote it as* $x \sim_r y$ *if* $\forall z \in \mathcal{U}\ r(x + z) = r(y + z)$.

Theorem 1 (Existence). *The minimal IS for the procedure* $p : \mathcal{D}^+ \to \mathcal{R}$ *exists and, up to isomorphism, coincides with the factormonoid* $(\mathcal{D}^+/\!\sim_p, q, r)$ *by the indistinguishability congruence on* $\mathcal{D}^+$ *with respect to* p. *Moreover,* $q : \mathcal{D}^+ \to \mathcal{D}^+/\!\sim_p$ *is the corresponding canonical epimorphism,* $q(x) = [x]_{\sim_p}$, $x \in \mathcal{D}^+$, *and the mapping* $r : \mathcal{D}^+/\!\sim_p \to \mathcal{R}$ *is defined as* $r([x]_{\sim_p}) = p(x)$ *for* $x \in \mathcal{D}^+$.

The following assertion gives a convenient minimality criterion.

Theorem 2 (Minimality criterion). *An IS* $(\mathcal{U}, q, r)$ *is minimal if all its elements are distinguishable with respect to the congruence* $\sim_r$.

The algebraic structure of the information space allows to naturally introduce the concept of *quality of information*. Namely, for elements x and y from the IS $\mathcal{U}$, we say that x is *better* (represents more information) than y and denote $x \succcurlyeq y$ if $\exists z \in \mathcal{U}\ x = y + z$. The quality relation on $\mathcal{U}$ is a preorder relation consistent with the algebraic structure, i.e., $x' \succcurlyeq x\ \&\ y' \succcurlyeq y \implies x' + y' \succcurlyeq x + y$ and $x \succcurlyeq 0$. Moreover, if $h : \mathcal{U}' \to \mathcal{U}$ is a transformation of ISs, then h preserves the quality ordering: $x \succcurlyeq y \implies h(x) \succcurlyeq h(y)$.

The use of the minimum IS allows to most efficiently parallelize the process of information accumulation within the MapReduce [3] distributed data analysis model and organize efficient processing without the need to transfer and accumulate the original data. In the context of this model, Map transforms the source data sets into elements of the IS, and Reduce combines all these pieces of partial information into one element representing all the source data, Fig. 1.

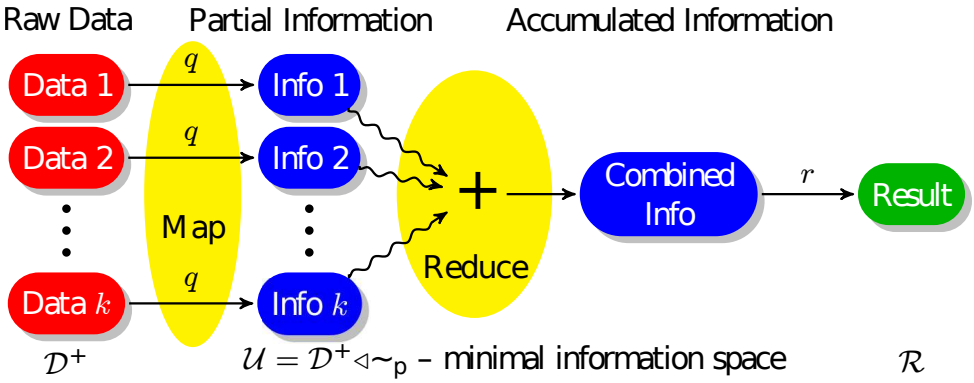


Fig. 1. Parallel processing of distributed data using minimal information space within the MapReduce model.

At the same time, the minimum information space determines the "optimal" mathematical structure for representing the information contained in the data for the MapReduce model and describes the "theoretical limit" of compactness of information representation.

As shown in this work, under very general assumptions, the problem of optimizing distributed data processing leads to a mathematical representation of the information contained in the data as an element of the minimum information space. And within such framework of ISs, the addition and quality of information are naturally expressed.

The concept of information has always been a subject of predominantly theoretical interest. Now the problems of big data require compact, efficient and well-organized forms of information representation. Such ideal forms can reflect the very essence of the information contained in the data. Therefore, the study of such forms and their properties can bring us closer to an adequate mathematical description of the very concept of information.

This research is funded by RFBR, grant 19-29-09044.

[1] *Golubtsov P. V.* The concept of information in big data processing // Autom. Doc. Math. Linguist., 2018. Vol. 52. No. 1. Pp. 38–43.

[2] *Golubtsov P.* Scalability and Parallelization of Sequential Processing: Big Data Demands and Information Algebras // Advances in Intelligent Systems and Computing, Springer, Cham. 2020. Vol. 1127. Pp. 274–298.

[3] *Dean J. and Ghemawat S.* MapReduce: simplified data processing on large clusters // Communications of the ACM, 2008. Vol. 51. No. 1. Pp. 107–113.

Вычислительная сложность двух задач когнитивного анализа данных

Кутненко Ольга Андреевна[1,2] olga@math.nsc.ru
[1]Новосибирск, Институт математики им. С. Л. Соболева
[2]Новосибирск, Новосибирский гос. университет

Доказана NP-трудность в сильном смысле двух задач когнитивного анализа данных: задачи таксономии (кластеризации) и задачи выбора подмножества типичных представителей классифицированной выборки, состоящей из объектов двух образов. Для количественной оценки качества множества выбранных типичных представителей выборки (прототипов) используется функция конкурентного сходства — FRiS-функция (Function of Rival Similarity) [1], с помощью которой оценивается сходство объекта с ближайшим типичным представителем выборки.

Задача таксономии (кластеризации) — это задача разбиения неклассифицированной выборки объектов на непересекающиеся подмножества (кластеры) таким образом, чтобы каждый кластер состоял из близких (похожих) по некоторому критерию объектов, непохожих на объекты других кластеров. В анализе данных эта проблема относится к классу задач обучения без учителя. Особенностью задачи таксономии неклассифицированной выборки является то, что априори неизвестны как принадлежность объектов выборки к тому или иному образу (классу), так и число таких образов. Для решения задачи используется редуцированная функция конкурентного сходства [2]:

$$F^*(z, \mathbf{A}) = \frac{\tau^* - \tau(z, \mathbf{A})}{\tau^* + \tau(z, \mathbf{A})},$$

где $\mathbf{A}$ — выборка, τ^* — константа, интерпретируемая как расстояние от каждого объекта $z \in \mathbf{A}$ до виртуального образа (или образа-конкурента), все объекты которого являются прототипами и расстояние от любого объекта $\mathbf{A}$ до ближайшего объекта образа-конкурента равно τ^*; $\tau(z, \mathbf{A}) = \min\limits_{x \in \mathbf{A} \setminus \{z\}} \tau(z, x)$ — расстояние от объекта z до множества $\mathbf{A} \setminus \{z\}$.

Обозначим через $\mathbf{S_A}$ множество прототипов выборки $\mathbf{A}$. Для оценки качества этого множества используется усредненная величина:

$$H(\mathbf{A}, \mathbf{S_A}) = \frac{1}{|\mathbf{A}|} \sum_{z \in \mathbf{A} \setminus \mathbf{S_A}} F^*(z, \mathbf{S_A}).$$

Для решения рассматриваемой задачи требуется найти множество $\mathbf{S_A}$ прототипов выборки $\mathbf{A}$, на котором достигается максимум функционала H.

Доказательство NP-трудности в сильном смысле задачи таксономии выполнено сведением известной NP-полной задачи о вершинном покрытии графа к

задаче выбора подмножества, на котором значение функционала H максимально.

Задача ВП (вершинное покрытие)[3]. Дан граф $G = (V, E)$ и положительное целое число $J \leqslant |V|$. Имеется ли в графе G вершинное покрытие не более чем из J элементов, то есть такое подмножество $V' \subseteq V$, что $|V'| \leqslant J$ и для каждого ребра $\{u, v\} \in E$ хотя бы одна из вершин u или v принадлежит V'?

В [4] доказана следующая

Теорема. Задача поиска наименьшего вершинного покрытия произвольного графа $G = (V, E)$ сводится к задаче выбора из некоторой искусственной выборки X_G множества объектов $\mathbf{S}^*_{\mathbf{A}}$, на котором достигается максимум функционала H. Причем выборка X_G строится по G за полиномиальное время и имеет полиномиальное количество объектов относительно $|V| + |E|$.

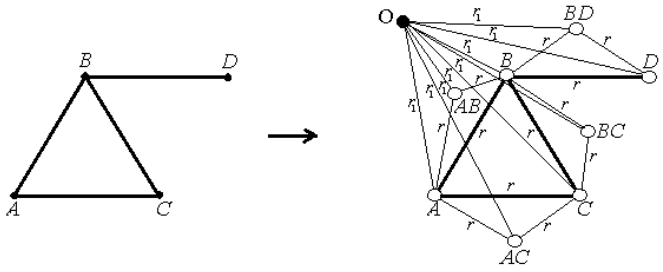


Рис. 1.

На рис. 1 приведено построение искусственной выборки X_G по графу G, состоящему из четырех вершин. Для r, r_1, R выполняются следующие неравенства:

$$0 < r < r_1 < R < 2r, \qquad \frac{2(r_1 - r)}{r_1 + r} + \frac{r_1 - R}{r_1 + R} < 0.$$

Задача выбора подмножества прототипов выборки, состоящей из объектов двух образов, является типичной в анализе данных. Предполагается, что разбиение на два класса задано, и каждый класс может описываться несколькими прототипами. Задачу таксономии можно рассматривать как частный случай данной задачи при условии, что один из образов состоит из одного объекта. В рассматриваемой постановке в качестве прототипов выборки, состоящей из объектов двух классов, выбираются объекты, на которые максимально похожи объекты из того же класса и не похожи объекты другого класса.

Обозначим через $\mathbf{S_A}$ и $\mathbf{S_B}$ множества прототипов образа $\mathbf{A}$ и образа $\mathbf{B}$, соответственно. Для оценки качества выбранных прототипов используется усред-

ненная величина:

$$H(\mathbf{A},\mathbf{S_A},\mathbf{B},\mathbf{S_B}) = \frac{1}{|\mathbf{A}\cup\mathbf{B}|}\left(\sum_{x\in\mathbf{A}\setminus\mathbf{S_A}}\frac{\tau(x,\mathbf{S_B})-\tau(x,\mathbf{S_A})}{\tau(x,\mathbf{S_B})+\tau(x,\mathbf{S_A})} + \sum_{x\in\mathbf{B}\setminus\mathbf{S_B}}\frac{\tau(x,\mathbf{S_A})-\tau(x,\mathbf{S_B})}{\tau(x,\mathbf{S_A})+\tau(x,\mathbf{S_B})}\right).$$

Для решения рассматриваемой задачи требуется найти множества $\mathbf{S_A}$ и $\mathbf{S_B}$ прототипов выборки, на которых достигается максимум функционала H.

Из доказанной теоремы получим

Следствие. Задача поиска множества прототипов выборки, представленной объектами двух классов, NP-трудна.

NP-трудность в сильном смысле рассмотренных задач когнитивного анализа данных обосновывает применение различных эвристических алгоритмов для решения задач выбора прототипов классифицированных, неклассифицированных и смешанных выборок, в которых для количественной оценки качества выбранных типичных объектов образов используется функция конкурентного сходства [5, 6].

Работа выполнена в рамках государственного задания ИМ СО РАН (проект № FWNF–2022–0015).

[1] *Zagoruiko N. G., Borisova I. A., Dyubanov V. V., Kutnenko O. A.* Methods of recognition based on the function of rival similarity // Pattern Recognition and Image Analisys. 2008. Vol. 18, No. 1. P. 1–6.

[2] *Борисова И. А.* Алгоритм таксономии FRiS-Tax // Научный вестник НГТУ. Новосибирск: Изд-во НГТУ. 2007. № 3. С. 3–12.

[3] *М. Гэри, Д. Джонсон.* Вычислительные машины и труднорешаемые задачи. М: Мир, 1982. 416 с.

[4] *Кутненко О. А.* Вычислительная сложность двух задач когнитивного анализа данных // Дискретный анализ и исследование операций. 2022. Т. 29, № 1. С. 18–32.

[5] *Борисова И. А., Дюбанов В. В., Загоруйко Н. Г., Кутненко О. А.* Сходство и компактность // Доклады 14-ой Всероссийской конференции Математические методы распознавания образов, ММРО–14, (Суздаль, Россия, 21–25 сентября, 2009). М: Макс Пресс. 2009. С. 89–92.

[6] *Загоруйко Н. Г., Борисова И. А., Дюбанов В. В., Кутненко О. А.* Количественная мера компактности и сходства в конкурентном пространстве // Сибирский журнал индустриальной математики. 2010. Т. 13, № 1. С. 59–71.

Computational complexity of two problems of cognitive data analysis

Kutnenko Olga[1,2] olga@math.nsc.ru

[1]Novosibirsk, Sobolev Institute of Mathematics

[2]Novosibirsk, Novosibirsk State University

NP-hardness in the strong sense of two problems of cognitive data analysis is proved, that is the problem of taxonomy (clustering) and the problem of selecting a subset of typical representatives of a classified sample consisting of objects of two images. To quantify the quality of the set of selected typical representatives of the sample (prototypes), the FRiS-function (Function of Rival Similarity) [1], is used, which evaluates the similarity of an object with the closest typical sample representative.

The problem of taxonomy (clustering) is the problem of dividing an unclassified sample of objects into non-overlapping subsets (clusters) in such a way that each cluster consists of objects that are close (similar) according to some criterion and dissimilar to the objects of other clusters. In data analysis this problem belongs to the class of unsupervised learning problems. A feature of the problem of taxonomy of an unclassified sample is that it is not known a priori whether the sample objects belong to a particular image (class), or the number of such images. To solve the problem, the reduced function of rival similarity is used [2]:

$$F^*(z, \mathbf{A}) = \frac{\tau^* - \tau(z, \mathbf{A})}{\tau^* + \tau(z, \mathbf{A})},$$

where A is a sample, τ^* is a constant interpreted as the distance from each object $z \in \mathbf{A}$ to the virtual image (or the image of a competitor), all objects of which are prototypes and the distance from any object $\mathbf{A}$ to the nearest object of the competitor image is equal to τ^*; $\tau(z, \mathbf{A}) = \min\limits_{x \in \mathbf{A} \setminus \{z\}} \tau(z, x)$ is the distance from the object z to the set $\mathbf{A} \setminus \{z\}$.

Denote by $\mathbf{S_A}$ the set of prototypes of the sample $\mathbf{A}$. To assess the quality of this set, the average value is used:

$$H(\mathbf{A}, \mathbf{S_A}) = \frac{1}{|\mathbf{A}|} \sum_{z \in \mathbf{A} \setminus \mathbf{S_A}} F^*(z, \mathbf{S_A}).$$

To solve the problem under consideration, it is required to find the set $\mathbf{S_A}$ of prototypes of the sample $\mathbf{A}$, on which the maximum of the functional H is achieved.

The proof of NP-hardness in the strong sense of the taxonomy problem is done by reducing the well-known NP-complete problem of graph vertex cover to the problem of choosing a subset on which the value of the functional H is maximal.

Problem VC (vertex cover)[3]. A graph $G = (V, E)$ and a positive integer number $J \leqslant |V|$ are given. Is there a vertex cover of at most J elements in G, i.e., a

subset $V' \subseteq V$ such that $|V'| \leqslant J$ and for each edge $\{u, v\} \in E$ at least one vertex u or v belongs V'?

The following theorem was proved in [4].

Theorem. The problem of finding the smallest vertex cover of an arbitrary graph $G = (V, E)$ is reduced to the problem of choosing a set of objects $\mathbf{S}^*_{\mathbf{A}}$ from some artificial sample X_G on which the maximum of the functional H is achieved. Moreover, the sample X_G is built according to G in polynomial time and has a polynomial number of objects with respect to $|V| + |E|$.

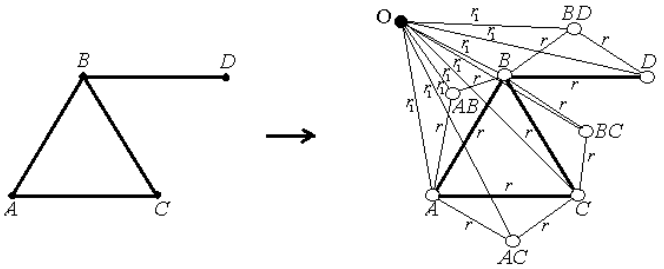


Fig. 1.

Figure 1 shows the construction of an artificial sample X_G according to the graph G, consisting of four vertices. For r, r_1, R the following inequalities hold:

$$0 < r < r_1 < R < 2r, \qquad \frac{2(r_1 - r)}{r_1 + r} + \frac{r_1 - R}{r_1 + R} < 0.$$

The problem of choosing a subset of sample prototypes, consisting of objects of two images, is typical in data analysis. It is assumed that the division into two classes is given, and each class can be described by several prototypes. The problem of taxonomy can be considered as a special case of this problem, provided that one of the images consists of one object. In the formulation under consideration, objects that are most similar to objects from the same class and are not similar to objects of another class, are chosen as prototypes of a sample consisting of objects of two classes.

Denote by $\mathbf{S_A}$ and $\mathbf{S_B}$ the sets of prototypes of pattern $\mathbf{A}$ and pattern $\mathbf{B}$, respectively. To evaluate the quality of the selected prototypes, the average value is

used:

$$H(\mathbf{A},\mathbf{S_A},\mathbf{B},\mathbf{S_B}) = \frac{1}{|\mathbf{A}\cup\mathbf{B}|}\left(\sum_{x\in\mathbf{A}\setminus\mathbf{S_A}}\frac{\tau(x,\mathbf{S_B})-\tau(x,\mathbf{S_A})}{\tau(x,\mathbf{S_B})+\tau(x,\mathbf{S_A})} + \sum_{x\in\mathbf{B}\setminus\mathbf{S_B}}\frac{\tau(x,\mathbf{S_A})-\tau(x,\mathbf{S_B})}{\tau(x,\mathbf{S_A})+\tau(x,\mathbf{S_B})}\right).$$

To solve the problem under consideration, it is required to find the sets $\mathbf{S_A}$ and $\mathbf{S_B}$ of sample prototypes on which the maximum of the functional H is achieved.

From the proved theorem we get

Corollary. The problem of finding a set of prototypes for a sample represented by objects of two classes is NP-hard.

NP-hardness in the strong sense of the considered problems of cognitive data analysis justifies the use of various heuristic algorithms for solving problems of selecting prototypes of classified, unclassified and mixed samples. The function of rival similarity is used to quantify the quality of selected typical objects of images [5, 6].

The research was carried out within the framework of the state order of the Institute of Mathematics of the Siberian Branch of the Russian Academy of Sciences (project no. FWNF-2022-0015).

[1] *Zagoruiko N. G., Borisova I. A., Dyubanov V. V., Kutnenko O. A.* Methods of recognition based on the function of rival similarity // Pattern Recognition and Image Analisys. 2008. Vol. 18, No. 1. P. 1–6.

[2] *I. A. Borisova.* Taxonomy algorithm FRiS-Tax // Nauchnye vectnik NGTU. (Izdatel'stvo NGTU, Novosibirsk, 2007) No. 3, 3–12 [Russian].

[3] *M. R. Garey and D. S. Johnson.* Computers and Intractability: A Guide to the Theory of NP-Completeness (Freeman, San Francisco, 1979; Mir, Moscow, 1982).

[4] *O. A. Kutnenko.* NP-hardness of some data cleaning problem // Diskretnyi analiz i issledovanie operatsii. 2022. Vol. 29, No 1. P. 18–32.

[5] *I. A. Borisova, V. V. Dyubanov, N. G. Zagoruiko, O. A. Kutnenko.* Similarity and Compactness // Proc. All-Russian Conf. Mathematical Methods for Pattern Recognition-14, Suzdal, Russia, September 21–25, 2009. (Maks Press, Moscow, 2009), pp. 89–92.

[6] *N. G. Zagoruiko, I. A. Borisova, V. V. Dyubanov, O. A. Kutnenko.* A quantitative measure of compactness and similarity in a competitive space // Siberian Journal of Industrial Mathematics, **13** (1), 59–71 (2010) [Russian]. Translated in J. Applied and Industrial Mathematics. **5** (1), 144–154 (2011).

Подход к аппроксимации границы невыпуклого множества достижимости управляемой динамической системы

Александр Юрьевич Горнов[1] gornov@icc.ru
Татьяна Сергеевна Зароднюк[1]⋆ tz@icc.ru
[1]Иркутск, Институт динамики систем и теории управления им. В.М. Матросова Сибирского Отделения РАН

Задача построения границы невыпуклых множеств достижимости управляемых динамических систем в общем случае считается в настоящее время нерешенной. Это связано как с объективной сложностью невыпуклых задач динамической оптимизации, так и с высокими требованиями к численным алгоритмам и их программным реализациям. Зная вид множества достижимости управляемой динамической системы, можно сводить исходные задачи оптимального управления к задачам поиска минимума целевого функционала на множестве заданной структуры.

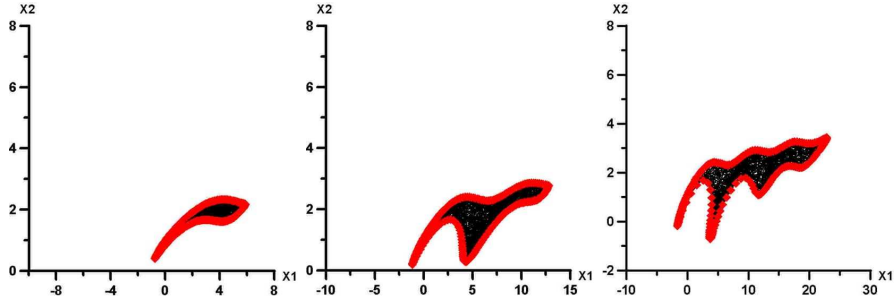
В работе предлагается подход, основанный на построении семейства допустимых управляющих воздействий, позволяющих получать точки на границе множества достижимости. Формируются управляющие воздействия с одной точкой переключения, «пробегающей» весь интервал изменения времени. В соответствии с bang-bang принципом на подобных управлениях достигаются граничные точки множества достижимости.

Программная реализация предлагаемого подхода протестирована с использованием задач с невыпуклыми множествами достижимости. В качестве примера приведем двумерную управляемую динамическую систему, рассмотренную на увеличивающихся интервалах изменения времени:

$$\dot{x}_1 = u_1 + x_2, \tag{1}$$

$$\dot{x}_2 = u_1 + \cos(x_2 - x_1), \tag{2}$$

$$x_1(0) = 0.0,\ x_2(0) = 0.0, |u(t)| \leqslant 0.5,\ t \in T = [0, t_1]. \tag{3}$$



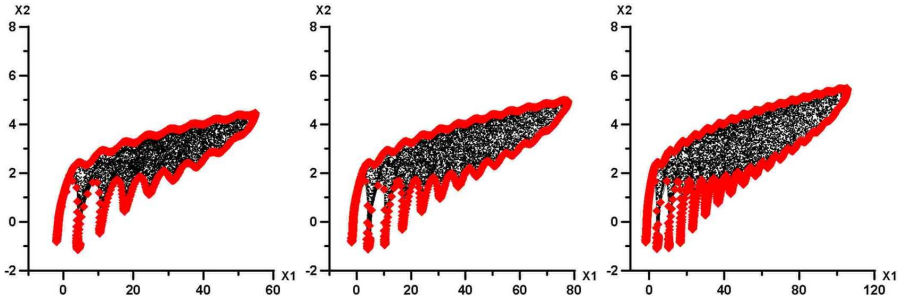


Рис. 1. Множества достижимости управляемой динамической системы (1)–(3), заданной на возрастающих интервалах времени: $t_1 = 2, 3, 5, 7, 9$ и 10.

На рисунке представлена последовательность множеств достижимости для задачи (1)–(3). Соответствующая динамическая система рассмотрена на разных интервалах времени: $[0, 2], [0, 3], [0, 5], [0, 7], [0, 9]$ и $[0, 10]$. Видно, как усложняется структура множества достижимости – появляются узкие области и если экстремальные точки будут находится в этих областях, то могут возникать дополнительные вычислительные трудности при численном поиске решения.

Проведено тестирование программной реализации предложенного подхода. Полученные результаты позволили оценить его применимость для аппроксимации границы множества достижимости управляемых динамических систем.

Работа выполнена за счет субсидии Минобрнауки России в рамках проекта «Теория и методы исследования эволюционных уравнений и управляемых систем с их приложениями» (№ гос. регистрации: 121041300060-4).

[1] *Gornov A., Zarodnyuk T., Anikin A., Sorokovikov P., Tyatyushkin A.* Software engineering for optimal control problems // Lecture Notes in Networks and Systems, 2022. Vol. 424. Pp. 415–426.

An approach to boundary approximation of a non-convex reachable set of controlled dynamical systems

Alexander Gornov[1] gornov@icc.ru
Tatiana Zarodnyuke12⋆ tz@icc.ru
[1]Irkutsk, Matrosov Institute for System Dynamics and Control Theory of Siberian Branch of RAS

The problem of constructing the boundary of non-convex reachable sets of controlled dynamical systems is generally considered to be currently unsolved. This is due both to the objective complexity of non-convex dynamic optimization problems and to the high requirements for numerical algorithms and their software implementations. Knowing the structure of the reachable set of the controlled dynamical system, it is possible to reduce the initial optimal control problems to the problems of finding the minimum of the objective functional on the known type set.

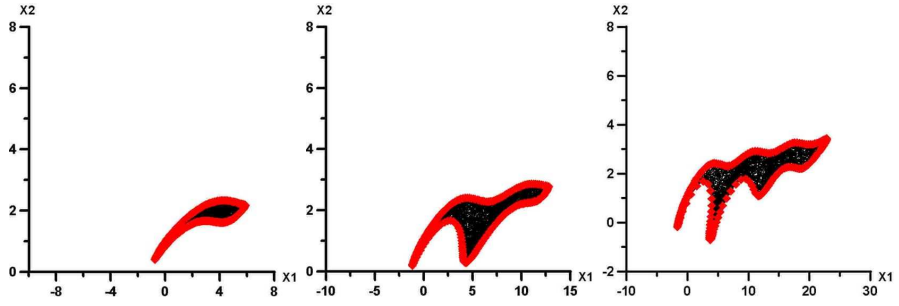
The paper proposes an approach based on the construction of a family of permissible controls that allow obtaining points on the boundary of the reachable set. The controls are formed with a single switching point that "runs through" the entire time interval. In accordance with the bang-bang principle, the boundary points of the reachable set are achieved on such controls.

The software implementation of the proposed approach is tested using problems with non-convex reachable sets. As an example, we give the two-dimensional controlled dynamical system considered at increasing time intervals:

$$\dot{x}_1 = u_1 + x_2, \tag{1}$$

$$\dot{x}_2 = u_1 + \cos(x_2 - x_1), \tag{2}$$

$$x_1(0) = 0.0,\ x_2(0) = 0.0, |u(t)| \leqslant 0.5,\ t \in T = [0, t_1]. \tag{3}$$



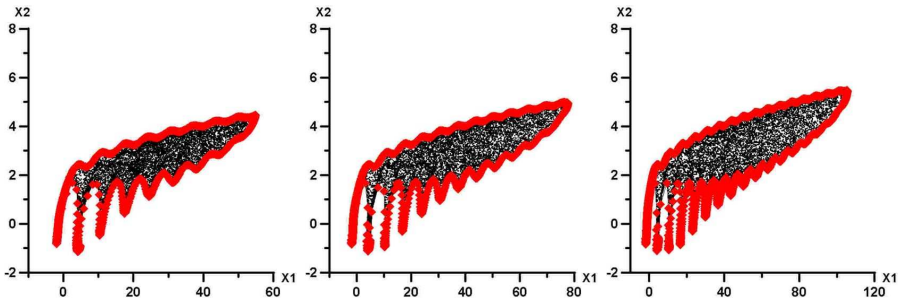


Fig. 1. Reachable sets of the controlled dynamical system (1)–(3), given at increasing time intervals: $t_1 = 2, 3, 5, 7, 9$ and 10.

The figure shows a sequence of reachable sets for problem (1)–(3). The corresponding dynamical system is considered at different time intervals: $[0, 2]$, $[0, 3]$, $[0, 5]$, $[0, 7]$, $[0, 9]$ and $[0, 10]$. It can be seen how the structure of the reachable set becomes more complicated, the narrow areas appear and if the extreme points are located in these areas, then additional computational difficulties may arise in the numerical search for the solution.

The software implementation of the proposed approach has been tested. The obtained results made it possible to evaluate its applicability for approximating the boundary of the reachable set of the controlled dynamical systems.

The work was supported by the grant from the Ministry of Education and Science of Russia within the framework of the project "Theory and methods of research of evolutionary equations and controlled systems with their applications" (state registration No 121041300060-4).

[1] *Gornov A., Zarodnyuk T., Anikin A., Sorokovikov P., Tyatyushkin A.* Software engineering for optimal control problems // Lecture Notes in Networks and Systems, 2022. Vol. 424. Pp. 415–426.

Регуляризация светрочной нейронной сети сингулярным разложенем для обучения на медицинских изображениях

Ваулин Николай Владимирович nvvaulin@gmail.com
Москва, МПГУ

Автоматизированный анализ медицинских изображений является одной из наиболее перспективных областей применения машинного обучения. Задачу классификации медицинских изображений можно рассматривать в контексте общей проблемы классификации в компьютерном зрении. Здесь так же применимы многие стандартные решения, такие как нейросетевые архитектуры, аугментация данных и стратегии обучения. Однако, отсутствие больших обучающих выборок с одновременной необходимостью высокой обобщающей способности модели, заставляет развивать ряд специфичных решений, апеллирующих к недостатку данных. Один из распространенных подходов, применяемых при обучении на малых выборках, является перенос знаний за счет использования предобученной модели. Это позволяет добиться лучшей сходимости и большей обобщающей способности модели, но эффект переобучения все еще значителен. В данной работе было предложено использовать сингулярное разложение весов сверток сети, что позволяет значительно уменьшить количество обучаемых параметров.

Разложение весов уже широко применяется в нейронных сетях. Например, в задаче распознавания речи удалось снизить количество параметров нейронной сети за счет обнуления малых сингулярных значений [1], уменьшив количество параметров в 8 раз и дообучив, точность модели практически не снизилась. Аналогично, снижение ранга матрицы использовалось в задаче дообучения (а именно, в self-supervised distillation). В [2] было предложено дистиллировать корреляцию между входными и выходными значениями блока, при этом внутренняя размерность входных и выходных признаков снижалась за счет обнуления сингулярных значений. Дообучение в режиме дистилляции к изначальной модели привело к увеличению точности.

По аналогии с полносвязными сетями, сингулярное разложение можно применить и к весам двухмерных сверток. В работе были рассмотрены несколько вариантов разложения параметров. Оптимизация только сингулярных значений уменьшает количество обучаемых параметров, что повышает стабильность сходимости и снижает эффект переобучения. Дополнительно, метод дообучения сингулярных значений обнуляет значительное количество параметров что можно использовать в задаче уменьшения количества весов сети.

[1] *Xue, Jian and Li, Jinyu and Gong, Yifan* Restructuring of deep neural network acoustic models with singular value decomposition. // Interspeech, 2013. — p. 2365–2369

[2] *Lee, Seung Hyun and Kim, Dae Ha and Song, Byung Cheol* Self-supervised knowledge distillation using singular value decomposition // Proceedings of the European Conference on Computer Vision (ECCV), 2018. — p. 335–350

Training with SVD regularization for medical imaging

Vaulin Nikolay nvvaulin@gmail.com
Moscow, MPSU

Automated analysis of medical images is one of the most promising applications of machine learning. Medical images classification can be considered in the context of the general problem of classification in computer vision. Many standard solutions are also applicable here, such as neural network architectures, data augmentation and optimization strategies. However, the lack of large training datasets and generalizability requirement for the model makes it necessary to develop specific solutions that targets lack-of-data problems. One of the common approaches used in training on small datasets is knowledge transfer by using pre-trained model. This improve convergence and generalization of the model, but the effect of overfitting is still significant. In this paper, it was proposed to use the singular value decomposition of the convolution weights, which can significantly reduce the number of trainable parameters.

Weight decomposition is already widely used in neural networks. For example, in the speech recognition problem, it was possible to reduce the number of parameters by fine-tuning model with only largest singular values [1], it reduce the number of parameters by 8 times with comerable accuracy. Similarly, matrix rank reduction was used in self-supervised distillation. In [2] it was proposed to distill the correlation between the input and output values of the block, while the internal dimension of the input and output features was reduced by eliminating small singular values. Additional training in the distillation mode to the original model increase accuracy.

By analogy with fully connected networks, the singular value decomposition can also be applied to the weights of two-dimensional convolutions. In proposed approach, several options for parameters decomposition were considered. Oprimizing the only singular values reduces the number of trainable parameters, which increases stability of convergence, reduces the effect of overfitting, and at the same time the network effectively learns a new task. Additionally, the singular value fine-tuning removes significant number of parameters, which can be used weights pruning

[1] *Xue, Jian and Li, Jinyu and Gong, Yifan* Restructuring of deep neural network acoustic models with singular value decomposition. // Interspeech, 2013. — p. 2365–2369

[2] *Lee, Seung Hyun and Kim, Dae Ha and Song, Byung Cheol* Self-supervised knowledge distillation using singular value decomposition // Proceedings of the European Conference on Computer Vision (ECCV), 2018. — p. 335–350

Классификация извлекаемых из панорам изображений нейронной сетью с модулем сдавливания-возбуждения

Филиппских Сергей Леонидович[1] philippsl@mail.ru

[1]Орел, Орловский филиал Федерального исследовательского центра "Информатика и управление" Российской Академии Наук (ОФ ФИЦ ИУ РАН)

В настоящее время все большее распространение получает подход к классификации полноцветных изображений (фотографий) с использованием сверточных нейронных сетей (СНС) и моделируемых шаблонов для распознавания выделяемых особенностей.

При разработке информационной технологии коррекции яркости и цвета (ИТ КЯЦ) панорамных изображений, полученных с помощью аэрофотосъемки, были выполнены задачи по нормализации яркости и цвету выбранных кадров, полученных с камеры БПЛА; сшивание подходящих кадров в панорамы; сравнение разновременных панорам и выявление возникающие аномалии [1]. Однако отнесение идентифицированных объектов к тому или иному классу отдавалось на рассмотрение оператору ПО [2]. Для автоматизации процесса классификации аномалий было принято решение применить современные нейросетевой модели [3].

При работе с данными, полученными с БПЛА, обучающая выборка часто содержит классы, сильно отличающиеся по количеству изображений в каждом. Из-за большого числа изображений малых размеров, нельзя использовать современные архитектуры глубоких и широких нейронных сетей в чистом виде. Использование устаревших архитектур также невозможно, так как они плохо работают на несбалансированных выборках и склонны к переобучению.

Для классификации аномалий был разработан шаблон SOTA-ConvNet. Шаблоны существенно упрощают разработку и проведение экспериментов с нейросетевыми моделями. Для построения нейронной сети необходимо указать три гиперпараметра. Всю остальную архитектуру автоматически достраивает шаблон.

Шаблон SOTA-ConvNet состоит из трех частей: стержневой компонент (stem), основная сверточная база (learner) и классификатор (task). Основная сверточная база – это основной компонент нейросетевой модели, в котором идет усвоение признаков, найденных при анализе изображений. Основная сверточная база состоит из нескольких сверточных групп (точное количество групп является гиперпараметром). Каждая группа состоит из плотного сверточного блока (dense convolution block) и плотного переходного блока (dense transition block). В последней группе отсутствует плотный переходной блок. Нейросетевая модель, построенная на базе шаблона SOTA-ConvNet, имеет следующие характеристики: 4 сверточные группы, коэффициент компрессии 0.8, базовая ширина сети 64 нейрона и 797980 параметров.

Для экспериментов с моделями машинного обучения был выбран набор данных (датасет) VisDrone2022 [4]. Общее количество изображений в наборе данных – 362762. В датасете собраны объекты шести классов (в скобках указана доля изображений каждого класса в процентах): автобусы (2,51 %), легковые автомобили (51,13 %), фотографии случайных участков изображений (играют роль ложных срабатываний при поиске аномалий) (3,36 %), мотоциклы (10,89 %), пешеходы (27,63 %), грузовики (4,48 %).

Результаты классификации изображений нейронной сетью на базе шаблона SOTA-ConvNet (метрика – точность): общая точность – 92,2 %, автобусы – 79,3 %, легковые автомобили – 96,1 %, фотографии случайных участков изображений – 67,0 %, мотоциклы – 88,1 %, пешеходы – 94,4 %, грузовики – 71,2 %.

Для дальнейшего улучшения качества классификации шаблона SOTA-ConvNet добавим в него связи сдавливания-возбуждения (squeeze-excitationlink или SE-link) [5]. SE-link определяет наиболее полезные карты признаков, усиливает их влияние на конечный результат и ослабляет влияние карт признаков, имеющих нейтральное или отрицательное воздействие. SE-link работает как дополнительный регуляризатор. Каждая карта признаков усредняется до скалярного значения, а затем, получившийся вектор пропускается через два полносвязных слоя. Первый слой сжимает информацию (степень сжатия определяется коэффициентом C), второй – распаковывает вектор до первоначального размера. Подобная операция приводит к потере информации и появлению шума. Получившийся вектор умножается на исходные карты признаков.

В нейронной сети на базе шаблона SOTA-ConvNet есть 3 варианта применения SE-link: после сверточных слоев плотного переходного блока (шаблон SE-SOTA-ConvNet), в остаточной связи (identity link) плотного переходного блока (шаблон SE1-SOTA-ConvNet) и в двух вышеперечисленных местах одновременно (шаблон SE2-SOTA-ConvNet).

В процессе исследований были спроектированы три нейросетевые модели с разными вариантами размещения модуля SE-link. Результаты классификации изображений на датасете VisDrone приведены ниже.

Результаты классификации изображений нейронной сетью на базе шаблона SE-SOTA-ConvNet (метрика – точность): общая точность – 94,3 %, автобусы – 83,1 %, легковые автомобили – 97,2 %, фотографии случайных участков изображений – 75,4 %, мотоциклы – 89,0 %, пешеходы – 97,3 %, грузовики – 76,2 %.

Результаты классификации изображений нейронной сетью на базе шаблона SE1-SOTA-ConvNet (метрика – точность): общая точность – 94,4 %, автобусы – 86,0 %, легковые автомобили – 96,5 %, фотографии случайных участков изображений – 77,1 %, мотоциклы – 87,1 %, пешеходы – 97,9 %, грузовики – 83,6 %.

Результаты классификации изображений нейронной сетью на базе шаблона SE2-SOTA-ConvNet (метрика – точность): общая точность – 94,1 %, автобусы – 85,4 %, легковые автомобили – 96,8 %, фотографии случайных участков изображений – 79,6 %, мотоциклы – 86,7 %, пешеходы – 96,7 %, грузовики – 81,4 %.

Все три нейросетевые модели с модулем SE-link показали примерно одинаковые результаты. Точность классификации изображений на малых классах выросла на 4 – 12 %. Однако в сети на базе шаблона SE2-SOTA-ConvNet используется два модуля SE-link вместо одного. Из-за этого данная нейросетевая модель имеет на 140862 параметра больше, чем две другие. Большее число параметров увеличивает время обучения на 20 % и ускоряет переобучение сети. По этой причина для классификации аномалий рекомендуется использовать нейросетевые модели на базе шаблонов SE-SOTA-ConvNet или SE1-SOTA-ConvNet.

[1] Методы и программные средства накопления и обработки данных 2019-2023. Часть 5. Описание предметной области информационной технологии коррекции яркости и цвета при создании панорамных изображений: отчет о НИР (промежуточный) // ФИЦ ИУ РАН; рук. Будзко В.И. – Москва, 2021. — 62 с. – No. ГР АААА-А19-119092490026-1.

[2] *Архипов П. О., Колесник А. В., Цуканов М. В.* Программа для ЭВМ Программная система обнаружения аномалий на разновременных панорамах обследуемой местности (CL_22). Свидетельство о государственной регистрации программы для ЭВМ No. 2022615139. – Зарегистрировано в Реестре программ для ЭВМ 30 марта 2022 г. – 1с.

[3] *Arkhipov P. O., Philippskih S. L.* Building an ensemble of convolutional neural networks for classifying panoramic images // Pattern Recognition and Image Analysis, 2022. Vol. 32, No.3, pp. 511–514.

[4] *Zhu P., Wen L., Du D., Bian X., Fan H., Hu X., Ling H.* Detection and Tracking Meet Drones Challenge // CoRR, 2020.

[5] *Hu J., Shen L., Albanie S., Sun G., Wu E.* Squeeze-and-Excitation Networks // CoRR, 2017.

Classification of images extracted from panoramas using a neural network with a squeeze-excitation module

Philippskih Sergey[1] philippsl@mail.ru

[1]Orel, Orel branch of Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences (OB FRC CSC RAS)

At present, an approach to the classification of full-color images (photos) using convolutional neural networks (CNN) and simulated patterns for recognizing distinguished features is becoming more widespread.

When developing the information technology for brightness and color correction (IT BCC) of panoramic images obtained using aerial photography, the tasks were performed to normalize the brightness and color of selected frames received from the UAV camera; stitching suitable frames into panoramas; comparison of panoramas at different times and identification of emerging anomalies [1]. However, the assignment of identified objects to a particular class was given to the software operator for consideration [2]. To automate the process of classifying anomalies, it was decided to apply modern neural network models [3].

When working with UAV data, the training sample often contains classes that differ greatly in the number of images in each. Due to the large number of images of small sizes, it is impossible to use modern architectures of deep and wide neural networks in their pure form. The use of outdated architectures is also impossible, since they do not work well on unbalanced samples and are prone to overfitting.

The SOTA-ConvNet template was developed to classify anomalies. Templates greatly simplify the development and experimentation with neural network models. To build a neural network, you must specify three hyperparameters. The rest of the architecture is automatically completed by the template.

The SOTA-ConvNet template consists of three parts: a stem, a learner, and a task. The learner is the main component of the neural network model, in which the assimilation of features, found during image analysis, takes place. The learner consists of several convolution groups (the exact number of groups is a hyperparameter). Each group consists of a dense convolution block and a dense transition block. The last group lacks a dense transition block.

The neural network model built on the basis of the SOTA-ConvNet template has the following characteristics: 4 learners, a compression ratio of 0.8, a base network width of 64 neurons, and 797,980 parameters.

For experiments with machine learning models, the data set VisDrone2022 [4] was chosen. The total number of images in the data set is 362,762. The dataset contains objects of six classes (in parentheses, the percentage of images of each class is indicated): buses (2.51%), cars (51.13%), photographs of random sections of images (play the role of false positives in the search for anomalies) (3.36%), motorcycles (10.89%), pedestrians (27.63%), trucks (4.48%).

The results of image classification by a neural network based on the SOTA-ConvNet template (metric – accuracy): overall accuracy – 92.2%, buses – 79.3%, cars – 96.1%, photographs of random sections of images – 67.0%, motorcycles – 88.1%, pedestrians – 94.4%, trucks – 71.2%.

To further improve the classification quality of the SOTA-ConvNet template, we will add squeeze-excitation links or SE-links to it [5]. SE-link identifies the most useful feature maps, amplifies their impact on the final result, and reduces the influence of feature maps that have a neutral or negative impact. SE-link works as an additional regularizer. Each feature map is averaged to a scalar value, and then the resulting vector is passed through two fully dense layers. The first layer compresses the information (the degree of compression is determined by the coefficient C), the second one decompresses the vector to its original size. Such an operation leads to loss of information and the appearance of noise. The resulting vector is multiplied by the original feature maps.

In a neural network based on the SOTA-ConvNet template, there are 3 options for using SE-link: after the convolution layers of the dense transition block (SE-SOTA-ConvNet template), in the identity link of the dense transition block (template SE1-SOTA-ConvNet) and in the above two places at the same time (template SE2-SOTA-ConvNet).

In the process of research, three neural network models were designed with different placement options for the SE-link module. The results of image classification on the VisDrone dataset are shown below.

Results of image classification by a neural network based on the SE-SOTA-ConvNet template (metric – accuracy): overall accuracy – 94.3%, buses – 83.1%, cars – 97.2%, photographs of random sections of images – 75.4 %, motorcycles – 89.0%, pedestrians – 97.3%, trucks – 76.2%.

Results of image classification by a neural network based on the SE1-SOTA-ConvNet template (metric – accuracy): overall accuracy – 94.4%, buses – 86.0%, cars – 96.5%, photographs of random sections of images – 77.1 %, motorcycles – 87.1%, pedestrians – 97.9%, trucks – 83.6%.

Results of image classification by a neural network based on the SE2-SOTA-ConvNet template (metric – accuracy): overall accuracy – 94.1%, buses – 85.4%, cars – 96.8%, photographs of random sections of images – 79.6 %, motorcycles – 86.7%, pedestrians – 96.7%, trucks – 81.4%.

All three neural network models with the SE-link module showed approximately the same results. The accuracy of image classification in small classes increased by 4–12%. However, a network based on the SE2-SOTA-ConvNet pattern uses two SE-link modules instead of one. Because of this, this neural network model has 140,862 more parameters than another two. More parameters increase training time by 20% and speed up network retraining. For this reason, it is recommended to use neural network models based on the SE-SOTA-ConvNet or SE1-SOTA-ConvNet templates to classify anomalies.

[1] Methods and software for data accumulation and processing 2019-2023. Part 5. Description of the subject area of information technology for brightness and color correction when creating panoramic images: research report (interim) // FRC CSC RAS; sup. Budzko V.I. – Moscow, 2021. — 62 p. – No. GR AAAA-A19-119092490026-1.

[2] *Arkhipov P. O., Kolesnik A. V., Tsukanov M. V.* Computer program Software system for detecting anomalies on multi-temporal panoramas of the surveyed area (CL_22). Certificate of state registration of the computer program No. 2022615139. – Registered in the Register of computer programs on March 30, 2022 – 1p.

[3] *Arkhipov P. O., Philippskih S. L.* Building an ensemble of convolutional neural networks for classifying panoramic images // Pattern Recognition and Image Analysis, 2022. Vol. 32, No.3, pp. 511–514.

[4] *Zhu P., Wen L., Du D., Bian X., Fan H., Hu X., Ling H.* Detection and Tracking Meet Drones Challenge // CoRR, 2020.

[5] *Hu J., Shen L., Albanie S., Sun G., Wu E.* Squeeze-and-Excitation Networks // CoRR, 2017.

Автоматизация анализа изображений с электронного микроскопа на основе метода экспоненциальной аппроксимции

Сулимова Валентина Вячеславовна[1] vsulimova@yandex.ru
Курбаков Михаил Юрьевич[1]⋆ muwsik@mail.ru
Середин Олег Сергеевич[1] oseredin@yandex.ru
Копылов Андрей Валериевич[1] and.kopylov@gmail.com
[1]Тула, Тульский Государственный Университет

Электронная микроскопия позволяет осуществлять увеличение объектов в миллионы раз и фксировать результат на изображениях, в результате чего находит активное применение для исследования микроскопических объектов [1].

В частности, одна из актуальных задач анализа изображений, полученных с электронного микроскопа, связана с анализом количества, размеров и взаимного расположения металлических наночастиц размера 1-5 нм, наносимых на поверхность исследуемых углеродных материалов. Полученная в результате такого анализа информация используется для определения значимых для многих промышленных процессов характеристик исследуемых материалов, выявления их особенностей, в частности, для выявления на поверхности катализаторов [2,3] скрытых дефектов, влияющих на их свойства, но не поддающихся обнаружению другими методами [4].

При этом для исследования только одного образца материала могут формироваться сотни и даже тысячи изображений, требуя автоматизации процесса детектирования наночастиц и последующего анализа их характеристик и взаимного расположения.

Визуально наночастицы представляют собой небольшие светлые области, как правило, округлой формы. Сложность их автоматического обнаружения заключается в том, что изображения, как правило, сильно зашумлены, а сами наночастицы имеют очень небольшой размер и, к тому же, несколько наночастиц могут располагаться очень близко и даже частично перекрывать друг друга.

В [5] нами предложен метод экспоненциальной аппроксимации для обнаружения наночастиц на изображении, осованный на аппроксимации небольших перекрывающихся фрагментов исследуемого изображения при помощи экспоненциальной функции яркости. Экспериментальное исследование на изображениях, размеченных вручную экспертами, показало, что предложенный подход позволяет получить наиболее высокое качество обнаружения наночастиц по сравнению с другими методами, включая нейросетевой подход [6].

Во второй части исследования мы применяем метод экспоненциальной аппроксимации для детектирования наночастиц для базе, содержащей изображения, полученные с электронного микроскопа для двух похожих материалов, а также проводим статистический анализ характеристик обнаруженных наноча-

стиц. Проведенный анализ показал, что статистические характериастики обнаруженных наночастиц могут служить признаком для дифференциации этих материалов, что является важным, в частности, для экспериментальных исследований в области изучения катализа.

Авторы благодарят Научную школу акад. В.П. Ананикова за тематику применения машинных методов в нанотехнологиях, полезные дискуссии и предоставленные экспериментальные данные.

Работа выполнена при финансовой поддержке Министерства науки и высшего образования РФ в рамках государственного задания FEWG-2021-0012.

[1] *Илюшин А.С., Орешко А.П.* Введение в дифракционный структурный анализ // Учебное пособие, Москва: физический факультет МГУ, 2008. — С. 335.

[2] *Liu X., Astruc D.* Development of the Applications of Palladium on Charcoal in Organic Synthesis // Advanced Synthesis and Catalysis, 2018. — С. 3818–3818.

[3] *Felpin F., Ayad T., Mitra S.* Pd/C: An Old Catalyst for New Applications – Its Use for the Suzuki–Miyaura Reaction. // European Journal of Organic Chemistry, 2006. — С. 2679–2690.

[4] *Boiko D.A., Pentsak E.O., Cherepanova V.A., Gordeev E.G., Ananikov V.P.* Deep neural network analysis of nanoparticle ordering to identify defects in layered carbon materials // Chem. Sci. 12, 2021. — С. 7428–7441.

[5] *Kopylov A.V., Kurbakov M.Y., Seredin O.S., Sulimova V.V., Boko D.A., Cherepanova V.A., Pentsak E.O., Ananikov V.P.* Automated recognition of nanoparticles in SEM images of Pd/C catalysts // Nanomaterials, 2022.

[6] *Paszke A.* PyTorch: An Imperative Style, High-Performance Deep Learning Library // Advances in Neural Information Processing Systems 32, 2019. — С. 8024–8035.

Automatic electron microscopes images analysis based on the exponential approximation method

Sulimova Valentina [1] vsuliova@yandex.ru
Kurbakov Mikhail[1]⋆ muwsik@mail.ru
Seredin Oleg[1] oseredin@yan-dex.ru
Kopylov Andrey[1] and.kopylov@gmail.com
[1]Tula, Tula State University

Electron microscopy makes it possible to magnify objects millions of times and fix the result on images, because of which it is actively used to study microscopic objects [1].

In particular, one of the urgent tasks of scanning electron microscope (SEM)'s images anaysis is related to the analysis of the number, size and relative position of 1-5 nm-sized metal nanoparticles, deposited on the surface of the carbon materials. The respective analysis helps to determine the significant for many industrial processes characteristics of the materials under study, to identify their features, in particular, to identify hidden defects on the surface of catalysts [2,3], that affect their properties, but cannot be detected by other methods[4].

At that, hundreds and even thousands of images can be formed to study only one material sample, requiring automation of the process of detecting nanoparticles and subsequent analysis of their characteristics and relative position.

Visually, nanoparticles are small bright areas, usually rounded. The complexity of their automatic detection lies in the fact that the images, as a rule, are very noisy, and the nanoparticles themselves are very small in size, and, moreover, several nanoparticles can be located very close and even partially overlap each other.

In [5], we proposed an exponential approximation method for detecting nanoparticles in SEM images. It is based on the approximation of small overlapping fragments of a studied image using an exponential brightness function. An experimental study on images marked manually by experts showed that the proposed approach allows obtaining the highest quality of nanoparticle detection in comparison with other methods, including the neural network approach [6].

In the second part of the study, we apply the exponential approximation method to detect nanoparticles for a database of SEM images obtained for two similar materials, and also perform a statistical analysis of the characteristics of the detected nanoparticles. The analysis performed showed that the statistical characteristics of the detected nanoparticles can differentiate these materials from each other, which is important, in particular, for experimental studies in the field of catalysis.

The authors thank the Scientific School of Autocad. V.P. Ananikov for the topic of application of machine learning methods in nanotechnologies, useful discussions and provided experimental data.

This research supported by Ministry of Science and Higher Education of the Russian Federation within the framework of the state task FEWG-2021-0012.

[1] *Ilyushin A., Oreshko A.* Introduction to Diffractive Structural Analysis // Tutorial, Moscow: Faculty of Physics, Moscow State University, 2008. — p. 335.

[2] *Liu X., Astruc D.* Development of the Applications of Palladium on Charcoal in Organic Synthesis // Advanced Synthesis and Catalysis, 2018. — p. 3818–3818.

[3] *Felpin F., Ayad T., Mitra S.* Pd/C: An Old Catalyst for New Applications – Its Use for the Suzuki–Miyaura Reaction. // European Journal of Organic Chemistry, 2006. — p. 2679–2690.

[4] *Boiko D.A., Pentsak E.O., Cherepanova V.A., Gordeev E.G., Ananikov V.P.* Deep neural network analysis of nanoparticle ordering to identify defects in layered carbon materials // Chem. Sci. 12, 2021. — p. 7428–7441.

[5] *Kopylov A.V., Kurbakov M.Y., Seredin O.S., Sulimova V.V., Boko D.A., Cherepanova V.A., Pentsak E.O., Ananikov V.P.* Automated recognition of nanoparticles in SEM images of Pd/C catalysts // Nanomaterials, 2022.

[6] *Paszke A.* PyTorch: An Imperative Style, High-Performance Deep Learning Library // Advances in Neural Information Processing Systems 32, 2019. — p. 8024–8035.

Концепция динамически структурированного изображения и объектов на изображении

Харинов Михаил Вячеславович[1]⋆ author_khar@iias.spb.su

[1]Санкт-Петербург, СПб ФИЦ РАН

В статье для адекватного компьютерного детектирования формализуется понятие *объектов*, которые описываются динамически структурированными множествами пикселей [1].

Множество пикселей считается *динамически* структурированным, если:

1. Представляется последовательностью приближений в 1, 2,. . . и т.д. цветах.
2. Описывается выпуклой зависимостью E_g суммарной квадратичной ошибки кусочно-постоянного приближения от числа g цветов в приближении изображения.
3. Поддерживается обратимое слияние таких множеств.

Адекватное определение объектов на изображении строится в рамках постановки и исследования оптимизационно-аппроксимационной задачи аппроксимации последовательных оптимальных приближений изображения, соответствующих последовательности наименьших *ошибок аппроксимации* (суммарных квадратичных ошибок), к которым приближается мажорантная последовательность ошибок бинарной иерархии субоптимальных приближений [1, 2].

Обе последовательности ошибок аппроксимации выпуклы, что ограничивает различие между ними сверху. Поскольку последовательность оптимальных приближений, вообще говоря, неиерархична, она не совпадает с бинарной иерархией субоптимальных приближений. Поэтому целевая бинарная иерархия строится неоднозначно и параметризуется подходящим управляющим параметром.

Динамически структурированные *объекты* состоят из вложенных меньших объектов и составляют бинарную иерархию приближений изображения.

Неиерархическая последовательность оптимальных приближений, описываемая выпуклой последовательностью E_g, представляет *изображение*.

Путем пересечения оптимальных разбиений изображения в расширяющемся диапазоне цветов генерируется нерегулярная иерархия *суперпикселей* (неделимых элементов изображения) таким образом, что оптимальные приближения в ограниченном числе цветов воспроизводятся путем слияния установленного числа суперпикселей. Разделение изображения на *суперпиксели* выбирается из числа приближений, которые составляют нерегулярную иерархию и описываются невыпуклой последовательностью E_g ошибок аппроксимации E.

Программная реализация концепции динамически структурированных объектов осуществляется тремя классическими методами кластерного анализа, а именно:

1. Кластеризацией пикселей по Уорду.
2. Методами разделения/слияния.

3. Методом K-средних.

Вместо перечисленных методов предлагается использовать их усиленные версии, а именно:

1. Обобщенную кластеризацию пикселей по Уорду по частям изображения.
2. Метод разделения/(обратимого слияния) (метод CI-Clustering Improvement).
3. метод K-meanless, предложенный С. Д. Двоенко.

Как работает концепция динамически структурированных множеств пикселей, иллюстрируется обнаружением объектов по их примерам на дополнительном изображении, которое объединяется с рассматриваемым изображением в единое целое.

Новизна исследования заключается в том, что предложено следующее.

1. Экспериментальное подтверждение эффективности аппроксимации изображения иерархией приближений, описываемой выпуклой последовательностью ошибок аппроксимации. Это объясняется тем, что последовательность минимальных ошибок оптимальных приближений изображения выпукла, как предполагается в концепции и, по нашему опыту, оказывается справедливым для реальных изображений.
2. Уточнение и модернизация классических методов кластерного анализа в сочетании друг с другом.
3. Формализация, интерпретация и количественная оценка эффекта упрощения детектирования объектов при объединении распознаваемого изображения и изображения с примерами объектов в единое совместное изображение.
4. Постановка и реализация программного эксперимента на примерах реальных изображений.

Исследование проводится в рамках модели кластеризации пикселей, которая, помимо концептуальных положений и модернизированных методов, объединяет следующие наработки.

1. Формализация понятий структурированных изображений, объектов и суперпикселей (элементов изображения), отличаемых компьютером друг от друга.
2. Постановка и разработка решения задачи аппроксимации неиерархической последовательности оптимальных приближений с помощью бинарной иерархии приближений, которая, как и оптимальные приближения, описывается выпуклой последовательностью ошибок аппроксимации и содержит оптимальное приближение в заданном числе цветов.
3. Определение и примеры вычисления нерегулярной иерархии суперпикселей, обеспечивающие безошибочное получение ряда оптимальных приближений путем слияния суперпикселей.
4. Вычисления с использованием структуры данных алгебраической многослойной сети (АМС), которая обеспечивает высокоскоростную кластеризацию пикселей изображения в терминах динамических деревьев Слейтора-Тарьяна и циклических графов.

Теоретическая значимость исследования заключается в том, что оно полезно не только для обработки изображений, но и для общего кластерного анализа, в котором системы методов минимизации ошибки аппроксимации или, что то же самое, стандартного отклонения не отработаны в совершенстве.

Практическая значимость исследования заключается в том, что обосновывается актуальность модернизации программных реализаций классических методов минимизации в общедоступных программных средствах, таких как MatLab.

Современные вычислительные средства по памяти и быстродействию превосходят ресурсы естественного зрительного восприятия, но по эффективности значительно уступают живым организмам. Для того, чтобы преодолеть отставание, необходимо найти адекватные методы обработки пиксельных множеств, которые обычно разрабатываются в процессе решения конкретных прикладных задач. При этом условию получения адекватных результатов сопутствует условие скоростных вычислений. При теоретических исследованиях имеет смысл сначала понять, как адекватно моделировать естественное зрительное восприятие, а затем, как это делать быстро.

Если игнорировать требование быстродействия, то для вычислений в рамках концепции динамически структурированных пиксельных множеств достаточно одного метода Уорда, реализованного в единственной программе, которая на выходе генерирует последовательность N иерархий приближений изображения из N пикселей. Так как исходные коды и исполняемые модули программ автор размещает в свободном доступе на своих веб-страницах сайтов ResearchGate и MachineLearning, то предлагаемую концепцию можно проверить на практике.

По поводу вычислений в реальном времени следует указать, что, на сегодняшний день, детально разработаны два механизма ускорения кластеризации по Уорду за счет укрупнения начальных множеств пикселей и за счет разделения изображения на части. В настоящее время разрабатывается третий механизм ускорения вычислений с учетом спектральных свойств изображения. Результаты исследования планируется опубликовать в очередной статье.

Работа выполнена в рамках бюджетной темы FFZF-2022-0006 “Теоретические и технологические основы оперативной обработки потоков больших гетерогенных данных в социокиберфизических системах”.

[1] *Kharinov M.* An Object in an Image as a Dynamically Structured Pixel Set // Pattern Recognition and Image Analysis, 32(3), Pleiades Publishing, Ltd., 2022. — p. 561–569.

[2] *Nenashev V., Khanykov I., Kharinov M.* A Model of Pixel and Superpixel Clustering for Object Detection // Journal of Imaging, 8(10), MDPI, 2022. — 274.

A Concept of Dynamically Structured Image and Objects in an Image

Kharinov Mikhail[1]⋆ khar@iias.spb.su

[1]St. Petersburg, SPC RAS

In the paper, for adequate computer detection, the concept of an *object* as a dynamically structured pixel set is formalized [1].

A pixel set is considered *dynamically* structured if:

1. It is represented by a sequence of approximations in 1, 2,... etc. colors.
2. It is described by a convex dependence E_g of total squared errors of piecewise constant image approximations on the number g of their colors.
3. Reversible merging of such sets is supported.

Attempts to adequately determine the object in the image are made within the framework of the formulation and study of the optimization-approximation problem of simulating successive optimal image approximations corresponding to the sequence of the smallest *approximation errors* (total squared errors), which are approached by the majorant sequence of errors of the binary suboptimal approximation hierarchy [1, 2]. Both sequences of approximation errors are convex, which limits the difference between them from above. Since the sequence of optimal approximations is, generally speaking, non-hierarchical, it does not coincide with the binary hierarchy of suboptimal approximations. Therefore, the target binary hierarchy is constructed ambiguously and parameterized by a suitable control parameter.

Dynamically structured *objects*, consisting of nested ones, are represented by binary hierarchy of image approximations.

The non-hierarchical sequence of optimal approximations, described by convex E_g sequence, represents the *image*.

By intersecting partitions of optimal image approximations in an expanding range of colors, an irregular hierarchy of *superpixels* (indivisible image elements) is generated, so that the limited series of optimal approximations can be accurately reproduced by merging of given number of superpixels. The image partitions into *superpixels* is chosen from a number of image approximations constituting an irregular hierarchy, described by non-convex sequence E_g of approximation errors E.

The software implementation of the concept of a dynamically structured object is supported by three classical methods of cluster analysis, namely:

1. Ward's pixel clustering.
2. Splitting/merging techniques.
3. K-means method.

by modernizing and jointly applying them.

As a result, it is proposed to use strengthened versions as the listed methods, namely:

1. Generalized Ward's pixel clustering by image parts.
2. Split/(reversible merge) method (CI-Clustering Improvement method).
3. Dvoenko's K-meanless method.

How does it works is illustrated by detecting objects identified by known instances in an additional image merged with the image in question into a single whole.

The novelty of the study lies in the fact that the following is proposed.

1. Experimental confirmation of the efficiency of image approximation by an approximation hierarchies described by a convex sequences of approximation errors. This is because the sequence of minimum errors of optimal image approximations is convex, as assumed in the concept and, in our experience, turns out to be true for real images.
2. Refining and modernizing the classical methods of cluster analysis in combination with each other.
3. Formalization, interpretation and quantification of the effect of simplifying object detection when combining a recognized image and an image with sample objects into a single joint image.
4. Developing up and implementing a software experiment using examples of real images.

The study is carried out within the framework of the pixel clustering model, which, in addition to conceptual provisions and modernized methods, combines the following developments.

1. Formalization of the concepts of structured images, objects and superpixels (image elements) that are distinguished by a computer from each other.
2. Stating and developing the solution to the problem of approaching a non-hierarchical sequence of optimal image approximations using a binary hierarchy of approximations, which, like optimal approximations, are described by a convex sequence of approximation errors and contain the optimal approximation with a given number of colors.
3. Definition and examples of computing an irregular hierarchy of superpixels, providing error-free obtaining of a number of optimal approximations by merging superpixels.
4. Computations using the Algebraic Multilayer Network (AMN) data structure, which provides high-speed clustering of image pixels in terms of dynamic Sleator-Tarjan trees and cyclic graphs.

The theoretical significance of the study lies in the fact that it is useful not only for image processing, but also for general cluster analysis, in which method systems for minimizing the approximation error or, which is the same most, the standard deviation are not perfectly developed.

The practical significance of the study lies in the fact that it substantiates the relevance of modernizing software implementations of classical minimization methods in publicly available software tools such as MatLab.

Modern computing facilities in terms of memory and speed surpass the resources of natural visual perception, but in terms of efficiency they are significantly inferior to living organisms. In order to overcome the lag, it is necessary to find adequate methods for processing pixel sets, which are usually developed in the process of solving specific applied problems. In this case, the condition of obtaining adequate results is accompanied by the condition of high-speed calculations. In theoretical studies, it makes sense to first understand how to adequately simulate natural visual perception, and then how to do this quickly.

If the speed requirement is ignored , then for calculations within the framework of the concept of dynamically structured pixel sets, one Ward method is sufficient, implemented in a single program that at output generates a sequence of N image approximation hierarchies, where N is the number of pixels in the image. Since the source codes and executable modules of programs, the author places in free access on its web pages of sites ResearchGate and MachineLearning, the proposed concept can be tested in practice by those whom it may concern.

With regard to real-time calculations, it should be noted that, to date, two mechanisms for accelerating Ward's clustering due to enlargement of initial pixel sets and by dividing the image into parts have been detaily developed . Currently, a third mechanism is being developed to speed up calculations, taking into account the spectral image properties. The results of the study are planned to be published in the next article.

This research was funded within the framework of the budgetary theme FFZF-2022-0006 "Theoretical and Technological Basis for Operational Processing of Big Heterogeneous Data Streams in Sociocyberphysical Systems".

[1] *Kharinov M.* An Object in an Image as a Dynamically Structured Pixel Set // Pattern Recognition and Image Analysis, 32(3), Pleiades Publishing, Ltd., 2022. — p. 561–569.

[2] *Nenashev V., Khanykov I., Kharinov M.* A Model of Pixel and Superpixel Clustering for Object Detection // Journal of Imaging, 8(10), MDPI, 2022. — 274.

Разработка технологии выделения области хориоидеи и ее количественного анализа на ОКТ изображениях для диагностики заболеваний глаза

Логинова Наталья Александровна[2]⋆ natashalogin99@gmail.com
Ильясова Наталья Юрьевна[1,2] ilyasova.nata@gmail.com
Демин Никита Сергеевич[1,2] volfgunus@gmail.com
[1]Самара, Институт систем обработки изображений - филиал ФНИЦ «Кристаллография и фотоника» РАН
[2]Самара, Самарский национальный исследовательский университет им. академика С.П. Королева

Изучение анатомических особенностей хориоидеи является основополагающей ролью во многих процессах функционирования и развития глаза и зрения. Хориоидея – богатая сосудами ткань организма, обеспечивает кислородом и питательными веществами пигментный эпителий и наружные слои сетчатки, поддерживает внутриглазное давление и температуру глазного яблока, принимает участие в фокусировке изображения на сетчатке, поэтому дефекты хориоидеи являются провоцирующим фактором ряда заболеваний глаза [1]. Диагностика хориоидеи проводится с помощью оптической когерентной томографии (ОКТ). Для количественной оценки хориоидеи были разработаны различные критерии оценки, включая толщину хориоидеи, объем хориоидеи и хориоидальный сосудистый индекс (CVI). Визуализация хориоидеи с помощью ОКТ является сложной задачей из-за маскирующего эффекта относительно непрозрачного пигментного эпителия сетчатки, традиционно визуализированные изображения ОКТ заднего сегмента содержат тени, которые влияют на визуализацию глубоких структур, в том числе хориоидеи [2]. Это затрудняет качественную оценку параметров хориоидеи. Задачей работы является разработка технологии выделения области хориоидеи и ее количественного анализа на ОКТ изображениях для выявления эндокринной офтальмопатии.

В данной работе предлагается технология выделения зоны интереса и подсчета хориоидального сосудистого индекса, основными этапами которой являются: 1) теневая компенсация, 2) бинаризация снимка, 3) фильтрация бинаризованного изображения, 4) подсчет хориоидального индекса.

Количественный анализ сосудистой оболочки затруднен из-за теней, отбрасываемых передними структурами, такими как сосуды сетчатки. Поэтому компенсация теней важна для успешного и надежного количественного ОКТ – анализа, особенно сосудистой оболочки и других субретинальных структур пигментного эпителия [2,3]. Компенсация теней выполнялась с использованием алгоритма Жирара. Каждое B – сканирование преобразуется в область необработанной интенсивности, а интенсивность каждого пикселя умножается на уникальный коэффициент компенсации, полученный на основе особенностей получения сигнала ОКТ. Этот метод также повышает контрастность изображения

кровеносных сосудов [4]. Непосредственно теневая компенсация изображения определялась следующим уравнением:

$$\frac{s(z)}{2\int_z^\infty s(u)du} = ar_s(z) = HDs(z),$$

где HD – оператор, переводящий заданный сигнал s(z) в его скомпенсированную форму, a – коэффициент затухания. Усиление контраста выполнялось до теневой компенсации изображения, которое описывается следующей формулой:

$$HDI^n(z) = \frac{I^n(z)}{2\int_z^\infty I^n(u)du},$$

где I – интенсивность пикселей, n – показатель степени.

Для выделения области просвета сосудов использовалась бинарная обработка. Первым шагом выполнялось адаптивное выравнивание гистограмм на уровне отдельных B-сканов для увеличения контраста сосудов сосудистой оболочки. Далее осуществлялась бинаризация изображения. Из-за разности распределения яркостей между отдельными снимками применение глобального порога проблематично. В связи с этим для бинаризации использовался метод адаптивной пороговой обработки Ниблэка. Изображение обрабатывается с помощью окна. Для каждого положения окна считается свой порог в соответствии с формулой:

$$treshold = mean + k \cdot s,$$

где mean – среднее значение яркости пикселей в окне, s – стандартное отклонение значений яркости в окне, а k – коэффициент для отделения объекта от фона. Соответственно алгоритм бинаризации имеет 2 параметра: размер окна и коэффициент k. В ходе исследований наулучшее среднее значение для максимальной разделимости классов дало окно размером 41 пиксель и параметр k=0.03.

Для избавления от шумов на бинаризованном изображении использовались методы морфологической обработки изображений [5]. Проводилось вскрытие (последовательное применение операций эрозии и дилатации) изображения с структурным элементом вида эллипс размером 5x5.

Хориоидальный сосудистый индекс определяентся как отношение площади просвета сосудов к общей площади сосудистого слоя. Расчеты проводились для изображений ОКТ с патологией и без патологии. Индекс рассчитывался на ограниченной области по 1000 нм влево и вправо от фовевы. Значения считались для всего набора данных для каждого класса, определялось среднее значение индекса для класса и стандартное отклонение. Исследования показали, что среднее значение индекса для класса нормы равно 0,52, а для патологии 0,56, при СКО 0,02.

В ходе работы была предложена технология выделения области хориоидеи и определения хориоидального сосудистого индекса на изображениях оптической когерентной томографии для выявления эндокринной офтальмопатии. Хориоидея является одной из наиболее васкуляризированных структур человеческого тела и играет незаменимую роль в питании фоторецепторов [6]. Таким образом, развитие и исследование новых методов хориоидеи играет важную роль в диагностике ЭОП, это даст возможность увеличить вероятность выявления у пациентов признаков ЭОП на более ранних стадиях.

[1] *Кудашева Г. Р.* Структурно-функциональные особенности хориоидеи в норме и при патологии зрительной системы, оптимальные методы исследования // Современные технологии в офтальмологии, Москва: Офтальмология, 2020. — С. 333–339.

[2] *Vupparaboina K. K.* Quantitative shadow compensated optical coherence tomography of choroidal vasculature // Scientific Reports, 2018. Vol. 8. No.6461.

[3] *Cheong H.* OCT-GAN: single step shadow and noise removal from optical coherence tomography images of the human optic nerve head // Biomed. Opt. Express, 2021. Vol. 12. — p. 1482–1498.

[4] *Girard M. J.* Shadow removal and contrast enhancement in optical coherence tomography images of the human optic nerve head // Investigative Ophthalmology & Visual Science, 2011. Vol. 52. — p. 7738–7742.

[5] *Шагалова П. А.* Исследование алгоритмов предобработки изображений для повышения эффективности распознавания медицинских снимков // Труды НГТУ им. Р.Е. Алексеева, Нижний Новгород: Нижегородский государственный технический университет им. Р.Е. Алексеева, 2020. — С. 25–32.

[6] *Agrawal M.* Exploring choroidal angioarchitecture in health and disease using choroidal vascularity index // Progress in Retinal and Eye Research, 2020. Vol. 77. No.. 100829.

Development of a technology for the selection of the choroidal region and its quantitative analysis on OCT images for the diagnosis of eye diseases

Loginova Natalya[2]⋆ natashalogin99@gmail.com
Ilyasova Natalya[1,2] ilyasova.nata@gmail.com
Demin Nikita[1,2] volfgunus@gmail.com
[1]Samara, IPSI RAS – Branch of the FSRC "Crystallography and Photonics" RAS
[2]Samara, Samara National Research University

The study of the anatomical features of the choroid is a fundamental role in many processes of functioning and development of the eye and vision. The choroid is a vascular tissue of the body that provides oxygen and nutrients to the pigment epithelium and the outer layers of the retina, maintains intraocular pressure and temperature of the eyeball, takes part in focusing the image on the retina, so choroidal defects are a provoking factor in a number of eye diseases [1]. Diagnosis of the choroid is performed using optical coherence tomography (OCT). Various scoring criteria have been developed to quantify the choroid, including choroidal thickness, choroidal volume, and choroidal vascular index (CVI). Visualization of the choroid with OCT is challenging due to the masking effect of the relatively opaque retinal pigment epithelium, traditionally rendered posterior segment OCT images contain shadows that affect visualization of deep structures, including the choroid [2]. This makes it difficult to qualitatively assess the parameters of the choroid. The objective of the work is to develop a technology for the selection of the choroidal region and its quantitative analysis on OCT images to detect endocrine ophthalmopathy.

This paper proposes a technology for selecting the area of interest and calculation the croidal vascular index, the main stages of which are: 1) shadow compensation, 2) image binarization, 3) binarized image filtering, 4) calculation of the croidal index.

Quantitative analysis of the choroid is difficult due to shadows cast by anterior structures such as retinal vessels. Therefore, shadow compensation is important for successful and reliable quantitative OCT analysis, especially of the choroid and other subretinal structures of the pigment epithelium [2,3]. Shadow compensation was performed using the Girard algorithm. Each B-scan is converted to a raw intensity region, and the intensity of each pixel is multiplied by a unique compensation factor derived from the acquisition characteristics of the OCT signal. This method also improves the image contrast of blood vessels [4]. Directly shadow compensation of the image was determined by the following equation:

$$\frac{s(z)}{2\int_z^\infty s(u)du} = ar_s(z) = HDs(z),$$

where HD is an operator that transforms a given signal s(z) into its compensated form, and a is the attenuation coefficient. Contrast enhancement was performed

before image shadow compensation, which is described by the following formula:

$$HDI^n(z) = \frac{I^n(z)}{2\int_z^\infty I^n(u)du},$$

where I is the pixel intensity, n is the exponent.

Binary processing was used to isolate the area of the vessel lumen. The first step was to perform adaptive histogram equalization at the level of individual B-scans to increase the contrast of choroidal vessels. Next, the image was binarized. Due to the difference in the distribution of brightness between individual images, the use of a global threshold is problematic. In this regard, Niblack's adaptive thresholding method was used for binarization. The image is processed using a window. For each window position, its own threshold is calculated in accordance with the formula:

$$treshold = mean + k \cdot s,$$

where mean is the average value of the brightness of the pixels in the window, s is the standard deviation of the brightness values in the window, and k is the coefficient for separating the object from the background. Accordingly, the binarization algorithm has 2 parameters: the window size and the coefficient k. In the course of research, the best average value for the maximum separability of classes was given by a window of 41 pixels and a parameter of k=0.03.

To get rid of noise in the binarized image, methods of morphological image processing were used [5]. An opening was performed (consecutive application of erosion and dilation operations) of an image with a structural element of the form of an ellipse sized 5x5.

The choroidal vascular index is defined as the ratio of the area of the vessel lumen to the total area of the vascular layer. Calculations were made for OCT images with and without pathology. The index was calculated on a limited area of 1000 nm to the left and right of the fovea. The values were calculated for the entire data set for each class, the average value of the index for the class and the standard deviation were determined. Studies have shown that the average value of the index for the normal class is 0.52, and for pathology 0.56, with an SD of 0.02.

In the course of the work, a technology was proposed for isolating the area of the choroid and determining the choroidal vascular index on images of optical coherence tomography to detect endocrine ophthalmopathy. The choroid is one of the most vascularized structures of the human body and plays an indispensable role in the nutrition of photoreceptors [6]. Thus, the development and research of new methods of the choroid plays an important role in the diagnosis of EOP, which will make it possible to increase the likelihood of detecting signs of EOP in patients at earlier stages.

[1] *Kudasheva G. R.* Structural and functional features of the choroid in normal and pathological conditions of the visual system, optimal research methods // Modern technologies in ophthalmology, Moscow: Ophthalmology, 2020. — p. 333–339.

[2] *Vupparaboina K. K.* Quantitative shadow compensated optical coherence tomography of choroidal vasculature // Scientific Reports, 2018. Vol. 8. No.6461.

[3] *Cheong H.* OCT-GAN: single step shadow and noise removal from optical coherence tomography images of the human optic nerve head // Biomed. Opt. Express, 2021. Vol. 12. — p. 1482–1498.

[4] *Girard M. J.* Shadow removal and contrast enhancement in optical coherence tomography images of the human optic nerve head // Investigative Ophthalmology & Visual Science, 2011. Vol. 52. — p. 7738–7742.

[5] *Shagalova P. A.* Study of image preprocessing algorithms to improve the efficiency of medical images recognition // Proceedings of NSTU im. R.E. Alekseeva, Nizhny Novgorod: Nizhny Novgorod State Technical University. R.E. Alekseeva, 2020. — p. 25–32.

[6] *Agrawal M.* Exploring choroidal angioarchitecture in health and disease using choroidal vascularity index // Progress in Retinal and Eye Research, 2020. Vol. 77. No.. 100829.

Глубокая нейронная сеть для диагностики острого инсульта на основе анализа бесконтрастных КТ-изображений мозга

Бериков Владимир Борисович[1] berikov@math.nsc.ru
Гривкин Андрей Артемович[2]⋆ a.grivkin@g.nsu.ru
[1]Новосибирск, Институт математики им. С.Л. Соболева СО РАН
[2]Новосибирск, Новосибирский государственный университет

В настоящее время актуальной задачей является разработка автоматизированных систем компьютерной диагностики различных заболеваний. Компьютерная томография (КТ) головного мозга, выполненная в первые часы после возникновения острого ишемического или геморрагического инсульта, оказывает значительное влияние на выбор тактики лечения пациента, а по сравнению с другими методами, бесконтрастная КТ не имеет значительных противопоказаний.

Интерпретация КТ изображений сопряжена с определенными трудностями, так как ранние изменения мозга выглядят как участки несколько измененной плотности, которые человеческий глаз в силу различных факторов не всегда может отличить от нормальных тканей. Кроме того, на изображениях часто видны артефакты (вызванные движениями пациента или сканером), которые могут выглядеть как инсультные поражения. Поэтому разработка автоматизированных процедур локализации и оценки объема пораженных тканей на основе КТ изображений, в том числе процедур на основе сверточных нейронных сетей (СНС), является актуальной задачей. Ряд проведенных исследований (см, например, [1, 2, 3]) показал, что методы на основе СНС сопоставимы с радиологами в с точки зрения чувствительности, специфичности и других метрик качества. Это позволяет использовать результаты, полученные на основе СНС в качестве вспомогательного инструмента в клинической практике.

Набор данных, использованный для исследования, содержит бесконтрастные КТ снимки головного мозга 80 пациентов с диагнозом острый ишемический инсульт. Данные получены из Международного томографического центра СО РАН. Все объемные изображения имеют одинаковое разрешение 512×512, но разное количество срезов от 306 до 505 в зависимости от пациента. Для каждого объемного изображения получена соответствующая ручная сегментация, выполненная двумя экспертами-рентгенологами.

Стандартной архитектурой СНС для задачи сегментации трехмерных медицинских изображений является 3D U-Net. В данной работе эта базовая модель была модифицирована для получения лучших результатов предсказания.

Первая модификация заключается в добавлении attention gate блоков. Данный механизм подавляет ненужные области во входном изображении, помогая алгоритму фокусироваться на важных зонах, что повышает чувствительность модели и улучшает ее точность. При получении окончательного предсказания также учитывался выход каждого блока декодера, исходя из предположения,

что это может улучшить чувствительность модели. Для этого карта признаков, полученная из блоков декодера, увеличивается до размера исходного изображения, после чего к нему применяется 3D-свертка 1x1x1 с одним выходным каналом. Эта маска принимает непосредственное участие в итоговом прогнозе с некоторым весом, который увеличивается к концу декодер-части сети. Конкретные значения весов были выбраны экспериментально.

Чтобы получить итоговое предсказание, все вероятности были бинаризованы с пороговым значением 0,5. Метриками качества предложенного алгоритма были выбраны коэффициент Dice, Чувствительность и Специфичность. Результаты, показанные в Таблице 1, представляются как среднее значение $\pm$ стандартное отклонение, полученные на валидационной выборке при перекрестной проверке на 5 частях.

Dice	Чувствительность	Специфичность
0.702 ± 0.161	0.760 ± 0.201	0.996 ± 0.006

Таблица 1. Значения целевых метрик на валидационной выборке при перекрестной проверке

На Рис. 1 можно видеть некоторые результаты работы алгоритма на валидационной выборке. Представлены срезы в аксиальной плоскости исходного изображения мозга; маски, которые были предоставлены экспертами; предсказанные бинаризованные маски.

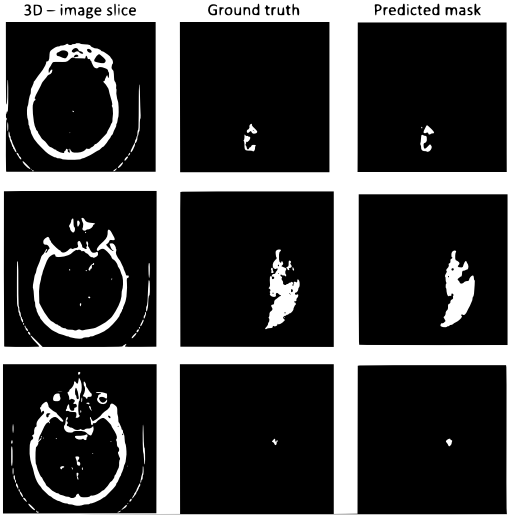


Рис. 1. Пример результатов работы алгоритма

Работа поддержана грантом РФФИ No. 19-29-01175.

[1] Nedel'ko V., Kozinets R., Tulupov A., Berikov V. Comparative Analysis of Deep Neural Network and Texture-Based Classifiers for Recognition of Acute Stroke using Non-Contrast CT Images // 2020 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT). IEEE, 2020. — p. 376-379.

[2] Kalmutskiy K., Tulupov A., Berikov V. Recognition of Tomographic Images in the Diagnosis of Stroke // Lecture Notes in Computer Science, Vol. 12665, 2021. — p. 166-171.

[3] Mikhailapov D., Tulupov A., Alyamkin S., Berikov V. Compression of Deep Neural Network for Acute Ischemic Stroke Segmentation // 2022 Ural-Siberian Conference on Computational Technologies in Cognitive Science, Genomics and Biomedicine (CSGB). IEEE, 2022. — p. 240-245.

Deep neural network for the diagnosis of acute stroke based on the analysis of non-contrast CT brain images

Berikov Vladimir[1] berikov@math.nsc.ru
Grivkin Andrey[2]⋆ a.grivkin@g.nsu.ru
[1]Novosibirsk, Sobolev Institute of Mathematics SB RAS
[2]Novosibirsk, Novosibirsk State University

Currently, an urgent task is the development of automated systems for computer diagnostics of various diseases. Computed tomography (CT) of the brain, performed in the first hours after the onset of an acute ischemic or hemorrhagic stroke, has a significant impact on the choice of patient treatment tactics, and compared to other methods, non-contrast CT has no significant contraindications.

Interpretation of CT images is associated with certain difficulties, since early brain changes look like areas of slightly reduced density, which the human eye, due to various factors, cannot always distinguish from the normal tissues. In addition, images often show artefacts (caused by patient movements or by the scanner) that may look like stroke lesions. Therefore, the development of automated procedures for localization and estimation of affected tissue based on CT images, including procedures based on convolutional neural networks (CNN), is an urgent task. A number of studies (see, for example, [1, 2, 3]) have shown that SNA-based methods are comparable to those of radiologists in terms of sensitivity, specificity, and other quality metrics. This makes it possible to use the results obtained on the basis of the CNN as an auxiliary tool in clinical practice.

The data set used for the study contains non-contrast CT images of the brains of 80 patients diagnosed with acute ischemic stroke. The data were obtained from the International Tomography Center SB RAS. All 3D images have the same resolution of 512×512, but a different number of slices from 306 to 505 depending on the patient. For each volumetric image, a corresponding manual segmentation was obtained, performed by two expert radiologists.

The standard CNN architecture for 3D medical image segmentation is 3D U-Net. In this paper, this basic model has been modified to obtain better prediction results.

The first modification is to add attention gate blocks. This mechanism suppresses unwanted areas in the input image, helping the algorithm to focus on important areas, which increases the sensitivity of the model and improves its accuracy. When obtaining the final prediction, the output of each decoder block was also taken into account, on the assumption that this could improve the sensitivity of the model. To do this, the feature map obtained from the decoder blocks is increased to the size of the original image, after which a 1x1x1 3D convolution with one output channel is applied to it. This mask is directly involved in the final prediction with some weight, which increases towards the end of the decoder part of the network. The specific values of the weights were chosen experimentally.

To get the final prediction, all probabilities were binarized with a threshold value of 0.5. The quality metrics of the proposed algorithm were the Dice coefficient, Sensitivity and Specificity. The results shown in Table 1 are presented as the mean ± standard deviation obtained from the validation set in a five-fold cross-validation.

Dice	Sensitivity	Specificity
0.702 ± 0.161	0.760 ± 0.201	0.996 ± 0.006

Table 1. Values of quality metrics on the validation sample during cross-validation

In Fig. 1 one can see some of the results of the algorithm on the validation set. There are presented slices in the axial plane of the brain image; masks provided by experts; predicted binarized masks.

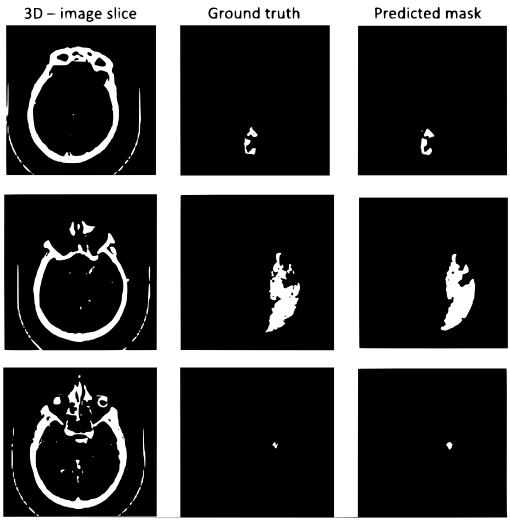


Fig. 1. An example of the results of the algorithm

This research is funded by RFBR, grant 19-29-01175.

[1] Nedel'ko V., Kozinets R., Tulupov A., Berikov V. Comparative Analysis of Deep Neural Network and Texture-Based Classifiers for Recognition of Acute Stroke using Non-Contrast CT Images // 2020 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT). IEEE, 2020. — p. 376-379.

[2] Kalmutskiy K., Tulupov A., Berikov V. Recognition of Tomographic Images in the Diagnosis of Stroke // Lecture Notes in Computer Science, Vol. 12665, 2021. — p. 166-171.

[3] Mikhailapov D., Tulupov A., Alyamkin S., Berikov V. Compression of Deep Neural Network for Acute Ischemic Stroke Segmentation // 2022 Ural-Siberian Conference on

Computational Technologies in Cognitive Science, Genomics and Biomedicine (CSGB). IEEE, 2022. — p. 240-245.

Извлечение признаков формы для классификации экологического состояния по изображениям листьев

Сенин Александр Николаевич[1]⋆ aaasenin@gmail.com
Местецкий Леонид Моисеевич[1] mestlm@mail.ru
Тирас Харлампий Пантелеевич[2] tiras@iteb.ru
[1]Москва, Московский государственный университет имени М.В. Ломоносова
[2]Пущино, Институт теоретической и экспериментальной биофизики РАН

Прямые количественные параметры (биодатчики, биоиндикаторы), отражающие состояние окружающей среды в регионе, ценны в задачах экологии. Такими датчиками–маркерами могут быть листья деревьев, например, липы. Современные технологии быстрого сканирования и обработки большого массива полученных изображений листьев в сочетании с алгоритмами вычисления параметров формы листьев открывают перспективу разработки простых биоиндикаторов, основанных на геометрических параметрах листьев.

Для проведения исследования были собраны листья липы в двух городах Московской области: в Пущино и Серпухове в июле 2021 г. Деревья принадлежат одному биологическому виду ***Tilia cordata***, но растут в разных по экологическому состоянию точках, отобранных по интенсивности локального автомобильного трафика. Будем условно называть эти районы «чистыми» и «грязными». Относительно грязные точки расположены на расстоянии 5–10 м от перекрестка автодорог с наиболее интенсивным движением в Пущино и Серпухове, а относительно чистые точки находятся на расстоянии не менее 500 м от таких мест.

Рассмотрим алгоритм вычисления простейших признаков формы листа — площади и периметра (длины границы). После сегментации листа на изображении необходимо удалить его черенок, иначе оцениваемые параметры будут зашумлены. Для удаления черенка строится скелет листа — связный геометрический граф, образованный центрами всех максимальных вписанных в фигуру окружностей. В каждой точке скелета определена радиальная функция, равная радиусу вписанной окружности с центром в этой точке [1]. В скелете листа выделяются те точки, в которых радиальная функция не превосходит порогового значения, подобранного экспертом. Эти точки образуют несвязный подграф скелетного графа. Выделяется наибольшая по числу вершин связная компонента подграфа, черенку будет отвечать именно эта компонента.

Современные библиотеки предлагают решения, вычисляющие площадь и периметр на основе подсчёта граничных пикселей в растровом изображении. Такой подход приводит к ошибкам в вычислении площади и периметра, вместо этого предлагается вычислять эти параметры на основе построения аппроксимирующего многоугольника. Для этого строится многоугольник минимального периметра, разделяющий чёрные и белые пиксели. Для найденного минималь-

ного разделяющего многоугольника длина границы (периметр) и площадь вычисляются по известным формулам [1].

Оценки площади и периметра листьев получаются с помощью описанных алгоритмов для всех листьев в двух городах. Отобразим листья точками в двумерном признаковом пространстве (см. пример для Пущино на Рис. 1). Точки разбиваются на кластеры в зависимости от экологического состояния. Несмотря на некоторое пересечение кластеров, существует связь между распределениями признаков и чистотой района. Чистому району отвечает кластер с большей дисперсией, средней площадью и большим средним периметром. Во всех случаях сохраняется зависимость: центр масс (точка, координаты которой равны выборочному среднему соответствующих признаков) чистого района лежит правее и выше центра масс грязного района.

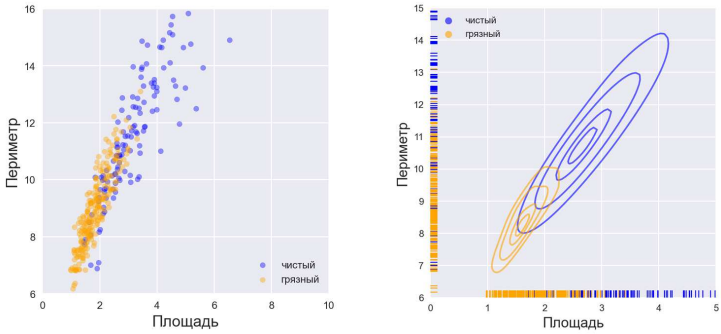


Рис. 1. Визуализация распределений признаков формы в зависимости от чистоты района в Пущино. На диаграмме рассеяния (слева) точка отвечает одному листу. Синие точки соответствуют чистым районам, оранжевые – грязным.

Подберем одномерный показатель, агрегирующий площадь и периметр, который при этом будет разделять выборки из разных по экологическому состоянию районов. Для описания удаленности центра масс от начала координат подойдут простейшие функции, например, среднее арифметическое или норма точки в двумерном признаковом пространстве. Можно показать, что средние нормы в двух выборках для чистого и грязного района после агрегации статистически значимо различимы.

Оценим, насколько «сильны» признаки площади и периметра с точки зрения классического машинного обучения. Поскольку объем данных является достаточно малым, существует риск переобучения, поэтому построим простой линейный классификатор – логистическую регрессию. Будем по площади и периметру одного листа предсказывать экологическое состояние района (класс), где растет дерево этого листа. Для корректного оценивания листья в обучении и тесте должны быть с разных деревьев. Пару признаков площади и периметра можно

считать достаточно сильной, классификатор дает на этой паре 0.938 ROC AUC. Отметим, что одномерная агрегация через среднее и норму сохраняет потенциал к качественной классификации (0.928 и 0.916 соответственно).

Помимо классификации экологического состояния по одному листу, предлагается строить классификатор по выборке листьев, собранных в одном районе. Такой подход позволяет улучшить качество классификации. Особенностью подхода является возможность оценить необходимый размер выборки листьев при заданной вероятности ошибки. Например, если задать вероятность ошибки ниже 0.05, то в Серпухове будет достаточно собрать с дерева 40 листьев, а в Пущино 20 листьев.

Подход с оцениванием состояния окружающей среды по признакам формы листьев выглядит перспективным, его можно применять без привлечения экспертов в области биологии и экологии, он не требует ресурсоемких и сложных расчетов, а самое главное – для его применения достаточно лишь собрать несколько листьев, отсканировать и применить алгоритм.

Работа поддержана грантом РФФИ No. 20-01-00664.

[1] *Местецкий Л. М.* Непрерывная морфология бинарных изображений // Москва: Физматлит, 2009.

[2] *Welch B. L.* The generalization of «Student's» problem when several different population variances are involved // Biometrika, 1947. — No.34 (1–2) С. 28–35.

[3] *Efron B.* Bootstrap methods: Another look at the jackknife // The Annals of Statistics, 1979. — No.7 (1) С. 1–26.

Shape feature extraction for environmental health classification using leaf images

Senin Alexander[1]⋆ aaasenin@gmail.com
Mestetskiy Leonid[1] mestlm@mail.ru
Tiras Kharlampiy[2] tiras@iteb.ru
[1]Moscow, Moscow State University
[2]Pushchino, Institute of Theoretical and Experimental Biophysics RAS

Simple quantitative parameters (biomarkers, bioindicators) that can reflect environmental state (health) in region are priceless for solving ecology problems. Linden leaves are an example of such markers. Modern technologies of fast scanning and processing of a large number of images give us the opportunity to develop simple biosensors based on the geometry of leaves.

Tilia cordata biological species was used. The trees were taken at different ecological points based on traffic intensity. These points are respectively called clean and dirty. Relatively dirty points are located at a distance of 5-10 m from the intersection of highways with the most intense traffic in Pushchino and Serpukhov, while relatively clean points are located at a distance of at least 500 m from such places.

Consider an algorithm for calculating the simplest signs of the leaf shape — the area and perimeter (the length of the border). The leaf stalk should be pruned after image segmentation. For pruning leaf skeleton is extracted [1]. After filtering of the skeleton nodes, the largest connected component is considered to be the leaf stalk.

Modern libraries can estimate area and perimeter via pixel calculation, but this approach leads to a decrease in quality. Instead, a polygon of the minimum perimeter is constructed separating the black and white pixels. After that the length of the border (perimeter) and the area are calculated [**?**].

Area and perimeter are estimated for all leafs in both cities. Leafs are considered as points in 2-dimensional feature space. Points are divided into clusters depending on the ecological state. These 2 features can be aggregated into 1-dimensional feature using mean or norm. It can be shown, that the average norms are statistically significantly different for clean and dirty places.

The strength of calculated features are estimated from the point of view of classical machine learning. Since the amount of available data is small the simple linear classifier (logistic regression) is chosen to predict ecological state of a place using only one leaf. The leaves in the training and test set should be taken from different trees. The trained classifier reaches 0.938 ROC AUC.

The quality of classification can be improved using approach of classification using a whole sample. This approach allows to estimate sample size with fixed error probability.

An approach to the creation of bioindicators using image processing methods is promising, it can be used without experts in biology and ecology, it does not require complex and resource intensive calculations.

This research is funded by RFBR, grant 20-01-00664.

[1] *Mestetskiy L. M.* Continuous morphology of binary images // Moscow: Fizmatlit, 2009.

[2] *Welch B. L.* The generalization of "Student's" problem when several different population variances are involved // Biometrika, 1947. — No.34 (1–2) p. 28–35.

[3] *Efron B.* Bootstrap methods: Another look at the jackknife // The Annals of Statistics, 1979. — No.7 (1) p. 1–26.

Построение трёхмерной модели объекта на основе карты расстояний до опорных плоскостей

Неделько Виктор Михайлович[1,3] nedelko@math.nsc.ru
Некрут Егор Олегович[2,3]⋆ e.nekrut@g.nsu.ru
[1]Новосибирск, Институт математики СО РАН
[2]Новосибирск, Новосибирский государственный университет, НГУ
[3]Новосибирск, ООО Экспасофт

Одной из актуальных задач является построение трёхмерной модели зданий (фрагментов жилой застройки) на основе фотографий, снятых с разных ракурсов. Для решения этой задачи существуют хорошо проработанные технологии и программные продукты, основанные на использовании ключевых точек [1]. Однако качество (точность) полученной модели критически зависит от качества и разрешения исходных снимков. При этом получение снимков высокого разрешения для большой территории — трудоёмкий и дорогостоящий процесс. Это обуславливает потребность в разработке методов, позволяющих строить качественные модели на основе снимков относительно низкого разрешения.

Можно выделить два направления в решении этой задачи с использованием нейросетевого подхода: «улучшение» исходных снимков (методы «сверхразрешения» и устранения помех) и непосредственное восстановление «карты» расстояний от камеры до точек объекта.

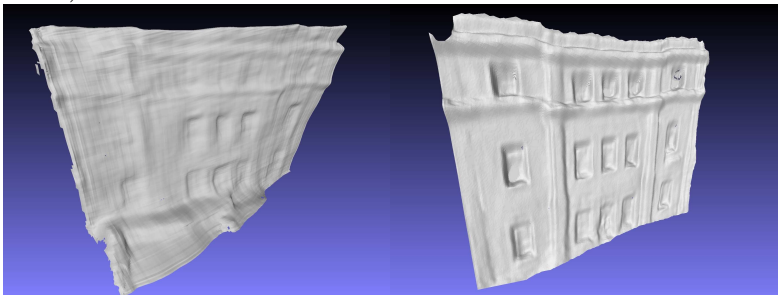
В рамках первого направления можно использовать известные (готовые), в том числе предобученные, архитектуры нейросетей. Однако подобные решения, хоть и повышают визуальное (субъективное) качество изображения, но также вносят искажения, поскольку фактически «дорисовывают картинку». По этой причине качество трёхмерной модели, построенной на основе снимков, обработанных подобным методом, оказывается выше, чем у исходной, но искажения остаются.

Второе направление предполагает использование нейронных сетей для оценки расстояний от камеры до точек сцены, соответствующим пикселям изображения. В последнее время использование полносвёрточных архитектур [2], а также моделей типа transformer [3], обученных на большом объеме данных, значительно увеличило качество предсказания карт глубин по одной фотографии.

Однако при оценке расстояний по изображению неизбежно получается значительная погрешность. Поэтому если строить модель непосредственно по полученной карте расстояний, то возникают значительные искажения (см. рис. 1), в частности, стены зданий могут получаться «изогнутыми», кроме того, мелкие детали могут теряться.

В данной работе предлагается существенное усовершенствование описанного подхода, основанное на использовании априорной информации, а именно того, что фасады (стены) зданий преимущественно плоские.

(a) Исходный снимок (низкого разре-(b) Модель на основе ключевых точек
шения)



(c) Модель на основе карты абсолют-
ных расстояний

(d) Предложенный метод

Рис. 1. Сравнение методов построения модели

Идея предлагаемого метода состоит в том, чтобы оценивать расстояния от точек объекта не до камеры, а до некоторой плоскости, аппроксимирующей часть объекта (например, стену). Иными словами, вместо карты абсолютных расстояний используется карта относительных расстояний.

Схема метода.

— На основе исходных снимков (низкого разрешения) строится грубая модель классическим методом на основе ключевых точек.
— Узлы полученной модели кластеризуются на группы точек, лежащих приблизительно в одной плоскости (метод RANSAC [4]). Назовём плоскости, аппроксимирующие найденные кластеры, опорными плоскостями.
— С помощью нейросети строится (оценивается) карта расстояний до опорных плоскостей.
— По карте расстояний строится фрагмент модели в окрестности данной плоскости.
— Фрагменты объединяются в общую модель.

Для обучения нейросети использовались карты расстояний, полученные на основе модели, построенной по снимкам высокого разрешения. При этом объём обучающей выборки был очень невелик (200 фрагментов фасадов).

При таком объёме выборки целесообразно использовать сеть простой архитектуры. Модели типа UNet [5] продемонстрировали впечатляющие результаты в области сегментации изображений. Была выбрана архитектура с 4-мя pooling-блоками и 1280 каналами в эмбединге. Использование блоков из архитектуры EfficientNet [6] в энкодере сети позволило увеличить точность оценки расстояний.

На рис. 1 приведены: фрагмент исходного снимка низкого разрешения; фрагмент модели, построенной по ключевым точкам; модель, построенная по карте абсолютных расстояний; модель (предлагаемая), построенная по карте относительных расстояний (от опорной плоскости).

Можно сделать вывод, что предложенный подход позволяет существенно улучшить качество модели для объектов, форма которых хорошо аппроксимируется плоскостями. К таким объектам относится подавляющее большинство зданий, что обуславливает достаточно широкую область применимости метода.

[1] *Schönberger J. Lutz, Frahm Jan-Michael* Structure-from-Motion Revisited // Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas: IEEE, 2016. pp. 4104-4113,

[2] *Ranftl René et al.* Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer // arXiv, 2019.

[3] *Ranftl René et al.* Vision Transformers for Dense Prediction // arXiv, 2021.

[4] *Martin A. Fischler, Robert C. Bolles* Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography // Comm. Of the ACM, June (vol. 24), 1981. — С. 381–395.

[5] *Ronneberger Olaf et al.* U-Net: Convolutional Networks for Biomedical Image Segmentation // arXiv, 2015.

[6] *Tan Mingxing, Le Quoc V.* EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks // arXiv, 2019.

Reconstruction of a 3D model based on a map of distances to reference planes

Nedel'ko Victor[1,3] nedelko@math.nsc.ru
Nekrut Egor[2,3]⋆ e.nekrut@g.nsu.ru
[1]Novosibirsk, Institute of mathematics SB RAS
[2]Novosibirsk, Novosibirsk State University, NSU
[3]Novosibirsk, Expasoft LLC

One of the urgent tasks is to build a three-dimensional model of buildings (fragments of residential buildings) based on photographs taken from different angles. To solve this problem, there are well-developed technologies and software products based on the use of key points [1]. However, the quality (accuracy) of the resulting model critically depends on the quality and resolution of the original images. At the same time, obtaining high-resolution images for a large area is a time-consuming and expensive process. This leads to the need to develop methods that allow you to build high-quality models based on relatively low-resolution images.

Two directions can be distinguished in solving this problem using a neural network approach: "improvement" of the original images (methods of "super-resolution" and interference elimination) and direct restoration of "map" distances from the camera to the points of the object.

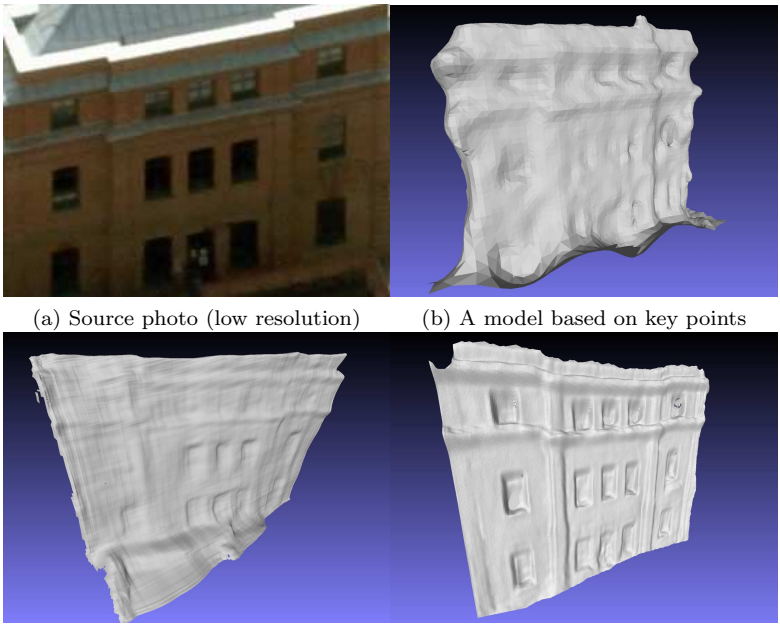
As part of the first direction, you can use well-known (ready-made), including pre-trained, neural network architectures. However, such solutions, although they increase the visual (subjective) image quality, they also introduce distortions, since they actually "finish the picture". For this reason, the quality of a three-dimensional model built on the basis of images processed by a similar method turns out to be higher than that of the original one, but distortions remain.

The second direction involves the use of neural networks to estimate the distances from the camera to the points of the scene corresponding to the pixels of the image. Recently, the use of full-convolutional architectures [2], as well as transformer models [3] trained on a large amount of data, has significantly increased the quality of predicting depth-maps from a single photo.

However, when estimating distances from an image, a significant error is inevitably obtained. Therefore, if you build a model directly from the obtained distance map, then significant distortions occur (see Fig. 1), in particular, the walls of buildings can be "curved", in addition, small details can be lost.

This paper proposes a significant improvement of the described approach based on the use of a priori information, namely that the facades (walls) of buildings are predominantly flat.

The idea of the proposed method is to estimate the distances from the points of the object not to the camera, but to some plane approximating a part of the object (for example, a wall). In other words, a relative distance map is used instead of an absolute distance map.



(a) Source photo (low resolution)

(b) A model based on key points

(c) A model based on an absolute distance map

(d) The proposed method

Fig. 1. Comparison of 3D-models construction methods

The scheme of the method.

— Based on the initial images (low resolution), a rough model is built using the classical method based on key points.
— The nodes of the resulting model are clustered into groups of points lying approximately in the same plane (the RANSAC method [4]). Let's call the planes approximating the found clusters reference planes.
— With the help of a neural network, a map of distances to the reference planes is built (evaluated).
— Based on the distance map, a fragment of the model is built in the vicinity of this plane.
— Fragments are combined into a common model.

Distance maps obtained on the basis of a model based on high-resolution images were used to train the neural network. At the same time, the volume of the training sample was very small (200 fragments of facades).

With such a sample size, it is advisable to use a simple architecture network. Models like UNet [5] have demonstrated impressive results in the field of image segmentation. An architecture with 4 pooling blocks and 1280 channels in embedding

was chosen. The use of blocks from the EfficientNet [6] architecture in the network encoder allowed to increase the accuracy of distance estimation.

Figure 1 shows: a fragment of the original low-resolution image; a fragment of a model constructed from key points; a model constructed from a map of absolute distances; a model (proposed) constructed from a map of relative distances (from the reference plane).

It can be concluded that the proposed approach makes it possible to significantly improve the quality of the model for objects whose shape is well approximated by planes. The vast majority of buildings belong to such objects, which determines a fairly wide area of applicability of the method.

[1] *Schönberger J. Lutz, Frahm Jan-Michael* Structure-from-Motion Revisited // Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas: IEEE, 2016.

[2] *Ranftl René et al.* Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer // arXiv, 2019.

[3] *Ranftl René et al.* Vision Transformers for Dense Prediction // arXiv, 2021.

[4] *Martin A. Fischler, Robert C. Bolles* Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography // Comm. Of the ACM, June (vol. 24), 1981. — p. 381–395.

[5] *Ronneberger Olaf et al.* U-Net: Convolutional Networks for Biomedical Image Segmentation // arXiv, 2015.

[6] *Tan Mingxing, Le Quoc V.* EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks // arXiv, 2019.

Применение вычислительно эффективных альтернатив поворота изображения в задаче поиска вращательной симметрии бинарных фигур

Ломов Никита Александрович [1,2] nikita-lomov@mail.ru
Ляхов Даниил Викторович[1]⋆ liakhov.daniil@mail.ru
Середин Олег Сергеевич[1] oseredin@yandex.ru
[1]Тула,Тульский государственный университет
[2]Москва, ФИЦ ИУ РАН

При решении задач нахождения точного положения фокуса вращательной квази-симметрии, основанной на мере Жаккара для бинарных растровых изображений, предполагается, что проверка найденных эффективных решений может быть проведена с использованием базовой экстенсивной процедура полного перебора по всем углам. При этом строится график (профиль) меры Жаккара в зависимости от угла $J(\varphi)$ [1]. Анализ таких профилей в разных точках изображения позволяет судить об положении фокуса вращательной симметрии и её порядка. Отметим, что также такой профиль может быть использован как самодостаточный дескриптор формы, инвариантный к масштабу и повороту изображения.

В работе показано, что полный перебор по диапазону построения профиля может быть ускорен.

Во-первых, обратим внимание на тот факт, что мера Жаккара вычисляется между исходной фигурой и ее повернутой на угол φ копией, что означает что вторая фигура повернута на угол $-\varphi$ относительно первой. Т.е. мера Жаккара для углов в φ и $2\pi - \varphi$ будет совпадать, что означает, что достаточно рассчитывать меру только для углов $\varphi \in [0, \pi]$.

Далее, пусть $\varphi \in [\frac{\pi}{2}, \pi]$, так как $J(\pi - \varphi) = J(\pi + \varphi)$, вычисление $J(\varphi)$ можно заменить вычислением $J(2\pi - \varphi)$, с учетом того, что матрица поворота $\mathbf{R}(2\pi - \varphi)$ равна $\mathbf{R}(\pi) * \mathbf{R}(\pi - \varphi)$, причем $\pi - \varphi \in [0, \frac{\pi}{2}]$. Таким образом, диапазон углов, необходимых для вычисления поворота, можно сузить до $[0, \frac{\pi}{2}]$.

Наконец, если $\varphi \in [\frac{\pi}{4}, \frac{\pi}{2}]$, вычисление $J(\varphi)$ заменяется вычислением $J(-\varphi) = J(\theta - \frac{\pi}{2})$ с матрицей поворота $\mathbf{R}(\theta - \frac{\pi}{2}) = \mathbf{R}(-\frac{\pi}{2}) * \mathbf{R}(\theta)$, где $\theta \in [0, \frac{\pi}{4}]$. В результате основной поворот в любом случае производится с углом из диапазона $[0, \frac{\pi}{4}]$.

Отдельно отметим, что повороты на углы, кратные $\frac{\pi}{2}$, сводятся к переупорядочению пикселей изображения и требуют значительно меньших трудозатрат по сравнению с поворотом на произвольный угол. Также примем во внимание, что повороты на один и тот же угол вокруг разных точек отличаются лишь смещением результатов друг относительно друга, что сводится к обрезке изображений, поэтому выбор конкретной точки поворота не носит принципиального характера.

Так как лишь в каждом четвёртом случае требуется трудозатратный поворот на новый угол, не кратный $\frac{\pi}{2}$, предложенный приём позволяет сократить время вычисления функции $J(\varphi)$ на отрезке $[0, \pi]$ почти в четыре раза по сравнению с экстенсивной процедурой полного перебора по всем углам.

Обратим внимание на тот факт, что мера Жаккара для изображений, полученных поворотами вокруг одной и той же точки на углы и зависит лишь от величины $|\varphi_2 - \varphi_1|$, так как пара таких фигур получается из исходной и повёрнутой на $|\varphi_2 - \varphi_1|$ дополнительным поворотом на $min(\varphi_1, \varphi_2)$, не изменяющим площади. В силу этого достаточно произвести повороты на углы, кратные $\frac{\pi}{2}$, только для исходного изображения и далее воспользоваться следующей схемой вычисления $J(\varphi)$:

1. Для $\varphi \in [0, \frac{\pi}{4}]$ вычисляется мера между изображением, повёрнутым на φ, и исходным.
2. Для $\varphi \in [\frac{\pi}{4}, \frac{\pi}{2}]$ вычисляется мера между изображением, повёрнутым на $\frac{\pi}{2} - \varphi \in [0, \frac{\pi}{4}]$, и изображением, повёрнутым на $\frac{\pi}{2}$.
3. Для $\varphi \in [\frac{\pi}{2}, \frac{3\pi}{4}]$ вычисляется мера между изображением, повёрнутым на $\varphi - \frac{\pi}{2} \in [0, \frac{\pi}{4}]$, и изображением, повёрнутым на $-\frac{\pi}{2}$.
4. Для $\varphi \in [\frac{3\pi}{4}, \pi]$ вычисляется мера между изображением, повёрнутым на $\pi - \varphi \in [0, \frac{\pi}{4}]$, и изображением, повёрнутым на π.

Во всех случаях абсолютная величина разности для сравниваемых углов принадлежит нужному диапазону.

В докладе приводится окончательный алгоритм с математическими выкладками, а также приводится сравнение подходов к ускорению расчета профиля, один из которых полагается на точное значение угла поворота, а другой – на абсолютную разность между двумя углами.

Проведены экспериментальные исследования по сравнению экстенсивной базовой процедуры полного перебора и предложенных в работе альтернативных версий на изображениях различного масштаба. Реализация выполнена на языке программирования C++ с использованием библиотеки OpenCV. Результаты замеров времени подтвердили предположение о почти четырехкратном ускорении. Для изображений размера порядка 800x600 пикселей полный расчет профиля не превышает трех сотых секунды.

Исследование выполнено при поддержке гранта Российского научного фонда No. 22-21-00575, `https://rscf.ru/project/22-21-00575/`

[1] *Seredin O., Liakhov D., Kushnir O., Lomov N.* Jaccard Index-Based Detection of Order 2 Rotational Quasi-Symmetry Focus for Binary Images. Pattern Recognition and Image Analysis, 2022. — No. 32(3), p. 672–681.

Application of computationally efficient alternatives to image rotation in the problem of searching for rotational symmetry on binary shapes

Lomov Nikita[1,2] nikita-lomov@mail.ru
Liakhov Daniil[1]⋆ liakhov.daniil@mail.ru
Seredin Oleg[1] oseredin@yandex.ru
[1]Tula, Tula State University
[2]Moscow, FRC CSC RAS

When solving the problems of finding the exact location of the focus of rotational quasi-symmetry based on the Jaccard measure for binary figures, it is assumed that the verification of the effective solutions can be carried out using the basic extensive procedure of a complete search at all angles. In this case, a graph (a profile) of the Jaccard measure is plotted depending on the angle $J(\varphi)$ [1]. The analysis of such profiles at different points of the image identifies the position and order of the rotational symmetry focus. Note that such a profile can also be used as a self-sufficient shape descriptor that is invariant to the scale and rotation of the image.

The paper shows that a complete search over the range of profile construction can be accelerated.

First, let's pay attention to the fact that the Jaccard measure is calculated between the original figure and its copy rotated by an angle φ, which means that the second figure is rotated by an angle $-\varphi$ relative to the first. That is, the Jaccard measure for angles φ and $2\pi - \varphi$ is the same which means that it is enough to calculate the measure only for the angles of $\varphi \in [0, \pi]$.

Next, let the angle $\varphi \in [\frac{\pi}{2}, \pi]$, since $J(\pi - \varphi) = J(\pi + \varphi)$, calculation of $J(\varphi)$ can be replaced by calculating $J(2\pi - \varphi)$, taking into account that the rotation matrix $\mathbf{R}(2\pi - \varphi)$ is equal to $\mathbf{R}(\pi) * \mathbf{R}(\pi - \varphi)$ and $\pi - \varphi \in [0, \frac{\pi}{2}]$. Thus, the range of angles required to calculate the rotation can be narrowed to $[0, \frac{\pi}{2}]$.

Finally, let the angle $\varphi \in [\frac{\pi}{4}, \frac{\pi}{2}]$, calculation of $J(\varphi)$ can be replaced by calculating $J(-\varphi) = J(\theta - \frac{\pi}{2})$ with the rotation matrix $\mathbf{R}(\theta - \frac{\pi}{2}) = \mathbf{R}(-\frac{\pi}{2}) * \mathbf{R}(\theta)$, where $\theta \in [0, \frac{\pi}{4}]$. As a result, the main rotation is in any case made with an angle in the range $[0, \frac{\pi}{4}]$.

We note separately, that turns at angles that are multiples of $\frac{\pi}{2}$ are reduced to reordering the pixels of the image and require significantly less labor compared to turning at an arbitrary angle. We also take into account that rotations at the same angle around different points differ only in the displacement of the results relative to each other, which comes down to cropping images, so the choice of a specific rotation point is not of a fundamental nature.

Since only in every fourth case a labor-intensive rotation to a new angle is required, not a multiple of $\frac{\pi}{2}$, the proposed technique reduces the calculation time of the function $J(\varphi)$ on the interval $[0, \pi]$ is almost four times as compared to the extensive procedure of a complete search through all corners.

Let's pay attention to the fact that the Jaccard measure for images obtained by rotations around the same point by angles depends only on the magnitude $|\varphi_2 - \varphi_1|$, since a pair of such figures is obtained from the original and rotated by the angle $|\varphi_2 - \varphi_1|$ additional rotation by the angle $min(\varphi_1, \varphi_2)$, which does not change the area. Because of this, it is enough to make turns at angles that are multiples of $\frac{\pi}{2}$only for the original image and then use the following scheme for calculating the function $J(\varphi)$:

1. For the angle $\varphi \in [0, \frac{\pi}{4}]$ the measure is between the image rotated by φ, and the original image.
2. For the angle $\varphi \in [\frac{\pi}{4}, \frac{\pi}{2}]$ the measure is between the image rotated by $\frac{\pi}{2} - \varphi \in [0, \frac{\pi}{4}]$, and the image rotated by $\frac{\pi}{2}$.
3. For the angle $\varphi \in [\frac{\pi}{2}, \frac{3\pi}{4}]$ the measure is between the image rotated by $\varphi - \frac{\pi}{2} \in [0, \frac{\pi}{4}]$, and the image rotated by $-\frac{\pi}{2}$.
4. For the angle $\varphi \in [\frac{3\pi}{4}, \pi]$ the measure is between the image rotated by $\pi - \varphi \in [0, \frac{\pi}{4}]$, and the image rotated by π.

In all cases, the absolute value of the difference for the compared angles belongs to the desired range.

The presentation provides the final algorithm with mathematical calculations, as well as a comparison of approaches to accelerating the calculation of the profile, one of which relies on the exact value of the angle of rotation, and the other on the absolute difference between the two angles.

Experimental studies have been conducted comparing the extensive basic procedure of full search and the alternative versions proposed in the work on images of different scales. The implementation is made in the C++ programming language using the OpenCV library. The results of time measurements confirmed the assumption of almost fourfold acceleration. For images of the size of 800x600 pixels, the full profile calculation does not exceed 0,03 of a second.

This study was supported by the Russian Science Foundation, Grant No. 22-21-00575, `https://rscf.ru/project/22-21-00575/`

[1] *Seredin O., Liakhov D., Kushnir O., Lomov N.* Jaccard Index-Based Detection of Order 2 Rotational Quasi-Symmetry Focus for Binary Images. Pattern Recognition and Image Analysis, 2022. — No. 32(3), p. 672–681.

NIGHT-HAZE-EXT: расширенный набор данных для оценки алгоритмов удаления тумана с изображений, полученных в темное время суток

Филин Андрей Игоревич[1]⋆ adnrewifilin@gmail.com
Копылов Андрей Валериевич[1] And.Kopylov@gmail.com
Холичева Ангелина Алексеевна[1] ang.hol@yandex.ru
Сурков Егор Эдуардович[1] eg-su@mail.ru
Курбаков Михаил Юрьевич[1] muwsik@mail.ru
Спицын Данила Александрович[1] danila.spitsyn@bk.ru
Грачева Инесса Александровна[1] gia1509@mail.ru
[1]Тула, Тульский государственный университет

Методы удаления тумана активно развиваются, но их сравнительная оценка затруднена недостатком данных. Сложность получения данных для разработки и исследования методов удаления тумана заключается в том, что такие данные представляют из себя пару идентичных изображений, с разницей лишь в том, что туман отсутствует на одном и присутствует на другом. В реальности сложно создать условия, которые позволили бы получить такую пару изображений, но относительно просто её можно получить, если искусственно наложить туман на исходное изображение без тумана. Подавляющее большинство данных для задачи удаления тумана с изображений получено именно таким образом.

Искусственно сгенерированные данные активно используются для решения задачи удаления тумана на изображении с применение методов машинного обучения. Тем не менее, они мало подходят для оценки методов удаления тумана, поскольку, как правило, имеются заметные искажения при наложении тумана, связанные с неточностью полученной карты глубины. Кроме того, физическая модель распространения тумана в полной мере не раскрывает сложность происходящих процессов.

Получить наборы данных, в которых оба изображения (с туманом/без тумана) являлись бы реальными фотографиями, значительно сложнее. Нами было найдено несколько небольших наборов такого рода [1, 2, 3, 4, 5, 6, 7]. Положение усугубляется ещё и тем, что для построения универсальных методов удаления тумана, способных работать в любых условиях без ручной перенастройки параметров, их оценку необходимо проводить как в светлое, так и в темное время суток. Среди указанных, лишь один набор данных включает изображения, которые имитируют темное время суток (получены при низком освещении и с наличием искусственного источника света).

В прошлой работе [8], нами был подготовлен набор NIGHT-HAZE, который делал шаг в сторону решения данной проблемы. Но при анализе нами было обнаружено, что точность устранения тумана по метрикам PSNR и SSIM возрастает вместе с увеличением степени сглаживания карты рассеивания. Дальнейшее исследование привело к такому выводу: из-за малой глубины сцены сложно по-

строить корректную карту рассеивания, т.к. слишком мало количество частиц между камерой и объектами на первом и дальнем планах. Поэтому значения карты рассеивания изменялись незначительно.

В текущей работе мы увеличили глубину сцен, а также число объектов на них. Кроме того, было расширено количество степеней тумана. Так же, как и в предыдущей работе, было подготовлено 2 сцены: на первой отсутствуют точечные источники света, на второй – присутствуют. Количество и сложность объектов примерно одинаковы в обоих сценах. Глубина сцен возросла примерно в 2 раза, объекты расположились на протяжении всей глубины. Как и в оригинальной работе, в кадре разместили устройства для калибровки (мишень SpyderLensCal, таблица для калибровки цветов SpyderCheckr, Datacolor SpyderCube для определения баланса серого, тестовая таблица по ISO 12233).

Для каждой сцены было сделано по 32 кадра (изменялись 4 степени освещенности и 8 степеней насыщенности туманом). Освещенность регулировалась количеством включенных ламп освещения. Минимальная освещенность подбиралась таким образом, чтобы при неизменных настройках фотоаппарата (ISO, диафрагма, выдержка), объекты на сцене оставались видимы.

Туман создавался с помощью генератора тумана Involight FM900. После размещения объектов на сцене и подготовки съемочной аппаратуры, делалось 4 снимка с разной степенью освещенности. Данные снимались со всей имеющейся аппаратуры.

После съемки ground truth изображений, на протяжении 15 минут нагнетался дым и был включен вентилятор для более равномерного его распределения. После этого вентилятор и дымогенератор выключались и делалась пауза в течении 30 секунд, чтобы частицы дыма замедлили перемещение. По прошествии паузы, делалась серия снимков с разной степенью освещенности. На то, чтобы сделать снимки на всех устройствах, уходило примерно 10 секунд.

После того, как снимки были получены, чтобы получить следующую ступень насыщенности туманом, делалась пауза 3 минуты на рассеивание дыма. По прошествии указанного времени, делалась следующая серия из 4 снимков. Всего было проведено 6 аналогичных циклов ожидания рассеивания тумана и получения снимков. В сумме, для одной сцены было получено 8 степеней варьирования туманом (1 без тумана, 7 с присутствием тумана различной интенсивности).

Когда были получены снимки для всех комбинаций освещенности и насыщенности туманом для одной сцены, оставшийся дым рассеивался, и формировалась следующая сцена. Когда новая сцена была сформирована, снимки получали аналогичным образом.

Каждый кадр снимался на следующее оборудование:

• фото были сделаны на камеру Canon 2000d в двух форматах: raw и jpg в разрешении 6000x4000 и глубиной цвета 24 bit;

• карта глубины с использованием Intel RealSence d435i;

• тепловая карта с использованием тепловизора Flir C2.

На полученном наборе, а так же паре других синтетических и реальных наборов, проведены эксперименты с использованием современных методов удаления тумана. Результаты показывают, что на синтезированных наборах данных метрики PSNR и SSIM значительно выше, чем на реальных наборах, причем SSIM для всех методов превышает 0.8, в то время как на реальных наборах значения этой метрики в большинстве случаев лежат в диапазоне 0.6-0.7. Вероятно, это возникает из-за того, что модель атмосферного рассеивания, которая использовалась при генерации тумана на соответствующих изображениях, содержит те же допущения, на которых построены методы удаления тумана. Таким образом, можно сделать вывод, что наборы данных с реальными изображениями показывают более объективные значения метрик.

Работа выполнена при поддержке РФФИ, гранты No. 20-07-00441, 20-07-00055.

[1] *Ancuti C. et al.* I-HAZE: a dehazing benchmark with real hazy and haze-free indoor images //International Conference on Advanced Concepts for Intelligent Vision Systems. – Springer, Cham, 2018. — С. 620-631.

[2] *Ancuti C. O., Ancuti C., Timofte R.* NH-HAZE: An image dehazing benchmark with non-homogeneous hazy and haze-free images //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. – 2020. — С. 444-445.

[3] *Ancuti C. et al.* O-haze: a dehazing benchmark with real hazy and haze-free outdoor images //Proceedings of the IEEE conference on computer vision and pattern recognition workshops. – 2018. — С. 754-762.

[4] *Ancuti C. et al.* Dense-haze: A benchmark for image dehazing with dense-haze and haze-free images //2019 IEEE international conference on image processing (ICIP). – 2019. — С. 1014–1018.

[5] *Bijelic M. et al.* Recovering the Unseen: Benchmarking the Generalization of Enhancement Methods to Real World Data in Heavy Fog //CVPR Workshops. – 2019. — С. 11-21.

[6] *El Khoury J. et al.* A Database with Reference for Image Dehazing Evaluation //Journal of Imaging Science and Technology. – 2018. — С. 10503-1-10503-13.

[7] *El Khoury J. et al.* A spectral hazy image database //Lecture Notes in Computer Science. – 2020. — С. 44-53.

[8] *Filin A. et al.* //Mathematical Methods for Pattern Recognition: Book of abstract of the 20th Russian National Conference with International Participation. – 2021. — С. 245-250.

NIGHT-HAZE-EXT: an extended dehazing benchmark with real hazy and haze-free low-light indoor images

Filin Andrei[1]⋆ adnrewifilin@gmail.com
Kopylov Andrei[1] And.Kopylov@gmail.com
Holicheva Angelina[1] ang.hol@yandex.ru
Surkov Egor[1] eg-su@mail.ru
Kurbakov Mihail[1] muwsik@mail.ru
Spitsyn Danila[1] danila.spitsyn@bk.ru
Gracheva Inessa[1] gia1509@mail.ru
[1]Tula, Tula State University

Haze removal techniques are rapidly evolving, but their comparative evaluation is difficult due to the lack of datasets. The difficulty in obtaining datasets for the development and research of haze removal methods is that the data is a pair of identical images, with the only difference being that haze is absent on one and present on the other. In reality, it is difficult to create conditions that would make it possible to obtain such a pair of images, but it can be obtained relatively easily if haze is artificially superimposed on the original haze-free image. The vast majority of datasets for the haze removal task was obtained in this way.

Artificially generated datasets are actively used in haze removal methods, built with the help of machine learning. However, they are not very suitable as dehazing benchmarks since there are usually noticeable distortions in the generated haze due to the inaccuracy of the original depth map. Moreover, the physical model of haze propagation does not fully reveal the complexity of the ongoing processes.

Obtaining datasets where both ground truth and hazy images are real photographs is much more difficult. We have found a few small datasets of this kind [1, 2, 3, 4, 5, 6, 7]. The situation even more aggravated by the fact that developing of universal haze removal methods (that can work in any conditions without manual reconfiguration), its' evaluation must be carried out both in day and nighttime. Among noticed, just one dataset includes images that simulate a nighttime (images, obtained in low light conditions and with the presence of an artificial light source).

In the previous work [8], we presented the NIGHT-HAZE dataset, which took a step towards solving this problem. But during it's analysis, we found that the haze removal accuracy by PSNR and SSIM metrics increases together with the increase smoothing of the scatter map. Further research led to the following conclusion: due to the small depth of the scene, it is difficult to build a correct scatter map, because the number of particles between the camera and objects in the foreground and background is too small. Therefore, the values of the scatter map changed little.

In this work, we have increased the depth of the scenes, as well as the number of objects on them. In addition, the number of haze levels has been expanded. Just like in the previous work, 2 scenes were prepared: on the first one there are no point light sources, on the second one they are present. The number and complexity of

objects are about the same in both scenes. The depth of the scenes has increased by about 2 times, the objects are located throughout the entire depth. As in the original work, calibration devices were placed in the frame (SpyderLensCal target, SpyderCheckr color calibration table, Datacolor SpyderCube for gray balance, ISO 12233 test chart).

For each scene, 32 frames were captured (with variation of 4 degrees of illumination and 8 degrees of haze intensity). Illumination was regulated by the number of included lighting lamps. The minimum illumination was tuned in such a way that the objects on the scene remained visible without changing the same camera settings (ISO, aperture, shutter speed).

The haze was generated using haze machine Involight FM900. After placing the objects on the stage and preparing the shooting equipment, 4 shots were taken with varying illumination degrees. Shots was taken from all available equipment one by one (it took about 10 seconds in sum for one cycle).

After shooting ground truth images, haze was generating in for 15 minutes and the fan was turned on to distribute it more evenly. After that, the fan and hazer were turned off and waiting for 30 seconds was made for the haze particles slowed down. After the delay, a series of shots were taken with varying illumination levels.

After the images were taken, to get the next level of haze intensity, a 3 minutes waiting was performed for the haze dissipating. After that, the next series of 4 shots was taken. Further, 6 more similar cycles were carried out. In total, 8 degrees of haze variation were obtained for one scene (1 haze-free (ground truth) a 7 hazy images with varying intensity).

When images were taken for all combinations of illumination and haze intensity for one scene, the remaining haze dissipated and the next scene was formed. When a new scene was formed, pictures were taken in a similar manner.

Each frame was taken with the following equipment:

- the photos were taken on a Canon 2000d camera in two formats: raw and jpg in 6000x4000 resolution and 24 bit color depth;
- depth map using intel RealSense d435i;
- heat map using thermal imager Flir C2.

Experimental research of the several state-of-the-art single image dehazing methods was performed on the presented, and also the several other synthetic and real datasets. The results show that PSNR and SSIM metrics on the synthesized datasets are significantly higher than on real datasets. Moreover, SSIM for all methods exceeds 0.8 on the synthetic datasets, while on the real datasets values of this metric in most cases lie in the range of 0.6-0.7. This is likely due to the fact that the atmospheric scattering model that was used to generate haze on the corresponding images, is based on the same assumptions, as the used haze removal methods. Thus, we can conclude that datasets, consisting with real images, show more objective metric values.

This research is founded by RFBR, grants No. 20-07-00441, 20-07-00055.

[1] *Ancuti C. et al.* I-HAZE: a dehazing benchmark with real hazy and haze-free indoor images //International Conference on Advanced Concepts for Intelligent Vision Systems. – Springer, Cham, 2018. — p. 620-631.

[2] *Ancuti C. O., Ancuti C., Timofte R.* NH-HAZE: An image dehazing benchmark with non-homogeneous hazy and haze-free images //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. – 2020. — p. 444-445.

[3] *Ancuti C. et al.* O-haze: a dehazing benchmark with real hazy and haze-free outdoor images //Proceedings of the IEEE conference on computer vision and pattern recognition workshops. – 2018. — p. 754-762.

[4] *Ancuti C. et al.* Dense-haze: A benchmark for image dehazing with dense-haze and haze-free images //2019 IEEE international conference on image processing (ICIP). – 2019. — p. 1014–1018.

[5] *Bijelic M. et al.* Recovering the Unseen: Benchmarking the Generalization of Enhancement Methods to Real World Data in Heavy Fog //CVPR Workshops. – 2019. — p. 11-21.

[6] *El Khoury J. et al.* A Database with Reference for Image Dehazing Evaluation //Journal of Imaging Science and Technology. – 2018. — p. 10503-1-10503-13.

[7] *El Khoury J. et al.* A spectral hazy image database //Lecture Notes in Computer Science. – 2020. — p. 44-53.

[8] *Filin A. et al.* //Mathematical Methods for Pattern Recognition: Book of abstract of the 20th Russian National Conference with International Participation. – 2021. — p. 245-250.

Метод шаблонного встраивания цифровых водяных знаков в изображения с использованием комплекса нейронных сетей

Евсютин Олег Олегович[1] oevsyutin@hse.ru
Джанашиа Кристина Малхазовна[1]⋆ kdzhanashia@hse.ru
[1]Москва, НИУ «Высшая школа экономики»

Цифровой водяной знак (ЦВЗ) – это информация о контейнере или его владельце, размещенная в цифровом объекте для охраны авторских прав или проверки подлинности. Выделяют три основных показателя, характеризующих эффективность методов встраивания ЦВЗ: вместимость, незаметность и робастность. Вместимость показывает максимальный объем ЦВЗ, который можно разместить в объекте. Незаметность определяет сложность обнаружения ЦВЗ, встроенного в объект, для внешнего наблюдателя. Робастность показывает, какие атаки (искажающие воздействия) объект может претерпеть и все еще быть пригодным для извлечения ЦВЗ. Достижение высоких значений всех трех показателей проблематично, так как повышение одного из показателей ведет к уменьшению двух других.

Шаблонное встраивание является одним из методов встраивания ЦВЗ в изображения, обеспечивающим высокую робастность [1]. Суть метода шаблонного встраивания состоит в том, что ЦВЗ, представленный изначально в виде двоичной последовательности, преобразуется в изображение-шаблон, состоящее из двоичных матриц-шаблонов, соответствующих нулевому и единичному биту, и это изображение-шаблон совмещается с контейнером по формуле:

$$\begin{cases} I'_R = I_R, \\ I'_G = I_G, \\ I'_B = I_B + \alpha \cdot Template - watermark, \end{cases} \tag{1}$$

где I_R, I_G, I_B – это исходные RGB каналы изображения, I'_R, I'_G, I'_B – это RGB каналы после встраивания ЦВЗ, α – это коэффициент силы встраивания, а $Template - watermark$ – это изображение-шаблон. Для извлечения встроенного ЦВЗ необходимо разбить контейнер на блоки, размер которых соответствует размеру шаблона, и определить значение встроенного бита для каждого блока. Для этого возможно использовать нейронные сети.

Основной проблемой шаблонного встраивания является выбор шаблонов, позволяющих достичь наилучшего баланса незаметности и робастности. В данной работе мы предлагаем использовать нейронные сети не только для извлечения ЦВЗ, но и для генерации шаблонов на этапе встраивания.

Предлагаемая модель генерирует шаблоны, использует их для встраивания ЦВЗ и учится правильно классифицировать блоки контейнера для извлечения встроенного ЦВЗ. Для этого создаются две нейронные сети: генератор, цель которого состоит в генерации наиболее незаметных шаблонов, и классификатор,

цель которого – правильно классифицировать блоки контейнера, содержащие нулевые и единичные биты ЦВЗ, встроенные с использованием сгенерированных шаблонов. Сети обучаются вместе. Для повышения робастности шаблонов в процесс обучения перед классифицирующей сетью добавляется блок имитации случайных атак (см. рис. 1).

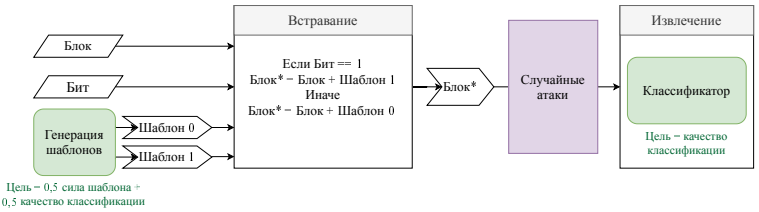


Рис. 1. Схема работы комплекса нейронных сетей для организации шаблонного встраивания ЦВЗ

Сеть генератора ничего не получает на вход, а на выходе выдает два шаблона, соответствующих нулю и единице. Обучаемыми параметрами данной сети являются две матрицы сдвигов одинакового размера, которые затем и используются как шаблоны. Целью обучения является минимизация разницы между блоками изображений без шаблонов и с ними $scoreGen$ и максимизация качества классификации $scoreClas$. Полная цель генератора: $totalScore = 0,5 \cdot \cdot scoreGen + 0,5 \cdot (1 - scoreClas)$. После обучения сеть генерации шаблонов не нужна – достаточно извлечь из нее шаблоны. Сеть классификации принимает на вход блок изображения–контейнера и выдает на выходе извлеченный бит, ее целью является максимизация качества классификации. Для обучения модели используется метод оптимизации Adam со скоростью обучения 0,002 на 30 эпохах и 60 000 тренировочными блоками.

Сравнение робастности шаблонов, полученных с помощью данной нейросетевой модели (далее – НС-шаблоны), и некоторых созданных вручную шаблонов [2] при одинаковом уровне незаметности приведено в таблице 1. Нумерация шаблонов взята из [2]. В таблице робастность оценивается с помощью метрики BER:

$$BER(W, W') = 1 - \frac{\sum_{i=1}^{n} W(i) \equiv W'(i)}{n}, \tag{2}$$

где W – это встраиваемый ЦВЗ, W' – извлеченный ЦВЗ, а n – длина ЦВЗ.

Шаблоны, полученные с использованием предложенной модели, однозначно превосходят созданные вручную шаблоны VI, превосходят шаблоны I и III по большинству атак и сравнимы с шаблонами II и V. НС-шаблоны обладают наилучшей устойчивостью к JPEG2000-сжатию и показывают второй результат по устойчивости к JPEG-сжатию. Визуально полученные НС-шаблоны напоми-

Таблица 1. Сравнение робастности полученных НС-шаблонов с предложенными ранее на одном уровне незаметности

Атака	I	II	III	IV	V	VI	НС-шаблоны
Нет	0,78	0	0	0	0,55	1,09	0
Шум Гаусса ($\sigma^2 = 0,01$)	10,39	2,97	0,86	4,77	15,71	15,47	0,7
Изменение резкости	0,86	0,08	0	0	0,79	0,86	0,08
Фильтр Гаусса ($\sigma = 1$)	0,86	0,08	30,85	0	3,75	2,19	0,23
Фильтр Винера	2,34	0,08	46,71	4,61	11,88	3,21	1,25
Поворот на 2^{circ} с обрезанием и возвращением	0,7	0	0	0	0,94	1,41	0
Увеличение в 4 раза с возвращением	0,7	0	0	0	0,63	0,86	0
Уменьшение в 2 раза с возвращением	0,62	0,79	33,67	0,24	5,4	2,66	1,25
JPEG ($QF = 90$)	6,32	22,97	39,53	27,35	31,57	15,55	14,84
JPEG ($QF = 100$)	0,46	2,19	19,14	0,63	7,1	3,56	2,26
JPEG2000 ($CR = 12$)	3,9	3,05	16,56	5,4	8,13	4,22	1,48

нают шум, что также является определенным преимуществом, поскольку это затрудняет статистический анализ изображений-контейнеров.

Таким образом, полученные шаблоны обладают достаточно высокими показателями эффективности и не уступают шаблонам, созданным вручную. Для достижения более высоких показателей возможно рассмотреть различные модификации сети генерации. Например, для повышения устойчивости к уменьшению можно попробовать генерировать шаблон меньше размера блоков и масштабировать его перед встраиванием. Также в будущих исследованиях планируется усложнить цель оптимизации генерирующей сети так, чтобы в ней были учтены свойства визуальной системы человека.

Работа поддержана грантом РНФ No. 21-71-10113.

[1] *Fang H. Chen D. Huang Q. Zhang J. Ma Z. Zhang W. Yu N.* Deep template-based watermarking // IEEE Transactions on Circuits and Systems for Video Technology, IEEE 2020. — С. 1436–1451.

[2] *Джанашиа К. М. Евсютин О. О.* Low complexity template-based watermarking with neural networks and various embedding templates // Computers and Electrical Engineering, Elsevier, 2022. — С. 108194.

Template-based image watermarking method with a complex of neural networks

Evsutin Oleg[1] oevsyutin@hse.ru
Dzhanashia Kristina[1]⋆ kdzhanashia@hse.ru
[1]Moscow, HSE University

A digital watermark is some information about the container or its owner placed in the object for copyright protection or authentication. There are three main indicators that define a watermarking method: capacity, imperceptibility, and robustness. Capacity shows the watermark of which size can be placed in the object. Imperceptibility demonstrates how hard it is to detect the watermark in the container. Robustness shows which attacks (modifications) a watermarked object can withstand without losing the ability to detect or extract the watermark. Achieving high results in all three indicators is problematic, as the enhancement of one of them usually leads to the degradation of the other two.

Template-based watermarking is one of the watermarking methods with good robustness [**?**]. Template-based watermarking's main idea is to transform the binary watermark into a template-watermark that consists of two matrix-templates corresponding to one and zero bits. Then, the template-watermark is combined with the container in accordance with:

$$\begin{cases} I'_R = I_R, \\ I'_G = I_G, \\ I'_B = I_B + \alpha \cdot Template - watermark, \end{cases} \tag{1}$$

I_R, I_G, I_B are RGB channels of container image, I'_R, I'_G, I'_B are RGB channels of watermarked image, α is the embedding strength coefficient. In template-based methods, extraction is performed through division of watermarked image into blocks of template's size and determination of which bit was embedded into them. This can be done with neural networks.

When using template-based watermarking, the question arises of how to choose templates to achieve the best imperceptibility-robustness balance. In this work, it is proposed to use neural networks not only during watermark extraction but also to generate matrix-templates.

The proposed model generates templates, uses them to embed the watermark, and learns to correctly classify container blocks to extract the embedded watermark. To do this, two neural networks are created: a generator, the purpose of which is to generate the most imperceptible templates, and a classifier, the purpose of which is to correctly classify container blocks containing zero and one bits of the watermark embedded using the generated templates. The networks are trained together. To increase the robustness of the templates, a block for simulating random attacks is added to the learning process in front of the classifying network (see fig. 1).

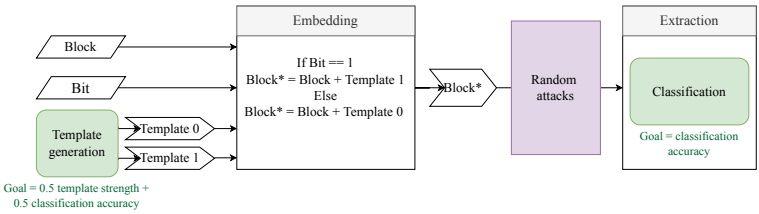


Fig. 1. The proposed complex of neural networks for watermarking

The generator network receives nothing as input and outputs two templates corresponding to zero and one bits. The learning parameters of this network are two bias matrices of the same size which are then used as templates. The goal of learning is to minimize the difference between image blocks without and with templates *scoreGen* and to maximize the quality of the *scoreClass* classification. The full goal of the generator is $totalScore = 0.5 \cdot scoreGen + 0.5 \cdot (1 - scoreClas)$. After training, the template generation network is not needed – it is enough to extract templates from it. The classification network takes as input a container image block and outputs the extracted bit, its goal is to maximize the quality of the classification. The model is trained using the Adam optimization method with a learning rate of 0.002 over 30 epochs and 60,000 training blocks.

The comparison of the robustness of templates obtained using this neural network model (hereinafter referred to as NN-templates) and some manually created [2] templates with the same level of imperceptibility is given in the table 1. Template's names are taken from [2]. In the table, robustness is evaluated using the BER metric:

$$BER(W, W') = 1 - \frac{\sum_{i=1}^{n} W(i) \equiv W'(i)}{n}, \tag{2}$$

where W is embedded watermark, W' is extracted watermark and n is length of the watermark.

The templates obtained using the proposed model clearly outperform manually created templates VI, outperform templates I and III in most attacks and are comparable to templates II and V. NN-templates have the best resistance to JPEG2000 compression and show the second best result in stability to JPEG compression. Visually obtained NN-templates resemble noise, which is also a definite advantage since it makes it difficult to statistically analyze container images.

Thus, the resulting templates have sufficiently high performance and are not inferior to manually created templates. To achieve higher performance, it is possible to consider various modifications to the generation network. For example, to increase robustness to shrinkage, it ts possible to try to generate a template smaller than the block size and scale it before embedding. Also, in future studies, it is planned

Table 1. Comparison of novel NN-templates with previous ones with one imperceptibility level

Attack	I	II	III	IV	V	VI	NN-templates
No	0.78	0	0	0	0.55	1.09	0
Gaussian noise ($\sigma^2 = 0.01$)	10.39	2.97	0.86	4.77	15.71	15.47	0.7
Sharpening	0.86	0.08	0	0	0.79	0.86	0.08
Gaussian filter ($\sigma = 1$)	0.86	0.08	30.85	0	3.75	2.19	0.23
Wiener filter	2.34	0.08	46.71	4.61	11.88	3.21	1.25
Rotation 2^{circ} with cropping and restoration	0.7	0	0	0	0.94	1.41	0
Scaling x4 with restoration	0.7	0	0	0	0.63	0.86	0
Scaling x0.5 with restoration	0.62	0.79	33.67	0.24	5.4	2.66	1.25
JPEG ($QF = 90$)	6.32	22.97	39.53	27.35	31.57	15.55	14.84
JPEG ($QF = 100$)	0.46	2.19	19.14	0.63	7.1	3.56	2.26
JPEG2000 ($CR = 12$)	3.9	3.05	16.56	5.4	8.13	4.22	1.48

to complicate the optimization goal of the generator network so that it takes into account the properties of the human visual system.

This research is funded by RNF, grant 21-71-10113.

[1] *Fang H. Chen D. Huang Q. Zhang J. Ma Z. Zhang W. Yu N.* Deep template-based watermarking // IEEE Transactions on Circuits and Systems for Video Technology, City: Publisher, 2020. — p. 1436–1451.

[2] *Dzhanashia K. Evsutin O.* Low complexity template-based watermarking with neural networks and various embedding templates // Computers and Electrical Engineering, Elsevier, 2022. — p. 108194.

Алгоритм подсчета нитей холстов картин на основе комбинирования Фурье-образов изображений, полученных в направленном свете

Мурашов Дмитрий Михайлович[1]⋆ d_murashov@mail.ru
Березин Алексей Владимирович[2] berezin_aleks@mail.ru
Иванова Екатерина Юрьевна[3] ivanova-e-u@yandex.ru
[1]Москва, ФИЦ ИУ РАН
[2]Москва, ГИМ
[3]Москва, РАЖВиЗ Ильи Глазунова

Предлагается новый алгоритм определения характеристик холстов произведений живописи по изображениям. Новизна алгоритма заключается в использовании пар изображений, зафиксированных в направленном свете, и комбинировании образов изображений в частотной области.

Одним из способов датировки произведений живописи является определение производителя и датировка производства материала основы, в частности, холста. К основным характеристикам холстов, которые используются в атрибуции картин, относятся плотность ткани по направлениям основы и утка, диаметры нитей основы и утка, а также зернистость холста. Для вычисления этих характеристик необходимо подсчитать количество нитей по двум направлениям и количество переплетений на единицу площади. Обычно для определения плотности ткани эксперты выполняют подсчет количества нитей вручную. Эта операция достаточно трудоемка и сопровождается ошибками, особенно если нити достаточно тонкие. В работе [1] предложена математическая модель переплетений нитей и разработан полуавтоматический алгоритм определения плотности нитей на рентгеновских снимках холста на основе анализа пиков Фурье-образа. Алгоритм показал точность в пределах одной нити на сантиметр на 95% изображений. Если в качестве грунта при создании картины использовались свинцовые белила, то рентгеновские снимки не дают требуемой информации для анализа. В этом случае необходимо использовать изображения, полученные в других спектральных диапазонах. Известен ряд алгоритмов подсчета нитей для контроля качества текстильного производства [2], [3], которые основаны на фильтрации изображений ткани в частотной области и пороговой бинаризации. Ошибка определения плотности ткани не превосходила десяти процентов. Однако применение этих алгоритмов, предназначенных для анализа изображений, полученных в проходящем свете, невозможно для определения характеристик тканых основ картин из-за наличия непрозрачного красочного слоя.

В работе [4] был предложен алгоритм подсчета нитей на изображениях холстов, полученных в направленном освещении. Направленное освещение подчеркивало границы нитей, что позволяло определить параметры холста с ошибкой, не превышающей одну нить на сантиметр. Такая величина ошибки соответствует известным алгоритмам [1], [2], [3]. Однако при анализе изображений ос-

нов картин после реставрации (например, “Портрет княжны М.А. Волконской” Ф.С. Рокотова) точность алгоритма снизилась из-за существенного искажения фвктуры холста после обработки реставрационными материалами. Так как в коллекциях музеев растет количество отреставрированных картин, возникает необходимость разработки алгоритмов, способных обеспечить приемлемую точность определения характеристик холста таких картин. В работе [4] плотность холста вычислялась по нескольким фрагментам одного и того же изображения. В представляемой работе для более сильного подчеркивания нитей холста в определенном направлении и повышения точности подсчета нитей, предлагается использовать комбинацию двух изображений холстов, зафиксированных при освещении с противоположных направлений. Предлагается комбинировать изображения в частотной области и определять количество нитей по пикам полученного комбинированного Фурье-спектра.

Алгоритм подсчета нитей. Предлагаемый алгоритм включает процедуры предобработки изображений холста, операции дискретного преобразования Фурье обработанных изображений, комбинирования полученных образов, операции постобработки и процедуру анализа пиков комбинированного спектра.

Процедура предобработки включает операции цветовой редукции, масштабирования, гауссовой фильтрации, эквализации гистограммы, компенсации неравномерной освещенности.Комбинирование Фурье-образов выполняется по правилу:

$$F_c\left(u,v\right) = G\left\{\mathfrak{F}\left[I_1\left(x,y\right)\right], \mathfrak{F}\left[I_2\left(x,y\right)\right]\right\},$$

где $F_c\left(u,v\right)$ - комбинированный спектр; $G\left\{\right\}$ - правило; $I_1\left(x,y\right)$ и $I_2\left(x,y\right)$ - изображения холста после предварительной обработки; x, y - пространственные координаты; $\mathfrak{F}\left[\right]$ - оператор дискретного преобразования Фурье; u, v - координаты в частотной области. Для подсчета количества нитей на изображении выполняются следующие операции. (1) Дискретное преобразование Радона, применяемое к изображению комбинированного спектра:

$$R\left(v\right) = \sum_{u=0}^{H-1} F_c\left(u,v\right),$$

где H - вертикальный размер изображения комбинированного спектра. (2) Определение пространственных частот v_1 и v_2, соответствующих пикам функции $R\left(v\right)$ вне интервала $\pm\delta$ от начала координат. (3) Поиск пространственных частот u_1 и u_2, соответствующих максимумам комбинированного спектра $F_c\left(u,v\right)$ на частотах v_1 и v_2. (4) Вычисление количества нитей на комбинируемых изображениях по значениям u_1 и u_2, v_1 и v_2.

Вычислительный эксперимент. С целью выбора параметров алгоритма и правила комбинирования, а также подтверждения эффективности предложенного алгоритма проведен вычислительный эксперимент. В качестве входных данных предложенного алгоритма использовались пары фотографий холстов,

снятых при направленном освещении с противоположных направлений. Съемка производилась с расстояния приблизительно $21 - 25$ см при положении осветительного прибора, обеспечивающим падение света в диапазоне углов $10 - 30$ градусов относительно плоскости картины. Свет направлялся последовательно снизу, сверху, слева и справа. Изображения холстов картин фиксировались цифровой фотокамерой NIKON D7100, установленной на штативе. Для коррекции искажений и выполнения измерений на холст накладывалась калибровочная сетка с шагом $2,5$ мм. На изображениях выделялись фрагменты размером приблизительно 2500×1900 пикселей, что соответствует области холста $7,5 \times 6$ см. При комбинировании Фурье-образов изображений в эксперименте применялось правило максимума и правило среднего значения. Подсчет количества нитей выполнялся как по одиночному изображению (без комбинирования) так и по паре изображений. Получены следующие результаты. Средняя ошибка подсчета количества нитей по одному изображению составила 0,3 нити на сантиметр. Средняя ошибка подсчета количества нитей по паре изображений указанного выше размера при использовании правила комбинирования по среднему значению составила $0,14$ нити на сантиметр. При этом ошибка определения толщины нити не превышала $0,0136$ мм при измерении по одиночному изображению и $0,007$ мм при измерении по паре изображений. В случае комбинирования Фурье-спектров по правилу максимума ошибки измерений уменьшились примерно в полтора раза. Контроль правильности подсчета количества нитей выполнялся вручную.

Выводы. Таким образом, предложенный алгоритм на основе комбинирования изображений позволяет определять характеристики холстов картин с искаженной фактурой. Точность измерений превосходит точность применявшихся ранее алгоритмов. Предложенный алгоритм позволяет упростить и ускорить процесс подготовки изображений для анализа, поскольку для измерений в одном направлении нитей (основы или утка) исследуется только один фрагмент вместо трех или четырех в алгоритме, описанном в работе [4].

[1] *Johnson D. H., Johnson C., Klein A. G., Sethares W. A., Lee H., Hendriks E.* A thread counting algorithm for art forensics // 2009 IEEE 13th Digital Signal Processing Workshop and 5th IEEE Signal Processing Education Workshop, 2009. P. 679–684.

[2] *Pan R., Gao W., Li Z., Gou J., Zhang J., Zhu D.* Measuring thread densities of woven fabric using the fourier transform // Fibres & Textiles in Eastern Europe, 2015.

[3] *Aldemir E., Özdemir H., Sarı Z.* An improved gray line profile method to inspect the warp–weft density of fabrics // The Journal of The Textile Institute, 110(1), 2019. P. 1–12.

[4] *Murashov D.M., Berezin A.V., Ivanova E.Yu.* Algorithms Based on Maximization of the Mutual Information for Measuring Parameters of Canvas Texture from Images // Lecture Notes in Computer Science, 12665. 2021. P. 77–89.

An algorithm for counting the threads of painting canvases based on combining the Fourier transforms of images obtained in raking light

Murashov Dmitry[1]⋆ d_murashov@mail.ru
Berezin Aleksey[2] berezin_aleks@mail.ru
Ivanova Ekaterina[3] ivanova-e-u@yandex.ru
[1]Moscow, FRC CSC RAS
[2]Moscow, State Historical Museum
[3]Moscow, Glazunov Academy

In this work, a new algorithm for calculating the characteristics of canvases of paintings from images is proposed. The novelty of the algorithm lies in using of pairs of images captured in raking light and combining image patterns in the frequency domain.

One of the methods for dating fine-art paintings is to identify the manufacturer and date the production of the canvas. The main characteristics of canvases that are used in the attribution of paintings are the density of the fabric in the directions of the warp and weft, the diameters of the warp and weft threads, as well as the graininess of the canvas. To compute the density of the fabric, it is necessary to count the number of threads in the sample. Traditionally, threads counting was carried out manually. This operation is quite laborious and is accompanied by errors, especially if the threads are thin enough. In work [1], a mathematical model of weave patterns was proposed, and a semi-automatic algorithm was developed for measuring the canvas density of paintings from X-ray images. The algorithm is based on filtering in the Fourier domain and analyzing the Fourier spectrum peaks. In [1] the authors reported that the best of several spectrum-based algorithms achieved an accuracy within one thread/cm in 95% of canvas density measurements. If lead white paint has been used for creating the painting, then X-rays may not provide the required information for analysis. In this case, it is necessary to use im-ages taken in other spectral ranges. To control the density of the fabric in textile production, a number of algorithms based on image filtering in the frequency domain and thresholding have been developed [2], [3]. The error of threads counting in fabric samples of plain structure did not exceed 10%. Despite the fact that the considered methods developed for analyzing images of fabrics obtained in a transmitted light provide acceptable results, they cannot be applied to analyze painting canvases, because the paint layer of paintings is opaque.

In work [4], an algorithm for counting threads on canvas images obtained in raking light was proposed. Raking light emphasized the boundaries of the threads, which made it possible to count canvas threads with an error not exceeding one thread per centimeter. This error value corresponds to known algorithms [1], [2], [3]. However, when analyzing canvas images of paintings after restoration (for example, “Portrait of Princess M.A. Volkonskaya” by F.S. Rokotov), the accuracy of

the algorithm decreased due to a significant distortion of the texture of the canvas after processing with restoration materials. As the number of restored paintings grows in museum collections, it becomes necessary to develop algorithms that can provide acceptable accuracy in computing the characteristics of the canvas of such paintings. In work [4], the density of the canvas was computed from several samples of the same image. In the present work, in order to emphasize more strongly the threads of the canvas in a certain direction and improve the accuracy of counting the threads, we propose to use a combination of two images of canvases fixed under illumination from opposite directions. We propose to combine images in the frequency domain and evaluate the number of threads from the peaks of the combined Fourier spectrum.

Thread counting algorithm. The proposed algorithm includes procedures for preprocessing canvas images, operations for discrete Fourier transform of processed images, combining the obtained patterns, post-processing operations, and a procedure for analyzing combined spectrum peaks.

The preprocessing procedure includes operations of color reduction, scaling, Gaussian filtering, histogram equalization, uneven illumination compensation. The combination of Fourier images is performed according to the rule:

$$F_c(u,v) = G\left\{\mathfrak{F}\left[I_1(x,y)\right], \mathfrak{F}\left[I_2(x,y)\right]\right\},$$

where $F_c(u,v)$ is the combined spectrum; $G\{\}$ is the rule; $I_1(x,y)$ and $I_2(x,y)$ are pre-processed canvas images; x, y are spatial coordinates; $\mathfrak{F}[]$ is the discrete Fourier transform operator; u, v are coordinates in the frequency domain.

To count the number of threads in an image, the following operations are performed. (1) Discrete Radon transform applied to the combined spectrum image:

$$R(v) = \sum_{u=0}^{H-1} F_c(u,v),$$

where H is the vertical size of the combined spectrum image. (2) Evaluating the spatial frequencies v_1 and v_2 corresponding to the peaks of the function $R(v)$ outside the interval $\pm\delta$ from the origin. (3) Evaluating the spatial frequencies u_1 and u_2 corresponding to the maxima of the combined spectrum $F_c(u,v)$ at the frequencies v_1 and v_2. (4) Computing the number of threads in the canvas images from the values of u_1 and u_2, v_1 and v_2.

Computational experiment. In order to select the parameters of the algorithm and the combination rule, as well as to confirm the effectiveness of the proposed algorithm, a computational experiment was carried out. As an input to the proposed algorithm, we used pairs of photographs of canvases taken under directional illumination from opposite directions. The shooting was carried out from a distance of approximately $21 - 25$ cm with the position of the lighting device illuminating the surface of the painting canvas at an angle of $10 - 30$ degrees. The light was

directed sequentially from below, above, left and right. Images of painting canvas were recorded with a NIKON D7100 digital camera mounted on a tripod. To correct distortions and perform measurements, a calibration grid with a step of 2.5 mm was superimposed on the canvas. Fragments of approximately 2500×1900 pixels in size were selected on the images, which corresponds to a canvas area of 7.5×6 cm. When combining the Fourier patterns of images in the experiment, the maximum rule and the mean value rule were applied. Thread counting was performed both in single images (without combining) and in pairs of images. The following results are obtained. The average thread count error for a single image is approximately 0.3 threads per centimeter. The average error in counting the number of threads for a pair of images of the mentioned above size when using the average value combination rule is equal to 0.14 threads per centimeter. In this case, the error in measuring the thread thickness did not exceed 0.0136 mm when measured from a single image and 0.007 mm when measured from a pair of images. In the case of combining the Fourier spectra according to the maximum value rule, the measurement errors decreased by about one and a half times. The control of the correctness of counting the number of threads was carried out manually.

Conclusions. The proposed algorithm based on combining images measures the characteristics of painting canvases with a distorted texture. The accuracy of measurements exceeds the accuracy of previously used algorithms. The proposed algorithm simplifies and speeds up the preparation of images for analysis, since for measurements in one direction of the threads (warp or weft), only one sample is examined instead of three or four in the algorithm described in the paper [4].

[1] *Johnson D. H., Johnson C., Klein A. G., Sethares W. A., Lee H., Hendriks E.* A thread counting algorithm for art forensics // 2009 IEEE 13th Digital Signal Processing Workshop and 5th IEEE Signal Processing Education Workshop, 2009. — p. 679–684.

[2] *Pan R., Gao W., Li Z., Gou J., Zhang J., Zhu D.* Measuring thread densities of woven fabric using the fourier transform // Fibres & Textiles in Eastern Europe, 2015.

[3] *Aldemir E., Özdemir H., Sarı Z.* An improved gray line profile method to inspect the warp–weft density of fabrics // The Journal of The Textile Institute, 110(1), 2019. — p. 1–12.

[4] *Murashov D.M., Berezin A.V., Ivanova E.Yu.* Algorithms Based on Maximization of the Mutual Information for Measuring Parameters of Canvas Texture from Images // Lecture Notes in Computer Science, 12665. 2021. — p. 77–89.

Анализ цифровых изображений солнечных пятен на основе непрерывной морфологической модели

Местецкий Леонид Моисеевич[1] Mestlm@mail.ru
Бербер Кирилл Андреевич[1]⋆ Berberkirill@mail.ru
[1]Москва, МГУ, кафедра ММП

Солнечные пятна – это относительно тёмные области на Солнце, хорошо наблюдаемые с Земли в различных обсерваториях. Количество, форма и расположение солнечных пятен используется для расчётов показателей солнечной магнитной активности. Оценка параметров пятен выполняется на основе анализа цифровых изображений солнечного диска высокого разрешения. При этом традиционно используются дикретные модели и методы обработки изображений пятен, основанные на просмотре, сегментации и подсчете пикселей изображения. Дискретные модели пятен сложно использовать для анализа формы пятен, вычисления количественных признаков, в частности, площади и периметра. Необходимо учитывать, что фактические размеры элементов солнечной поверхности, соответствующих пикселям, существенно зависят от положения пикселей на солнечном диске. При формировании групп пятен обычно пятна рассматриваются как точечные объекты, их группирование осуществляется по расстоянию между этими точками. При таком подходе могут получаться некорректные результаты в тех случаях, когда пятна имеют большие размеры и неправильную форму.

Предлагаемый подход к анализу формы и расположения солнечных пятен основывается на построении непрерывной морфологической модели [1] для изображения солнечного диска. Непрерывная морфологическая модель представляет собой многосвязную многоугольную фигуру и её медиальное представление, включающее срединную ось (множество точек-центров вписанных окружностей) и радиальную функцию, равную радиусу вписанной окружности в точках срединной оси. Многоугольная фигура используется для аппроксимации и анализа формы солнечных пятен, а срединная ось – для формирования групп пятен.

Многоугольная фигура строится путём аппроксимации бинарного изображения солнечного диска так называемыми разделяющими многоугольниками минимального периметра. В результате такой аппроксимации все солнечные пятна описываются непересекающимися простыми многоугольниками. Координаты вершин многоугольников далее пересчитываются из системы координат изображения в систему координат солнечной полусферы.

Скелет фигуры представляет собой плоский геометрический граф на солнечном диске. Каждое пятно лежит внутри отдельной грани этого графа, а рёбра состоят из точек центров пустых окружностей, касающихся многоугольников, описывающих границы пятен. Окружности, имеющие радиус меньше заданного порогового значения, касающиеся двух разных многоугольников - разных

пятен, называются опорными. Именно опорная окружность определяет расстояние между парой пятен. По этим попарным расстояниям между соседними пятнами далее производится группирование пятен (Рис. 1).

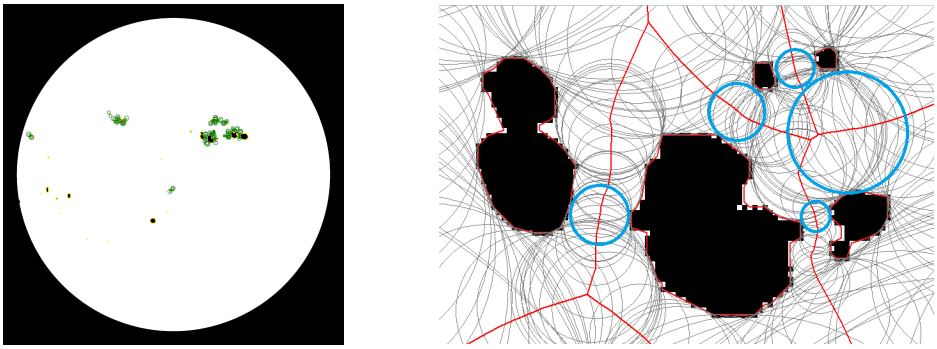


Рис. 1. Группирование на основе внешней скелетезации пятен (слева). Минимальные окружности, касающиеся различных кластеров (справа).

Эксперименты показали, что такая модель формирования пятен позволяет более адекватно описать соседство пятен по сравнению с существующими методами группирования [2] в случае, когда их размеры сильно отличаются от точечных объектов. Моделирование пятен с помощью аппроксимирующих многоугольников позволяет более тонко проанализировать форму пятен, в частности, выделить те из них, которые имеют особо сложную форму. В качестве параметра для оценки сложности формы исследовался коэффициент $k = \frac{P^2}{4\pi S}$, где P - периметр пятна, S - его площадь.

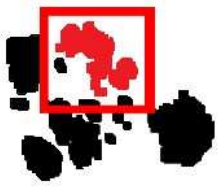


Рис. 2. Степень вытянутости пятна 3.76.

Этот коэффициент, условно названный коэффициентом вытянутости, легко вычисляется для аппроксимирующего многоугольника пятна с учётом коррекции вершин при пересчёте в систему координат солнечной полусферы (Рис. 2).

В исследовании на основе анализа сложности формы солнечных пятен выдвинута гипотеза о связи этого показателя с показателями активности Солнца.

Предлагается сопоставлять каждому изображению Солнца наибольший из всех коэффициентов k всех пятен на этом изображении. Предварительные эксперименты, основанные на сравнении общепринятых индексов измерения солнечной активности с предложенным в работе, показали, что имеется значительная корреляция между солнечной активностью и морфологией пятен: чем больше солнечная активность, тем более неправильными будут формы пятен и тем большими будут их коэффициенты степени вытянутости.

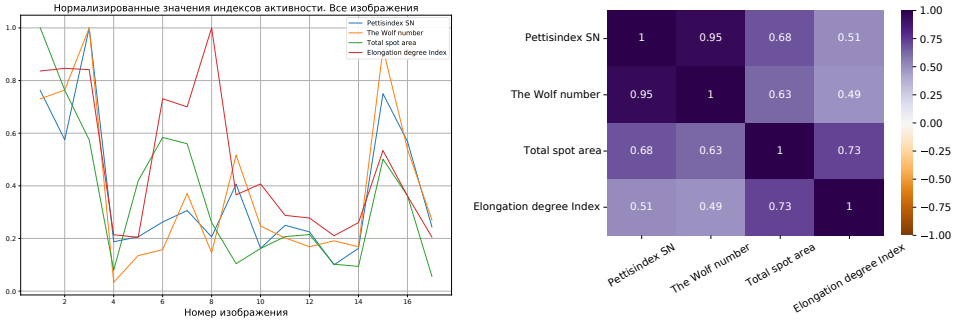


Рис. 3. Нормализированные значения индексов солнечной активности (слева). Матрица корреляции Пирсона (справа).

В финальной части работы представлено сравнение описанного выше признака активности Солнца с уже существующими классификациями солнечной активности, а также подсчет соответствующих корреляций (Рис. 3). Матрица корреляции Пирсона строится между нормализованными значениями индексов солнечной активности и предложенного в работе индекса, основанного на коэффициенте степени вытянутости пятен. В работе прослеживается корреляция между данными величинами. Более масштабные эксперименты предполагается провести в дальнейших исследованиях.

Работа поддержана грантом РФФИ No. 20-01-00664.

[1] *Местецкий Л. О* Непрерывная морфология бинарных изображений: фигуры, скелеты, циркуляры., Москва: ФИЗМАТЛИТ, 2009.

[2] *Trung T.* A Hybrid System for Learning Sunspot Recognition and Classification, Cheju, Korea: IEEE, 2006.

Analysis of sunspots on digital images based on a continuous morphological model

Mestetskiy Leonid[1] Mestlm@mail.ru
Berber Kirill[1]⋆ Berberkirill@mail.ru
[1]Moscow, MSU, department of CMC

Sunspots are relatively dark areas on the Sun that are easily visible from Earth at various observatories. The number, shape and location of sunspots are used to calculate solar magnetic activity. The estimation of sunspot parameters is based on the analysis of high-resolution digital images of the solar disk. In this case, discrete models and methods for processing images of spots are traditionally used, based on viewing, segmentation, and counting image pixels. Discrete spot models are difficult to use for analyzing the shape of spots, calculating quantitative features, in particular, area and perimeter. It should be taken into account that the actual size of the solar surface elements corresponding to pixels depend significantly on the position of the pixels on the solar disk. When forming groups spots are usually considered as point objects, their grouping is carried out according to the distance between these points. With this approach, incorrect results can be obtained in cases where the spots are large and irregular in shape.

The proposed approach to the analysis of the shape and location of sunspots is based on the construction of a continuous morphological model [1] for the image of the solar disk. A continuous morphological model is a multiply connected polygonal figure and its medial representation, including the median axis (a set of points-centers of inscribed circles) and a radial function equal to the radius of the inscribed circle at the points of the median axis. The polygonal figure is used to approximate and analyze the shape of sunspots, and median axis is used to form sunspot groups (Fig. 1).

The polygonal figure is constructed by approximating the binary image of the solar disk with the so-called separating polygons of the minimum perimeter. As a result of such an approximation, all sunspots are described by non-intersecting simple polygons. The polygon vertex coordinates are recalculated from the image coordinate system to the solar hemisphere coordinate system.

The skeleton of the figure planar geometric graph on the solar disk. Each spot lies inside a separate face of this graph, and the edges consist of the points of the centers of empty circles tangent to the polygons that describe the boundaries of the spots. Circles with a radius less than a given threshold value, touching two different polygons - different spots, are called reference. It is the reference circle that determines the distance between a pair of spots. According to these pairwise distances between adjacent spots, the spots are further grouped.

Experiments have shown that such a model of spot formation makes it possible to more adequately describe the neighborhood of spots, in comparison with the

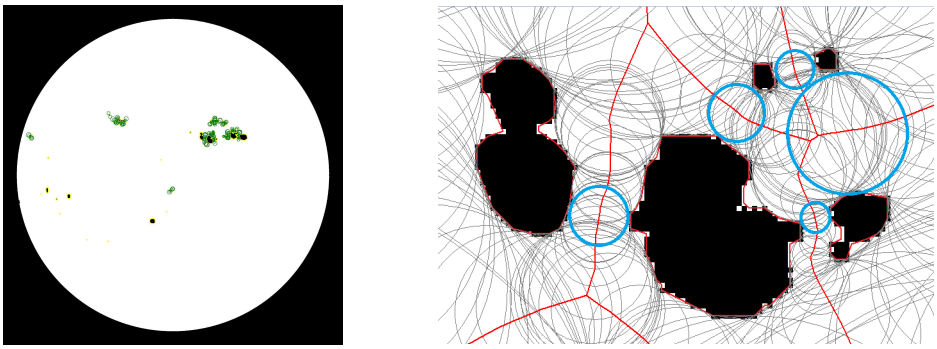


Fig. 1. Grouping based on the external skeletonization of spots (left). Minimal circles tangent to different clusters (right).

existing [2] grouping methods, in the case when their sizes are very different from point objects.

Modeling spots using approximating polygons makes it possible to analyze the shape of spots more finely, in particular, to identify those that have a particularly complex shape. The coefficient $k = \frac{P^2}{4\pi S}$, where P is the perimeter of the spot, S is its area, was studied as a parameter for estimating the complexity of the shape. This coefficient, conventionally called the elongation coefficient, is easily calculated for the approximating sunspot polygon, taking into account the correction of the vertices when recalculated into the coordinate system of the hemisphere sun (Fig. 2).

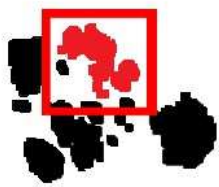


Fig. 2. The elongation coefficient 3.76.

In the study, based on the analysis of shape of sunspots, a hypothesis was put forward about the relationship of this indicator with indicators of solar activity.

It is proposed to assign to each image of the Sun the largest of all coefficients k of all spots in this image. Preliminary experiments based on a comparison of generally accepted indices for measuring solar activity with those proposed in the work showed that there is a significant correlation between solar activity and sunspot morphology: the greater the solar activity, the more irregular the sunspot shapes and the greater their elongation.

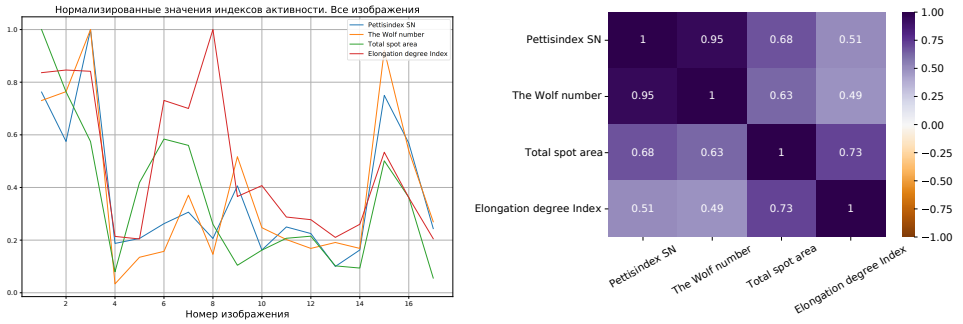


Fig. 3. Normalized values of solar activity indices (left). Pearson's correlation matrix (right).

In the final part of the work, the sign of solar activity described above is compared with the already existing classifications of solar activity, as well as the calculation of the corresponding correlations (Fig. 3). The Pearson correlation matrix is constructed between the normalized values of the solar activity indices and the index proposed in the paper, based on the sunspot elongation coefficient. The paper traces the correlation between these values. Larger scale experiments are expected to be carried out in future studies.

This research is funded by RFBR, grant No. 20-01-00664.

[1] *Mestetskiy L.* Continuous morphology of binary images: figures, skeletons, circulars., Moscow: FIZMATLIT, 2009.

[2] *Trung T.* A Hybrid System for Learning Sunspot Recognition and Classification, Cheju, Korea: IEEE, 2006.

Оценка равномерности моментов отсчетов во временных рядах

Хисматуллин Владимир Витальевич[1]⋆ vladimirkhismatullin@gmail.com
Майсурадзе Арчил Ивериевич[1] maysuradze@cs.msu.ru
[1]Москва, МГУ имени М. В. Ломоносова

Современное производственное оборудование, например на промышленных предприятиях, содержит сотни различных датчиков, постоянно контролирующих производственный процесс. Информация с датчиков собирается и обрабатывается на программируемых логических контроллерах (ПЛК), управляющих процессом производства. Так как системы сбора данных, например SCADA, работают с ПЛК, а не с датчиками напрямую, то значения с датчика поступают в систему не через равные промежутки времени. Поэтому при многократном запросе данных можно говорить о получении временного ряда с неравномерными моментами отсчетов.

Существует несколько способов получать данные с ПЛК, однако скорость и качество передачи данных каждого из них сильно зависит от загруженности и пропускной способности локальной сети, а также нагрузки на опрашиваемое устройство. Именно поэтому после запуска любого решения нужно иметь возможность оценить качество сбора данных. Одним из основных показателей качества сбора временных рядов является равномерность моментов времени передачи данных.

Введем обозначения. Пусть $\{(t_i, X(t_i))\}_{i=1}^n$ — анализируемый временной ряд. Далее значения $X(t_i)$ не представляют интереса, анализируются только моменты времени прихода данных t_i. Пусть $\{\Delta_i = t_{i+1} - t_i\}_{i=1}^{n-1}$ — длины задержек между приходом последующих блоков информации. Ряд называется равномерным, если моменты t_i образуют арифметическую прогрессию, иными словами, если все Δ_i равны. Данное условие во многих приложениях слишком сильное, поэтому на практике проверяется статистическая равномерность: $t_1, t_2, ..., t_{n-1}, t_n \sim \mathbb{U}(a, b)$.

Неудобство состоит в том, что традиционные статистические критерии равномерности плохо интерпретируемы. При анализе равномерности прихода данных выделяется два основных типа нарушений: задержки в приходе информации, выражаемые выбросами среди Δ_i и качественные изменения скорости прихода информации, выражаемые изменением распределения Δ_i. В работе предлагается новый способ оценки равномерности, обладающей большей информативностью для пользователей.

Предлагаемый метод основан на известных подходах к разбиению ряда на «однородные» сегменты. В частности, предлагаемое решение использует идею поиска точек изменения (change point detection) [1], которая состоит в следующем. Пусть набору индексов $1 = i_0 < i_1 < i_2 < \cdots < i_m < i_{m+1} = n+1$ сопоставляется разбиение временного ряда на $m + 1$ сегмент $x_{i_{k-1}:i_k} = \{t_j | j \in [i_{k-1}, i_k -$

$-1]\}$. Тогда функционал качества разбиения ряда на однородные сегменты имеет следующий общий вид:

$$Q(i_1, ..., i_m) = \sum_{k=1}^{m+1} \mathcal{C}(x_{i_{k-1}:i_k}) + \mathcal{R}(m),$$

где $\mathcal{R}(m)$— член, регулирующий сложность модели (количество индексов), а неотрицательный функционал качества разбиения сегмента $\mathcal{C}(x_{i:j})$ выбирается исходя из содержания задачи. Для анализа равномерности в качестве функционала мы выбрали среднеквадратичное отклонение на сегменте.

Результатом минимизации функционала $Q(i_1, ..., i_k)$ по всевозможным наборам индексов является набор точек изменения ряда

$$(m^*, i_1^*, i_2^*, ..., i_{m^*}^*) = \underset{\substack{m,(i_1,i_2...,i_m)\\ 1<i_1<i_2...<i_m<n}}{\arg\min} Q(i_1, ..., i_m).$$

В качестве меры равномерности предлагается взять число, являющееся отношением качества разбиения с учетом точек изменения к значению функционала в целом

$$0 \leqslant \frac{1}{\mathcal{C}(x_{1:n})} \left(\sum_{k=1}^{m^*+1} \mathcal{C}(x_{i_{k-1}^*:i_k^*}) \right) \leqslant 1.$$

Преимущество данной меры заключается в использовании информации о систематических изменениях, а именно точках перехода и учете связи между соседними значениями ряда длин задержек. Приведём простую иллюстрацию. Допустим, что некоторое время информация с ПЛК приходила с одной задержкой, а после сбоя длина задержки изменилась. В этом случае ряд можно разбить на два равномерных сегмента, т. е. $\Delta_1 = \Delta_2 = ... = \Delta_i \neq \Delta_{i+1} = \Delta_{i+2} = ... = \Delta_{n-1}$. Для данного неравномерного ряда введенная мера равномерности будет близка к нулю, в то время как p-value принадлежности $t_1, ..., t_n$ равномерному распределению может достигать единицы и зависит от значения отношения $\frac{\Delta_i}{\Delta_{i+1}}$.

Экспериментальное исследование было произведено на реальных данных, полученных с промышленного производства напитков. Предлагаемый метод показал лучшее качество в терминах интерпретируемости результата. С одной стороны, после правильного подбора функционала регуляризации, алгоритм способен учесть все систематические выбросы. С другой, на сегментах ряда, содержащих сбои, выражающиеся изменением скорости времени сбора данных, мера оказывается близка к нулю. Единственным недостатком относительно базовых методов является необходимость подбора гиперпараметров регуляризации.

Работа выполнена при поддержке НОШ МГУ «Мозг, когнитивные системы, искусственный интеллект», НИР МГУ 5.1.21, гранта РФФИ No. 20-01-00664.

[1] *R. Killick, P. Fearnhead, and I. A. Eckley* Optimal detection of changepoints with a linear computational cost // Journal of the American Statistical Association, vol. 107, no. 500, pp. 1590–1598, 2012.

Sampling Uniformity Evaluation of Moments in Time Series

Khismatullin Vladimir[1]⋆ vladimirkhismatullin@gmail.com
Maysuradze Archil[1] maysuradze@cs.msu.ru
[1]Moscow, Lomonosov Moscow State University

Modern production equipment, such as the one used in production lines, contains hundreds of different sensors controlling the manufacturing process. The information from these sensors is collected and processed with the use of programmable logic controllers (PLCs), which are also responsible for the control of the manufacturing process. Since data acquisition systems, such as SCADA, interact directly with PLCs, sensors measurements may enter the system with different delays between timestamps. Thus, repeated process of collecting data from these devices results in a time series with irregular sampling time.

There are several ways of collecting data from PLC, each of which has different speed and quality of transmission. Both characteristics are heavily dependent on the network congestion and capacity as well as the capacity of the device itself. That is why in order to compare implemented algorithms, it is necessary to have the ability to evaluate the quality of resulting time series. The uniformity of sampling time is one of the main indicators of quality of data collection process.

In order to proceed, we first need to clarify some definitions. Assume $\{(t_i, X(t_i))\}_{i=1}^n$ is the analyzed time series. Values $X(t_i)$ are not considered in this work. The only analyzed object is the series of data arrival timestamps t_i. The sequence $\{\Delta_i = t_{i+1} - t_i\}_{i=1}^{n-1}$ denotes the series of delays between consecutive data arrival timestamps. In this notation a time series is called uniform, if arrival times t_i form an arithmetic progression, i. e. Δ_i is constant. This condition is very strong and in practice is substituted with the condition of statistical uniformity : $t_1, t_2, ..., t_{n-1}, t_n \sim \mathbb{U}(t_1, t_n)$.

The main drawback of traditional statistical uniformity measure is poor interpretability. There are two main types of irregularities studied in the analysis of time series unevenness. The first one being the considerable delays in data arrival times, i. e. outliers in terms of Δ_i. The other being significant changes in overall speed of data transferring, i. e. changes in the distribution of Δ_i. This work introduces a new, more informative measure of uniformity.

The proposed method is based on well-known approaches to the problem of dividing a time series into homogeneous segments. In particular, our solution uses the idea of change point detection [1], which consists in the following. Let each set of indexes $1 = i_0 < i_1 < i_2 < \cdots < i_m < i_{m+1} = n + 1$ divide a time series into $m + 1$ segments $x_{i_{k-1}:i_k} = \{t_j | j \in [i_{k-1}, i_k - 1]\}$. Then the general form of the cost function used to detect change points is as follows:

$$Q(i_1, ..., i_m) = \sum_{k=1}^{m+1} \mathcal{C}(x_{i_{k-1}:i_k}) + \mathcal{R}(m),$$

where $\mathcal{R}(m)$ is the regularization term, and the non-negative functional of complexity of a segment $\mathcal{C}(x_{i:j})$ is chosen based on the given problem. In our study the complexity of a segment is the standard deviation of the values contained in the segment.

The result of minimization of the cost function $Q(i_1, ..., i_k)$ over all possible sets of indexes is the set of change points:

$$(m^*, i_1^*, i_2^*, ..., i_{m^*}^*) = \underset{\substack{m,(i_1,i_2...,i_m) \\ 1<i_1<i_2...<i_m<n}}{\arg\min} Q(i_1, ..., i_m).$$

We define the measure of uniformity as the ratio of minimal complexity with respect to change points, to complexity of the original segment

$$0 \leqslant \frac{1}{\mathcal{C}(x_{1:n})} \left(\sum_{k=1}^{m^*+1} \mathcal{C}(x_{i_{k-1}^*:i_k^*}) \right) \leqslant 1.$$

The advantage of this measure is the utilization of both the information about systematical changes, namely the change points, and the information about connections between close (index-wise) values. Let us give a simple illustration of these properties. Suppose the data had been arriving at the same rate for the first i moments and then the arrival frequency has changed. In this case the series of time delays can be divided into two uniform segments, i. e. $\Delta_1 = \Delta_2 = ... = \Delta_i \neq = \Delta_{i+1} = \Delta_{i+2} = ... = \Delta_{n-1}$. For this uneven example the proposed measure of uniformity will be close to zero, while the p-value of $t_1, ..., t_n$ belonging to the uniform distribution may appear close to one and in general depends on the value of $\frac{\Delta_i}{\Delta_{i+1}}$.

The experimental study has been carried out on data from several PLCs installed on an industrial production line of soft beverages. The proposed measure has shown the best quality in terms of interpretability. Firstly, with the properly selected regularization term the algorithm has succeeded in detecting all outliers. Secondly, the uniformity measure for time series segments containing significant structural changes has appeared to be close to zero. The only downside of the method is the necessity to find suitable regularization hyperparameters.

The research is supported by Scientific and educational school of Moscow State University "¡Brain, cognitive systems, artificial intelligence"¿, research work of Moscow State University 5.1.21, RFBR grant No. 20-01-00664.

[1] *R. Killick, P. Fearnhead, and I. A. Eckley* Optimal detection of changepoints with a linear computational cost // Journal of the American Statistical Association, vol. 107, no. 500, pp. 1590–1598, 2012.

Применение инструментов марковских моделей с групповым обслуживанием к решению задач синхронизации данных в распределенных системах

Азарнова Татьяна Васильевна[1]⋆ ivdas92@mail.ru
Полухин Павел Валерьевич[1] alfa_force@bk.ru
[1]Воронеж, Воронежский Государственный Университет

Применение методов математического моделирования для исследования процессов синхронизации распределенных систем обработки данных позволяет существенно оптимизировать данные процессы и обеспечить требуемый уровень их согласованности. Для анализа стохастических процедур распределенной синхронизации необходимы инструменты математического моделирования, способные адекватно отражать их стохастический характер и формировать вероятностную оценку временных характеристик владения разделяемым ресурсом для каждого из параллельных процессов. Проведенное в рамках работы исследование показывает, что инструменты стохастического моделирования способны решать целый ряд сложных проблем в данной сфере и, в частности, провести адаптацию алгоритмов синхронизации для использования в рамках неоднородных вычислительных систем с разным набором аппаратных компонентов, обеспечивающую приемлемое время синхронизации большого объема данных. В работе анализируются в основном марковские модели синхронизации с групповым обслуживанием и рассматриваются инструменты приведения к данным моделям в случае произвольного времени обслуживания.

Среди алгоритмов синхронизации выделяют алгоритмы на основе событий и передачи маркера. Алгоритмы на основе событий (волновые алгоритмы) осуществляют широковещательную рассылку нового значения синхронизируемого параметра всем потокам, запросившим доступ к данному ресурсу. Синхронизация на основе маркера выполняется в случае получения токена одним из процессов, остальные процессы ожидают получение токена в порядке первичного запроса на получение доступа к общему ресурсу. В распределенной системе Spark, наиболее простой алгоритм синхронизации основывается на полном копировании содержимого переменной между процессами. В этом случае каждый процесс хранит локальную копию всех переменных внутри блоков параллельных функций в распределенной памяти (RDD). Наиболее оптимальным подходом является предварительная рассылка переменных и сохранение их в RDD. В качестве встроенного механизма синхронизации Spark используется реализация протокола BitTorrent. Основным элементом данного протокола является «трекер», хранящий информацию относительно всех доступных узлов, для которых необходимо провести синхронизацию. Каждый из узлов может обмениваться данными непосредственно с соседними узлами и осуществлять рассылки метаинформации относительно тех блоков данных, которые он уже имеет у себя. В отличии от классического алгоритма BitTorrent, в Spark производится ва-

лидация не отдельных блоков данных, а общего блока на завершающем этапе выполнения алгоритма. Это дает возможность сократить накладные расходные, связанные с вычислением хэш-сумм передаваемых фрагментов, однако при возникновении коллизий, требуется повторная пересылка всего блока данных с последующим пересчетом контрольной суммы. В данном исследовании будет рассмотрена синхронизация на основе применения Марковских моделей теории массового обслуживания. Представление процесса синхронизации в виде классической цепи Маркова применимо только в том случае, если распределенная вычислительная система является однородной и время обслуживания иметь показательный закон распределения. Для представления систем синхронизации с произвольным временем обслуживания наиболее применимым является метод вложенных цепей Маркова (ВЦМ), предложенный Кендаллом. Марковская модель переходов, соответствующая начальной и завершающей стадии процесса синхронизации на основе передачи маркера между распределенными процессами $Q = (Q_1, Q_2, ..., Q_n)$ с параметрами входных потоков $\lambda = (\lambda_1, \lambda_2, ..., \lambda_3)$ и параметрами обслуживания $\mu = (\mu_1, \mu_2, ..., \mu_n)$ будет иметь следующий вид:

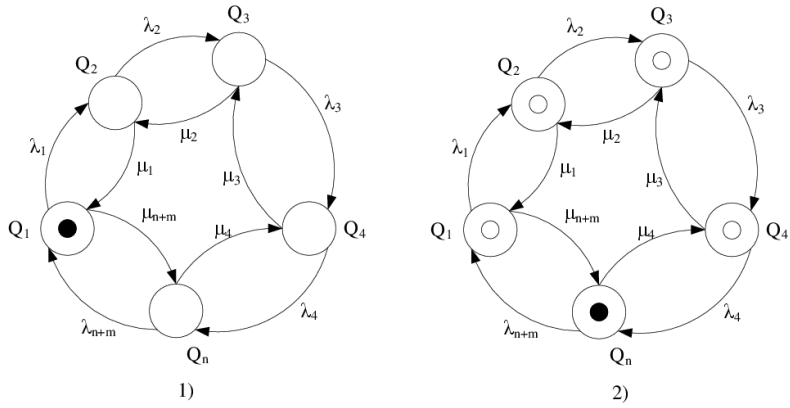


Рис. 1. Марковские модели, соответствующие началу и концу синхронизации распределенного ресурса.

Применение Марковских моделей с групповым обслуживанием (ММГО) для решения задач синхронизации обладает рядом преимуществ, связанных с возможностью: моделирования анализируемых процессов в условиях разнородной распределенной системы, оценки состояния системы, прогнозирования времени синхронизации. В данной работе рассматриваются модели синхронизации с пуассоновским потоком требований и более широким спектром распределений времени обслуживания за счет применения метода ВЦМ. В рамках модели передачи маркера и реализации процедуры рассылки широковещательных

сообщений для каждого из процессов используется распределенный алгоритм Сузуки-Касами. В рамках практической части исследования произведено сравнение существующих алгоритмов синхронизации для платформ Spark и Hadoop и алгоритма синхронизации на основе Марковской модели с групповым обслуживанием. Для этого произведено развертывание кластера Apache Hadoop и Spark, состоящего из 10 узлов. Приведем результаты вычислительного эксперимента по сравнению алгоритмов синхронизации: BitTorrent (встроенный алгоритм Spark), HDFS[5] (встроенный в программную модель распределенной файловой системы HDFS 5 реплик), HDFS[10] (Алгоритм HDFS 10 реплик), ММГО (Разработанная система синхронизации на основе марковской модели группового обслуживания). Зависимость времени от числа потоков-получателей приведем для объема данных до 100 ГБ:

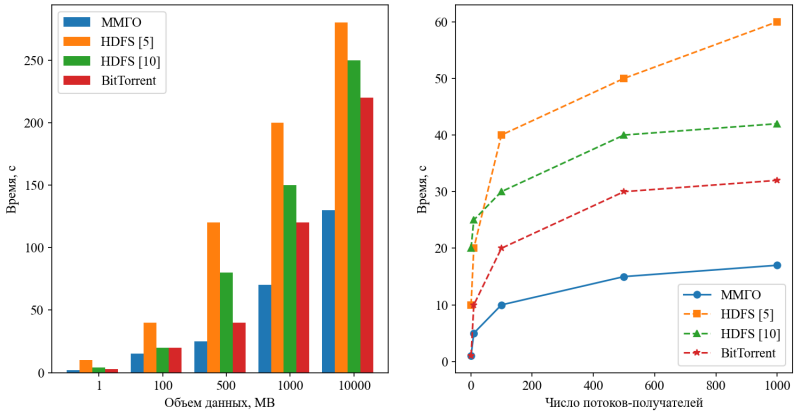


Рис. 2. Оценка алгоритмов синхронизации для платформы Spark.

Как показывают результаты исследования применение алгоритм ММГО оптимизирует распределение данных между областями RDD-записей каждого их узлов кластера. В свою очередь ММГО позволяет обеспечить требуемый уровень устойчивости к росту потоков-получателей, что повышает горизонтальную масштабируемость вычислительной системы. Для случая разнородной аппаратной платформы использование предложенного инструментария предоставляет возможность произвести оптимальную настройку конфигурации для каждого из узлов путем расчета параметров соответствующих моделей массового обслуживания.

[1] *Azarnova T. V., Polukhin P. V.* Distributed computing systems synchronization modeling for solving machine learning tasks // IOP Conf. Ser., Bristol: IOP Publishing, 2021.

Application of Markov models tools with group services for solving tasks of data synchronization in distributed systems

Azarnova Tatyana[1]⋆ ivdas92@mail.ru
Polukhin Pavel[1] alfa_force@bk.ru
[1]Voronezh, The Voronezh State University

Appling mathematical modeling methods for analyses synchronization processes of distributed data processing systems makes it possible to significantly optimize these processes and guarantee the required level of consistency. To analyze stochastic distributed synchronization procedures, mathematical modeling tools are needed to adequately reflect their stochastic nature and form a probabilistic time characteristics assessment of the shared resource ownership distributed between the parallel processes. The research carried out within the current paper shows that stochastic modeling tools are able to solve a number of complex problems in this area and in particular to adapt synchronization algorithms for use in heterogeneous computing systems with a different set of hardware components and providing an acceptable synchronization time for a large amount of data. The paper essentially analyzes Markov synchronization models with group maintenance and considers tools for bringing these models to the case of arbitrary service time.

Among the synchronization algorithms, algorithms based on events and token transmission are distinguished. Event-based algorithms (wave algorithms) broadcast the new value of the synchronized parameter to all streams requesting access to the resource. Token-based synchronization is performed when a token is received by one of the processes, the remaining processes wait for the token to be received in the order of the primary request to access the share. In the Spark distributed system, the simplest synchronization algorithm is based on a full copy of the content of the variable between processes. In this case, each process stores a local copy of all variables within the parallel function blocks in distributed memory (RDD). The best approach is to pre-send variables and store them in the RDD. The built-in Spark synchronization mechanism uses the BitTorrent protocol implementation. The main element of this protocol is the tracker, which stores information about all available nodes for which synchronization is required. Each of the nodes can communicate directly with neighboring nodes and distribute meta-information about those blocks of data that it already has. Unlike the classic algorithm BitTorrent, Spark does not validate individual data blocks, but a common block at the final stage of the algorithm. This makes it possible to reduce the overhead associated with calculating the hash sums of the transmitted fragments, however, when collisions occur, it is necessary to resend the entire data block and then recalculate the checksum. This study will consider synchronization based on the application of Markov models of mass service theory. The presentation of the synchronization process in the form of a classic Markov circuit is applicable only if the distributed computing system is homogeneous and the service time has an indicative distribution law. To represent random

service time synchronization systems, the most applicable method is the embeddeb Markov chains (EMC) method proposed by Kendall. Markov transition model corresponding to the initial and final stage of the synchronization process based on token transfer between distributed processes $Q = (Q_1, Q_2, ..., Q_n)$ with input stream parameters $\lambda = (\lambda_1, \lambda_2, ..., \lambda_3)$ with service parameters $\mu = (\mu_1, \mu_2, ..., \mu_n)$ will have the following form:

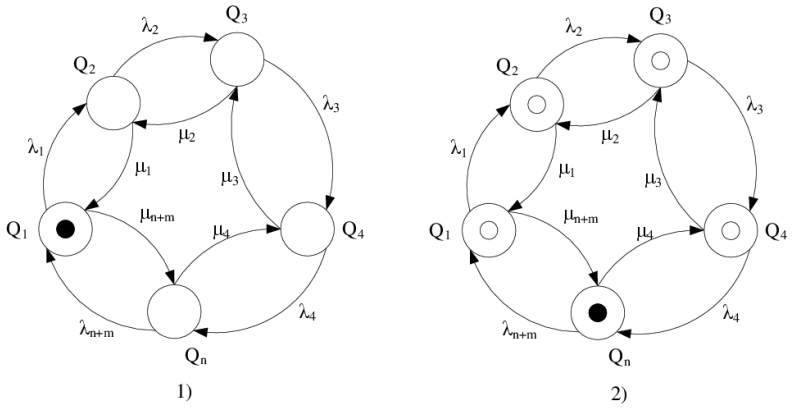


Fig. 1. Markov models corresponding to the beginning and end distributed resource synchronization.

The application of Markov models with group processing (MMGP) to solve synchronization problems has a number of advantages associated with the ability to model analyzed processes in a heterogeneous distributed system, evaluate the state of the system, predict the timing of synchronization. This paper discusses synchronization models with a Poisson flow of requirements and a wider spectrum of service time allocation through the use of the EMC method. A distributed Suzuki-Kasami algorithm is used for each process within the token transmission model and broadcast message distribution procedure. As part of the practical part of the study, a comparison of the existing synchronization algorithms for the Spark and Hadoop platforms and the synchronization algorithm based on the Markov model with group service was made. This is done by deploying a 10-node Apache Hadoop and Spark cluster. Let's give the results of a computational experiment compared to synchronization algorithms: BitTorrent (integrated algorithm Spark), HDFS [5] (built into the software model of the distributed file system HDFS 5 replicas), HDFS [10] (algorithm HDFS 10 replicas), MMGP (Developed synchronization system based on the Markov group service model). The time dependence on the number of recipient threads is given for up to 100 GB of data:

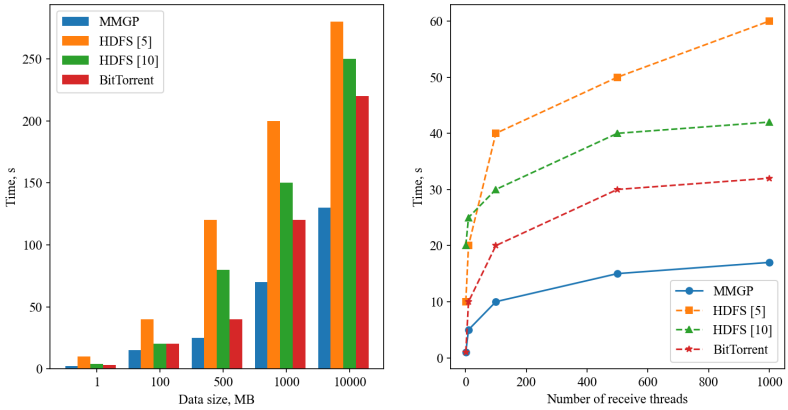


Fig. 2. Evaluation of synchronization algorithms for the Spark and Hadoop platforms.

As the results of the research show tha application of the MMGP algorithm optimizes the distribution of data between the regions of RDD records for each of their cluster nodes. In turn, MMGP allows you to provide the required level of resistance to the growth of recipient flows, which increases the horizontal scalability of the computer system. In the case of a heterogeneous hardware platform, the use of the proposed tools provides the opportunity to optimally configure each node by calculating the parameters of the respective mass service models.

[1] *Azarnova T. V., Polukhin P. V.* Distributed computing systems synchronization modeling for solving machine learning tasks // IOP Conf. Ser., Bristol: IOP Publishing, 2021.

Автоматическая транскрипция мелодики речи с использованием музыкальной нотации на основе модели восприятия человеком высоты звука

Трифонов Иван Николаевич[1]⋆ ivan14trifonov@mail.ru
Копылов Андрей Валериевич[1] andkopylov@gmail.com
[1]Тула, Тульский государственный университет

Важной составляющей анализа человеческой речи, помимо содержащейся в ней текстовой информации, является определение особенностей произношения, дополнительных по отношению к основной артикуляции звуков речи и не выделяющихся при членении речи на фонемы. Такие особенности получили название просодических характеристик речи. Важнейшей просодической характеристикой является система изменений относительной высоты тона речи, называемая мелодикой речи [1]. Физической характеристикой, соответствующей мелодике речи, является контур частоты основного тона (ЧОТ). Полезным инструментом изучения просодических характеристик, в том числе мелодики речи, является просодическая транскрипция. Под транскрипцией будем понимать графическую фиксацию звуковых характеристик речи. Пит Мертенс [2], в зависимости от подхода к интерпретации просодических характеристик, выделяет три основных типа просодической транскрипции:

1) символьная транскрипция;
2) транскрипция, основанная на акустических характеристиках речи;
3) транскрипция, выполненная с помощью слухового анализа.

Отдельного внимания заслуживает разновидность символьной транскрипции, основанная на музыкальной нотации. Музыкальная нотация была выработана в течении многовековой истории музыки, и поэтому отражает особенности восприятия человека и является удобным способом фиксации характеристик звука. Запись мелодии речи в музыкальной нотации позволяет изучать взаимосвязь речи и музыки [3]. Представляли речь с использованием музыкальной нотации и изучали просодические характеристики речи с музыкальной точки зрения композиторы Бах, Бетховен, Яначек, Шёнберг, исследователи Джошуа Стил, Роман Якобсон и другие [4]. Современные технологии позволяют автоматизировать процесс транскрипции мелодики речи. В [5] (2016) представлена методика просодической транскрипции с использованием нотной записи, а в [4] (2017) — с использованием музыкальной системы, расширенной до 24-х полутонов. Аудиозапись речи разделяется на гласные единицы, для каждой из них вычисляется среднее значение частоты. Сравнивая эту частоту с дискретными значениями частот в четвертитоновой системе, получают запись мелодики речи в музыкальной нотации. Однако контур ЧОТ речи, хотя и является точным физическим описанием речевого сигнала, не является наиболее точным представлением мелодики речи в том виде, в каком она воспринимается человеком.

В ходе эмпирических исследований установлено, что восприятие высоты тона в речи подвержено нескольким перцептивным преобразованиям (преобразованиям восприятия). Одним из них является перцептивное разделение контура ЧОТ на единицы размером со слог из–за быстрых спектральных и амплитудных колебаний речевого сигнала. Второе — временная интеграция ЧОТ внутри слога, которая состоит в том, что, если контур ЧОТ внутри слога короткий и имеет относительно небольшие изменения, слушатели воспринимают одну высоту тона, представляющую собой средневзвешенное по времени значение ЧОТ внутри слога. Если величина изменения ЧОТ внутри слога превышает некоторый «порог глиссандо», то воспринимается скольжение (или несколько скольжений) [6, 2]. Отметим, что в [4] учтена только первая из названных выше особенностей. Разработанная нами система автоматической транскрипции мелодики речи с использованием музыкальной нотации, в отличие от уже существующей системы, описанной в [4], учитывает особенности восприятия человеком высоты звука, описанные как в [2], так и в [7]. Полученные в ходе экспериментов транскрипции мелодики речи в музыкальной нотации по результатам безэталонной экспертной оценки показали лучшее соответствие полученной нотной записи восприятию человека.

Работа выполнена при финансовой поддержке Министерства науки и высшего образования РФ в рамках государственного задания FEWG-2021-0012.

[1] *Трифонов И. Н., Копылов А. В.* Автоматическая транскрипция мелодики речи с использованием музыкальной нотации на основе модели восприятия человеком высоты звука // Известия Тульского государственного университета. Технические науки, 2022., Тула: Издательство ТулГУ, 2022.

[2] *Mertens P.* The prosogram: Semi-automatic transcription of prosody based on a tonal perception model // Speech prosody 2004, international conference, Nara, Japan, March 23-26, 2004.

[3] *Трифонов И. Н.* Просодическая транскрипция как инструмент анализа речи // XVII Региональная магистерская научная конференция (25 – 29 апреля 2022 года): сб. докладов. В 2 ч. Ч. I, Тула: Издательство ТулГУ, 2022. —с. 273.

[4] *Meireles A. R., Simões A. R., Ribeiro A. C., de Medeiros B. R.* Musical Speech: A New Methodology for Transcribing Speech Prosody // Interspeech, 2017. — P. 334–338.

[5] *Simões A. R., Meireles A. R.* Speech prosody in musical notation: Spanish, Portuguese and English // Proceedings of the 8th International Conference on Speech Prosody, Boston, USA, 2016.. — P. 212–216.

[6] *Patel A. D.* An Empirical Method for Comparing Pitch Patterns in Spoken and Musical Melodies: A Comment on JGS Pearl's "Eavesdropping with a Master: Leos Janáček and the Music of Speech." // Empirical Musicology Review, Vol. 1, No. 3, 2006. — P. 166–169.

[7] *d'Alessandro C., Mertens P.* Automatic pitch contour stylization using a model of tonal perception // Computer Speech and Language, Vol. 9, No. 3, Academic Press, 1995. — P. 257–288.

Automatic transcription of speech melody using musical notation based on human pitch model perception

Trifonov Ivan[1]⋆ ivan14trifonov@mail.ru
Kopylov Andrei[1] andkopylov@gmail.com
[1]Tula, Tula State University

An important component of the analysis of spoken language, in addition to the textual information contained in it, is the determination of pronunciation features that are additional to the main articulation of speech sounds and do not stand out when speech is divided into phonemes. Such features are called prosodic characteristics of speech. The most important prosodic characteristic is the system of changes in the relative pitch, called the melody of speech [1]. The physical characteristic corresponding to the melody of speech is the pitch contour. A useful tool for studying prosodic characteristics, including the melody of speech, is prosodic transcription. By transcription we will understand the graphic fixation of the sound characteristics of speech. Piet Mertens [2], depending on the approach to the interpretation of prosodic characteristics, distinguishes three main types of prosodic transcription:

1) character transcription;
2) transcription based on the acoustic characteristics of speech;
3) transcription performed using auditory analysis.

A variety of symbolic transcription based on musical notation deserves special attention. Musical notation was developed during the centuries-old history of music, and therefore reflects the characteristics of human perception and is a convenient way to fix the characteristics of sound. Recording a speech melody in musical notation makes it possible to study the relationship between speech and music [3]. Composers Bach, Beethoven, Janacek, Schoenberg, researchers Joshua Steele, Roman Jakobson and others [4] presented speech using musical notation and studied the prosodic characteristics of speech from a musical point of view. Modern technologies make it possible to automate the process of transcription of speech melody. Paper [5] (2016) presents a prosodic transcription technique using music notation, and [4] (2017) uses a musical system extended to 24 semitones. The audio recording of speech is divided into vowel units, for each of them the average frequency value is calculated. Comparing this frequency with discrete values of frequencies in the quarter–tone system, a recording of the melody of speech in musical notation is obtained. However, the contour of the fundamental frequency (F0), although an accurate physical description of the speech signal, is not the most accurate representation of speech melody as perceived by humans. Empirical research has established that pitch perception in speech is subject to several perceptual transformations (perceptual transformations). One of them is the perceptual division of F0 into syllable-sized units due to fast spectral and amplitude fluctuations of the speech signal. The second is the temporal integration of the intra-syllable F0, which is that if the intra-syllable F0

contour is short and has relatively little change, listeners perceive one pitch, which is the time-weighted average of the intra-syllable F0. If the change in F0 within a syllable exceeds a certain "glissando threshold", then a slip (or several slips) is perceived [6, 2]. Note that [4] takes into account only the first of the above features. The system we have developed for automatic transcription of speech melody using musical notation, unlike the already existing system described in [4], takes into account the peculiarities of human perception of pitch described both in [2] and [7]. The transcriptions of speech melody in musical notation obtained during the experiments, based on the results of an expert assessment, showed the best correspondence of the received musical notation to human perception.

This research is funded by the Ministry of Science and Higher Education of the Russian Federation within the framework of the state task FEWG-2021-0012.

[1] *Trifonov I., Kopylov A.* Automatic transcription of speech melody using musical notation based on human pitch model perception // Proceedings of Tula State University. Technical sciences, 2022., Tula: Publishing house of TSU, 2022.

[2] *Mertens P.* The prosogram: Semi-automatic transcription of prosody based on a tonal perception model // Speech prosody 2004, international conference, Nara, Japan, March 23-26, 2004.

[3] *Trifonov I.* Prosodic transcription as a tool for speech analysis // XVII Regional master's scientific conference (April 25 – 29, 2022): Conference Proceeding. Part I, Tula: Publishing house of TSU, 2022. — 2022. —c. 273 (in Russian).

[4] *Meireles A. R., Simões A. R., Ribeiro A. C., de Medeiros B. R.* Musical Speech: A New Methodology for Transcribing Speech Prosody // Interspeech, 2017. — P. 334–338.

[5] *Simões A. R., Meireles A. R.* Speech prosody in musical notation: Spanish, Portuguese and English // Proceedings of the 8th International Conference on Speech Prosody, Boston, USA, 2016.. — P. 212–216.

[6] *Patel A. D.* An Empirical Method for Comparing Pitch Patterns in Spoken and Musical Melodies: A Comment on JGS Pearl's "Eavesdropping with a Master: Leos Janáček and the Music of Speech." // Empirical Musicology Review, Vol. 1, No. 3, 2006. — P. 166–169.

[7] *d'Alessandro C., Mertens P.* Automatic pitch contour stylization using a model of tonal perception // Computer Speech and Language, Vol. 9, No. 3, Academic Press, 1995. — P. 257–288.

Применение методов морфологического анализа к исследованию временных рядов

Бизин Владислав Константинович[1]⋆ bizin.vlad19@physics.msu.ru
Чуличков Алексей Иванович[1,2] achulichkov@gmail.com
Газарян Варвара Арамовна[1,3] vagazaryan@fa.ru
Шапкина Наталья Евгеньевна[1,4] neshapkina@mail.ru
[1]МГУ им. М. В. Ломоносова, физический факультет, Москва, Россия
[2]ИФА им. А. М. Обухова РАН, Москва, Россия
[3]Финансовый университет при правительстве РФ, Москва, Россия
[4]ИТПЭ РАН, Москва, Россия

Природные и жизненные процессы имеют свойство периодичности в силу циклического обращения Земли вокруг своей оси и Земли вокруг Солнца. Это проявляется во многих биологических, метеорологических и иных физических процессах. В частности, на поведение временных рядов метеорологических параметров Земли влияют, в основном, сезонные и суточные колебания параметров атмосферы. Для подробного изучения менее значимых вкладов и факторов, определяющих временной ряд, возникает необходимость выделения (фильтрации) его суточной или сезонной составляющей.

К существующим методам фильтрации относятся скользящее среднее [1], фильтрация при помощи преобразования Фурье [2], усреднение данных по предполагаемому периоду цикла [3], методы вейвлет анализа [4] и другие. Недостатки этих методов состоят в предположении об априорной известности периода цикла. Однако, природные процессы, вообще говоря, являются квазипериодическими. Кроме того, сам вид временного ряда может отличаться от синусоидального.

Для разрешения этих проблем в данной работе предложен метод морфологической фильтрации временного ряда, выделяющий его циклическую компоненту с переменным периодом. Построение метода основывается на методах морфологического анализа данных, разработанных и развиваемых проф. Ю. П. Пытьевым [5].

Методы морфологического анализа предназначены для изучения структуры сигналов, сохраняющейся при преобразованиях, принадлежащих некоторому классу преобразований. Так, если сигналы представляют собой последовательность участков, выпуклых вниз и выпуклых вверх (или, иными словами, последовательность локальных максимумов и минимумов), то их можно отнести к сигналам одной формы, если положения максимумов и минимумов совпадают. Очевидно, эти свойства сигналов сохраняются при монотонных и выпуклых преобразованиях.

В методах морфологического анализа форма сигнала $f(t)$, $t \in T$ понимается как множество V_f сигналов, полученных из $f(t)$ всевозможными монотонными преобразованиями. Математическую модель сигнала и множество преобразо-

ваний, сохраняющих форму, обычно выбирают так, чтобы для любого предъявленного для анализа сигнала $g(t)$, $t \in T$, была разрешима задача его наилучшего приближения элементами множества V_f, решение которой называют проекцией g на V_f [5]. Проекция ищется как решение задачи на минимум по набору аргументов, называемых параметрами формы.

В настоящей работе строится модель квазипериодического сигнала и рассматриваются разные способы задания его формы (параметризации формы). Для каждого из них предлагается алгоритм построения проекции, который затем применяется к временному ряду концентрации двуокиси углерода, измеренного на метеорологической станции "AsiaFlux"во Вьетнаме и предоставленных ИПЭЭ им. А.Н. Северцова РАН [6]. Результаты фильтрации сравниваются и делается вывод о зависимости вида квазициклической составляющей временного ряда от способа параметризации формы сигнала. Прикладной аспект морфологической фильтрации проявляется в стационарности остаточного ряда, который можно исследовать статистическими методами.

Перспективами развития работы являются усовершенствование вычислительного алгоритма построения проекций, а также введение иных способов задания формы.

Работа поддержана грантом РФФИ No. 19-29-09044.

[1] *Gazaryan V. A., Kurbatova J. A., Ovsyannikov T. A. et al.* Contemporary climate changes in the southwest of the valdai hills: A statistical analysis of the long-term dynamics of the air temperature. // Moscow University Physics Bulletin, 2015. — Vol. 70, no. 5. — p. 346—352.

[2] *Kurbatova J. A., Aleshnovskij V. S., Kuricheva O. A. et al.* Seasonal and interannual variability of co2 above the moist tropical forest of southern vietnam. // IOP Conference Series: Earth and Environmental Science, 2020. — Vol. 606. — p. 1–8.

[3] *Gazaryan V. A., Kurbatova J. A., Ovsyannikov T. A. et al.* A statistical analysis of cyclical changes in the time series of meteorological parameters in the southwest of the valdai hills. // Moscow University Physics Bulletin, 2018. — Vol. 73, no. 1. — p. 61—67.

[4] *Ziuzina N. A., Kurbatova V. A. et al.* Study of time series of meteorological parameters by wavelet analysis. // IOP Conference Series: Earth and Environmental Science, 2020. — Vol. 606. — p. 012069.

[5] *Pytyev Yu. P.* Morphological Image Analysis. // Pattern Recognition and Image Analysis: Advances in Mathematical Theory and Applications, 1993. — Vol. 3, no. 1 — p. 19–28.

[6] *Дещеревская О. А., Авилов В. К., Ба Зуй Динъ, Конг Хуан Чан, Курбатова Ю. А.* Современный климат национального парка Кат Тьен (Южный Вьетнам): использование климатических данных для экологических исследований. // Геофизические процессы и биосфера, 2013. — т. 12, №. 2. — С. 5—33.

Application of morphological analysis methods to the study of time series

Bizin Vladislav[1]⋆ bizin.vlad19@physics.msu.ru
Chulichkov Alexey[1,2] achulichkov@gmail.com
Gazaryan Varvara[1,3] vagazaryan@fa.ru
Shapkina Natalia[1,4] neshapkina@mail.ru

[1]Faculty of Physics, Lomonosov Moscow State University, Moscow Russia
[2]Obukhov Institute of Atmospheric Physics, Russian Academy of Sciences, Moscow Russia
[3]Financial University under the Government of the Russian Federation, Moscow Russia
[4]Institute of Theoretical and Applied Electromagnetics, Russian Academy of Sciences, Moscow Russia

Natural and life processes have the property of periodicity due to the cyclic rotation of the Earth around its axis and the Earth around the Sun. This property takes place in many biological, meteorological and other physical processes. In particular, the behavior of meteorological parameters of the Earth is mainly influenced by seasonal and diurnal variations of atmospheric parameters. In order to study less significant contributions and factors determining the time series one needs to single out (filter) its diurnal or seasonal component.

Existing filtering methods include moving average [1], Fourier transform [2], averaging over the assumed period [3], wavelet transform [4], and others. The disadvantage of these methods is the assumption that the period is known a priori. However, generally speaking, natural processes are quasi-periodic. More than that, the form of the time series itself may differ from sinusoidal.

To solve this problem, a method of morphological filtering of time series is proposed, which singles out its cyclic component with a variable period. The method is based on the methods of morphological analysis, developed by Prof. Y. P. Pytyev [5].

The purpose of morphological analysis methods is to study the structure of signals preserved by transformations belonging to some class. Thus, if each of signals is a sequence of convex and concave segments (i.e. a sequence of maxima and minima), they are considered to be signals of the same form if the positions of maxima and minima coincide. It is clear that these properties of signals are preserved in monotonic and convex transforms.

In morphological analysis methods, the shape of a signal $f(t)$, $t \in T$ is understood as a set of V_f signals obtained from $f(t)$ by all possible monotonic transforms. The mathematical model of the signal and the set of shape-preserving transforms are usually chosen in such a way that for any signal $g(t)$, $t \in T$ to be analyzed, the problem of its best approximation by elements of the set V_f is solvable. The solution is called the projection of g on V_f [5]. The projection is sought as a solution of the minimum problem where the functional is varied by a set of variables called shape parameters.

In this paper we build a model of quasi-periodic signal and consider different ways of setting its shape. For each of them we propose an algorithm for constructing a projection, which is then applied to the time series of carbon dioxide concentration measured at the meteorological station "AsiaFlux" in Vietnam and provided by A. N. Severtsov IPEE RAS [6]. The filtering results are compared and a conclusion about the dependence of quasi-cyclic component on the method of setting the shape is made. The applied aspect of morphological filtering manifests in the stationarity of the residual series, which can be investigated by statistical methods.

Prospects for the development include computational algorithm improvement, as well as introduction of other ways of setting the form.

This research is funded by RFBR, grant 19-29-09044.

[1] *Gazaryan V. A., Kurbatova J. A., Ovsyannikov T. A. et al.* Contemporary climate changes in the southwest of the valdai hills: A statistical analysis of the long-term dynamics of the air temperature. // Moscow University Physics Bulletin, 2015. — Vol. 70, no. 5. — p. 346—352.

[2] *Kurbatova J. A., Aleshnovskij V. S., Kuricheva O. A. et al.* Seasonal and interannual variability of co2 above the moist tropical forest of southern vietnam. // IOP Conference Series: Earth and Environmental Science, 2020. — Vol. 606. — p. 1–8.

[3] *Gazaryan V. A., Kurbatova J. A., Ovsyannikov T. A. et al.* A statistical analysis of cyclical changes in the time series of meteorological parameters in the southwest of the valdai hills. // Moscow University Physics Bulletin, 2018. — Vol. 73, no. 1. — p. 61—67.

[4] *Ziuzina N. A., Kurbatova V. A. et al.* Study of time series of meteorological parameters by wavelet analysis. // IOP Conference Series: Earth and Environmental Science, 2020. — Vol. 606. — p. 012069.

[5] *Pytyev Yu. P.* Morphological Image Analysis. // Pattern Recognition and Image Analysis: Advances in Mathematical Theory and Applications, 1993. — Vol. 3, no. 1 — p. 19–28.

[6] *Deshcherevskaya O. A., Kurbatova J. A., Avilov V. K., Dinh B. D., Tran C. H.* MODERN CLIMATE OF THE CáT TIêN NATIONAL PARK (SOUTHERN VIETNAM): CLIMATOLOGICAL DATA FOR ECOLOGICAL STUDIES. // Izvestiya, Atmospheric and Oceanic Physics, 2013. — Vol. 49, no. 8. — p. 819—838.

Применение машинного обучения в нейрофизиологии для выявления новых функциональных паттернов в многомерных временных рядах

Сидоров Леонид Станиславович[1]⋆ leon.sidorov@gmail.com
Майсурадзе Арчил Ивериевич[1] maysuradze@cs.msu.ru
[1]Москва, Московский государственный университет имени М.В. Ломоносова

Во многих предметных областях исследуемые данные имеют вид многомерных временных рядов. Это могут быть показания датчиков на производственных линиях, акции на фондовом рынке или истории денежных транзакций. Иногда такие временные ряды возникают как результат дополнительной обработки, например, видео может преобразовываться в пучок траекторий специфических точек. При этом на многомерных временных рядах решаются как традиционные задачи классификации, так и специфические задачи поиска аномалий или точки изменения тренда. Соответственно, для всех этих прикладных областей и задач разрабатывают различные модели машинного обучения (МО), которые принимают многомерные временные ряды.

В наши дни исследователи ждут от моделей МО не просто решения прикладных задач, но и интерпретируемости результатов. Модели должны явно демонстрировать пользователям закономерности, которые они нашли. При этом анализ многомерных временных рядов во многих случаях сводится к выявлению так называемых *функциональных паттернов* [2], то есть особенностей поведения временного ряда, соответствующих некоторым интересующим исследователей состояниям системы. На протяжении многих лет эксперты из различных предметных областей уже выявили некоторое количество функциональных паттернов. Задачей данной работы является является предложить методику автоматического поиска подобных паттернов в многомерных временных рядах без обладания какими-либо априорными знаниями о предметной области. Идея исследования состоит в том, чтобы предложить модель, которая как частный случай найдет уже известные паттерны. Развивая подобный подход, в будущем исследователи смогут использовать модели машинного обучения, чтобы быстрее находить новые виды паттернов и закономерностей в исходных данных многомерных временных рядов.

В данной работе в качестве предметной области взята нейрофизиология. Соответственно, работа методики будет продемонстрирована на выявлении одного из самых известных функциональных паттернов в электроэнцефалограммах (ЭЭГ), а именно *волны P300*. Нетрудно видеть, что традиционные паттерны, такие как волна P300, удобно выявлять с помощью скользящих фильтров или операции свёртки [1]. Подобные операции уже реализованы в традиционных нейросетевых моделях МО.

Для экспериментов мы использовали данные, описанные в [2]. Набор данных «Матрица» был записан с помощью матрицы символов, наблюдая за которой

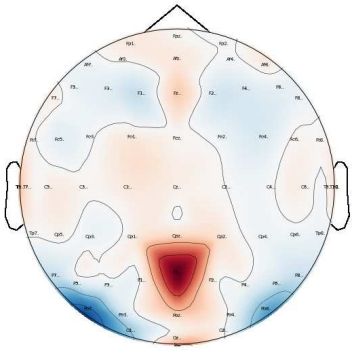


Рис. 1. Карта важности каналов.

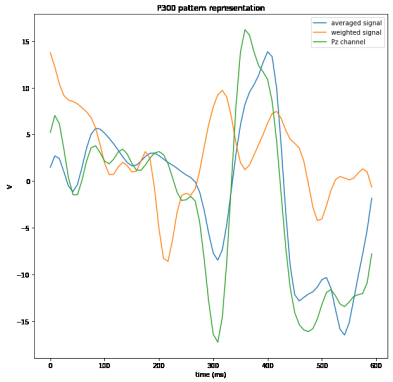


Рис. 2. оптимальные комбинации каналов.

человек может писать слова, концентрируя своё внимание на отдельных символах. Во время эксперимента строки и столбцы матрицы подсвечиваются в случайном порядке, реакция на эти неожиданные для респондента подсвечивания и позволяет понять, какой символ загадал человек. Исходно эти данные были собраны с целью понять, насколько хорошо алгоритмы могут предсказать целевой символ из данных о мозговой активности. Соответственно, в [2] также описаны существующие подходы из классического МО, которые уже достигают достаточно высокого уровня качества на рассматриваемом наборе данных. Но в то время исследователи ещё не рассматривали, что именно находят их модели.

Мы фокусируемся не на исходной прикладной задаче распознавания символов, а на выявлении функциональных паттернов и интерпретации полученных результатов. Для этого мы использовали уже известный в литературе переход от распознавания целевых символов к распознаванию строк и столбцов матрицы, содержащих целевой символ. В таком случае задача может рассматриваться как бинарная классификация. На вход модели МО подаётся фрагмент ЭЭГ, соответствующий подсвечиванию, а выходом является бинарная метка класса (содержит ли текущее подсвечивание целевой символ). Понятно, что волна P300 может быть детектирована в масштабе одной секунды. В экспериментах мы работали с фрагментами по 600 мс.

Согласно исследованиям нейрофизиологов ожидается, но не гарантируется, что подсвечивание целевого символа вызывает на ЭЭГ волну P300. Хотя из обучающей разметки мы знаем, когда загорался целевой символ, эксперимент по постановке не гарантирует, что во время воздействия возникнет волна P300, однако шанс этого события крайне велик. Соответственно, модель МО не получает информацию о наличии волны P300 напрямую. Основная задача данной работы заключается в проверке гипотезы, что модель МО всё-таки сможет сама найти функциональный паттерн, который нейрофизиологам уже известен.

Отметим, что в нейрофизиологии базовым методом определения паттерна является усреднение временных рядов по каждому классу. Даже для ярко выраженных паттернов такой подход требует привлечения десятков респондентов. В наших данных присутствует только два респондента, соответственно базовый метод не использовался ни в [2], ни в нашем исследовании.

Предложенная архитектура нейронной сети поочерёдно применяет фильтры к каналам и временным промежуткам на исходных данных, выявляя тем самым важность тех или иных электродов на голове человека или моментов времени во время записи.

На на рис. 2 видно, что модель самостоятельно выявила оптимальную линейную комбинацию исходных электродов. Pc7 и Pc8 были взяты с отрицательным знаком, чтобы учесть негативные значения потенциала, предшествующие P300, а электрод Pz, наоборот, с положительным, чтобы уловить пик самой волны. Таким образом, модель самостоятельно научилась выявлять волну P300. Качество решения задач бинарной классификации и классификации символов сразу же оказалось на уровне лучших решений из [2].

Карта важности каналов ЭЭГ представлена на рис. 1. Видно, что модель предложила использовать малое число каналов ЭЭГ.

Таким образом, в работе предложена методика, помогающая исследователям по серии экспериментов автоматически выявить функциональный паттерн в многомерных временных рядах. При этом было достаточно формализовать исходную задачу в терминах машинного обучения и не требовалось углубляться в предметную область. Работоспособность методики продемонстрирована в области нейрофизиологии для данных, где уже известен ожидаемый паттерн. Для дальнейшего развития данной идеи представляет интерес применение предложенной методики в других предметных областях, например, показателям с датчиков на промышленных конвейерах или банковским транзакциям. Для подобных задач рассмотренный подход может быть доработан и расширен.

Работа выполнена при поддержке НОШ МГУ «Мозг, когнитивные системы, искусственный интеллект», НИР МГУ 5.1.21, гранта РФФИ No. 20-01-00664.

[1] *Cecotti H., Graser A.* Convolutional neural networks for p300 detection with application to brain-computer interfaces. // IEEE transactions on pattern analysis and machine intelligence, Piscataway: IEEE, 2010. — p. 433–445.

[2] *Blankertz B., Muller K R., Krusienski D J., Schalk G., Wolpaw J R., Schlogl A., Pfurtscheller G., Millan J R., Schroder M., Birbaumer N.* The BCI competition III: Validating alternative approaches to actual BCI problems // IEEE transactions on neural systems and rehabilitation engineering, Piscataway: IEEE, 2006. — p. 153–159.

Application of machine learning in neurophysiology to identify new functional patterns in multivariate time series

Sidorov Leonid[1]⋆ leon.sidorov@gmail.com
Maysuradze Archil[1] maysuradze@cs.msu.ru
[1]Lomonosov Moscow State University

In many subject areas, the data under study have the form of multivariate time series. These can be sensor readings on production lines, stocks on the market or the history of monetary transactions. Sometimes such time series arise as a result of additional processing, for example, a video can be transformed into a bundle of trajectories of specific points. At the same time, both traditional classification problems and specific problems of finding anomalies or trend change points are solved on multivariate time series. Accordingly, a lot of different machine learning (ML) models and architectures have been developed for all these applied areas and tasks.

Nowadays, researchers expect ML models not only to solve applied problems, but also to interpret their results. Models should clearly demonstrate to users the patterns they have found. At the same time, the analysis of multivariate time series in many cases boils down to identifying the so-called *functional patterns* [2], that is, the behavior features of the time series corresponding to some states of the system of interest to researchers. Over the years, experts from various subject areas have already identified a number of functional patterns. The objective of this work is to propose a methodology for automatic search for such patterns in multivariate time series data without having any a priori knowledge about the subject area. The idea of the study is to propose a model that, as a special case, will find already known patterns. By developing such an approach, in the future researchers will be able to use machine learning models in order to quickly find new types of patterns in the source data of multivariate time series.

In this paper, neurophysiology is taken as the subject area. Accordingly, the technique will be demonstrated by identifying one of the most well-known functional patterns in electroencephalograms (EEG), namely *P300 waves*. It is not difficult to see that traditional patterns, such as the P300 wave, are easy to find with sliding filters or the convolution operation [1]. Similar operations have already been implemented in traditional neural network architectures.

We used the data described in [2] for the experiments. The data set "Matrix" was recorded using a character matrix, observing which a person can write words, concentrating his attention on individual characters. During the experiment, the rows and columns of the matrix are highlighted in random order, the reaction to these unexpected highlights for the respondent makes it possible to understand which symbol the person made. Initially, these data were collected in order to understand how well algorithms can predict the target symbol from brain activity data. Accordingly, the [2] also describes existing approaches from the classical MO, which already

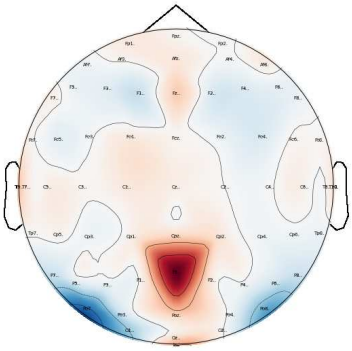


Fig. 1. Electrodes importance map.

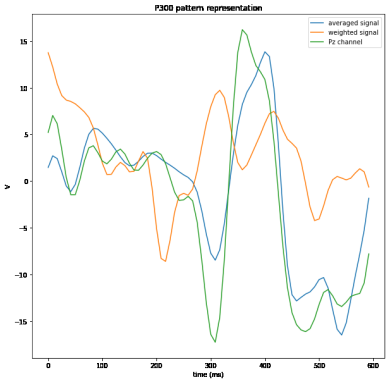


Fig. 2. Optimum channels combinations.

achieve a sufficiently high level of quality on the data set in question. But at that time, researchers had not yet considered what exactly their models were finding.

We focus not on the initial applied problem of character recognition, but on identifying functional patterns and interpreting the obtained results. In order to do this, we used the task transformation already known in the literature from recognition of target characters to recognition of rows and columns of the matrix containing these characters. In this case, the problem can be considered as a binary classification. An EEG fragment corresponding to the highlighting is fed to the input of the ML model, and the output is a binary class label (whether the current highlighting contains the target symbol). It is clear that the P300 wave can be detected on the scale of one second. In the experiments, we worked with fragments of 600 ms.

According to neurophysiologists research, it is expected, but not guaranteed, that highlighting of the target symbol causes a P300 wave on the EEG. Although we know from the training labels when the target symbol is highlighted, the experiment design does not guarantee that a P300 wave will occur during the exposure, however, the chance of this event is extremely high. So, the ML model does not receive information about the presence of the P300 wave directly. The main task of this work is to test the hypothesis that the ML model will still be able to find the functional pattern, which already have been discovered by neurophysiologists.

Note that in neurophysiology, the basic method of determining a pattern is the averaging of time series for each class. This approach requires the involvement of dozens of respondents even for clearly expressed patterns. There are only two respondents in our dataset, so this method was used neither in [2] or in our study.

The proposed neural network architecture applies filters to channels and time intervals on the source data, thereby reveals the importance of certain electrodes on a person's head or moments of time during recording.

In fig. 2 it can be seen that the model independently identified the optimal linear combination of the initial electrodes. Pc7 and Pc8 were taken with a negative weight in order to take into account the negative potential values preceding P300, and the Pz electrode, on the contrary, with a positive weight in order to catch the peak of the wave itself. Thus, the model independently learned to detect the P300 wave. The quality of binary classification and character recognition immediately turned out to be at the level of the best solutions from [2].

Importance map of EEG channels is shown in fig. 1. It can be seen that the model proposed to use a small number of EEG channels.

Thus, the paper proposes a technique that helps researchers to automatically identify functional patterns in multivariate time series based from a series of experiments. It was enough to formalise the original task in terms of machine learning and there was no need for intense exploration of the subject area. The efficiency of the technique has been demonstrated in the field of neurophysiology for data where the expected pattern is already known. For further development of this idea, it is of interest to apply the proposed methodology in other subject areas, for example, data from sensors on production lines or banking transactions. For such tasks, the considered approach can be refined and expanded.

The research is supported by Scientific and educational school of Moscow State University "¡Brain, cognitive systems, artificial intelligence"¿, research work of Moscow State University 5.1.21, RFBR grant No. 20-01-00664.

[1] *Cecotti H., Graser A.* Convolutional neural networks for p300 detection with application to brain-computer interfaces. // IEEE transactions on pattern analysis and machine intelligence, Piscataway: IEEE, 2010. — p. 433–445.

[2] *Blankertz B., Muller K R., Krusienski D J., Schalk G., Wolpaw J R., Schlogl A., Pfurtscheller G., Millan J R., Schroder M., Birbaumer N.* The BCI competition III: Validating alternative approaches to actual BCI problems // IEEE transactions on neural systems and rehabilitation engineering, Piscataway: IEEE, 2006. — p. 153–159.

Восстановление данных во временных рядах метеорологических показателей и CO2 методами математического моделирования

Фадеева Пелагея Александровна[1]⋆ fadeeva.pa20@physics.msu.ru
Безрукова Александра Владимировна[2] aleksandra.bezrukova@effem.com
Газарян Варвара Арамовна[1,3] vagazaryan@fa.ru
Шапкина Наталья Евгеньевна[1,4] neshapkina@mail.ru
Чуличков Алексей Иванович[1,5] achulichkov@gmail.com

[1]Москва, Московский государственный университет имени М.В.Ломоносова, физический факультет
[2]Москва, Leadership Program Specialist, Mars Inct
[3]Москва, Финансовый университет при правительстве РФ
[4]Москва, ИТПЭ РАН
[5]Москва, ИФА им. А.М. Обухова РАН

Метеорологические показания и их изменения оказывают глобальное влияние на функционирование экосистем и развитие экономической деятельности человека, поэтому получение качественных данных является важнейшей задачей современных климатических и экологических исследований.

Данные представляют собой упорядоченную последовательность значений метеорологического показателя $f(t)$, называемую временным рядом. Временной ряд - совокупность величин, представляющая собой значение какого-либо параметра, изменяющегося во времени, при этом каждое значение ряда соответствует значению параметра в определенный момент времени.

Однако измерительные приборы иногда работают со сбоями, которые чаще всего связаны с техническими неисправностями, вследствие чего в рядах существуют пропуски, что не позволяет обеспечить непрерывность регистрации измеряемых параметров. В этой ситуации восстановление рядов данных возможно осуществить на основе математического моделирования.

Рассмотрены варианты восстановления временных рядов, основанных на статистических методах и методах машинного обучения, а также модифицированного варианта, основанного на построении интегрированной модели авторегрессии ARIMA, позволяющих восстанавливать ряды динамики. Статистический и модифицированные методы берут в основу модель ARIMA(p,d,q) - интегрированную модель авторегрессии - скользящего среднего, которая используется для прогнозирования и работы с нестационарными временными рядами и приводит их к стационарному виду путем взятия разности d-го порядка. Модель имеет следующий вид[2]:

$$\Delta^d Y_t = \sum_{i=1}^{p} \alpha_i \Delta^d Y_{t-i} + \sum_{j=1}^{q} \beta_j \varepsilon_{t-j} - \varepsilon_t,$$

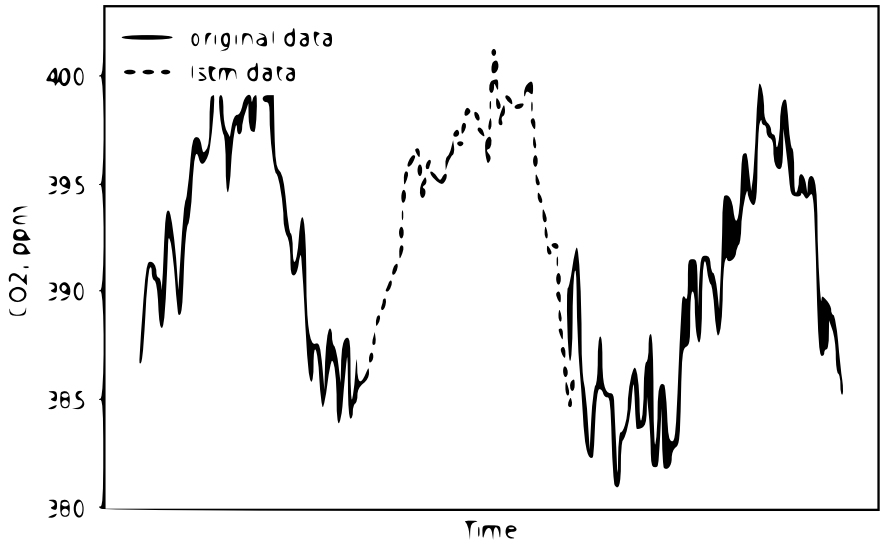


Рис. 1. На графике представлена зависимость исходных данных концентрации CO2 и данных, построенных с помощью модели LSTM от времени.

где Y_t - значение временного ряда в t-ый момент времени, Y_{t-i}- значения временного ряда в предшествующие моменты времени, $\alpha_1, .., \alpha_p$ - коэффициенты авторегрессии, $\beta_1, .., \beta_q$ - параметры модели, ε_t - шумовая компонента, ε_{t-j} - значения шумовой компоненты в предыдущие моменты времени, Δ^d - оператор взятия разности d-го порядка (дифференцирования).

В случае вариантов восстановления на основе методов машинного обучения были использованы сети прямого распространения сигнала, сверточная нейронная сеть (convolutional neural network; CNN), нейронная сеть долгой кратковременной памяти (long short-termmemory; LSTM), двунаправленные LSTM и гибридная методика восстановления временных рядов ARIMA+LSTM.

Полученные с помощью предложенных методов результаты оказались сравнимы по качеству с результатами классических методов прогнозирования [3], а некоторые из моделей способствовали их улучшению.

[1] *Газарян В.А., Курбатова Ю.А., Овсянников Т.А., Шапкина Н.Е. // ВМУ.Серия 3. ФИЗИКА. АСТРОНОМИЯ. 2015. No.5.*

[2] *Bianchi Marco. X-12 - ARIMA (Beta Version 1.1a) // The Economic Journal. Vol. 107. No.. 444. Sep. 1997. p. 1613 - 1620*

[3] *Kurbatova J., Tatarinov F., Molchanov A. et al. // Environ. Res. Lett. 2013. No.8.045028.*

Data recovery in time series of meteorological parameters and CO2 by mathematical modeling

Fadeeva Pelageya[1]⋆ fadeeva.pa20@physics.msu.ru
Bezrukova Alexandra[2] aleksandra.bezrukova@effem.com
Gazaryan Varvara[1,3] vagazaryan@fa.ru
Shapkina Natalia[1,4] neshapkina@mail.ru
Chulichkov Alexey[1,5] achulichkov@gmail.com

[1]Moscow, Lomonosov Moscow State University, Faculty of Physics
[2]Moscow, Leadership Program Specialist, Mars Inc
[3]Moscow, Financial University under the Government of the Russian Federation
[4]Moscow, Institute of Theoretical and Applied Electrodynamics, RAS
[5]Moscow, Atmospheric Physics of the Russian Academy of Sciences

s

Meteorological parameters and their changes have a global impact on the functioning of ecosystems and the development of human economic activity, therefore, obtaining high-quality data is the most important task of modern climate and environmental research.

The data is an ordered sequence of values of a meteorological parameters $f(t)$ called a time series. A time series is a set of values representing the value of a parameter that changes over time, while each value of the series corresponds to the value of the parameter at a certain point in time.

However, measuring instruments sometimes work with failures, which are most often associated with technical malfunctions. As a result, there are gaps in the series, which does not allow for continuous registration of the measured parameters. In this situation, data series recovery can be carried out on the basis of mathematical modeling.

The variants of time series reconstruction based on statistical methods and machine learning methods, as well as a modified version based on the construction of an integrated ARIMA autoregression model, allowing to restore the dynamics series, are considered. Statistical and modified methods are based on the ARIMA(p,d,q) model - an integrated autoregression - moving average model, which is used to predict and work with non-stationary time series and brings them to a stationary form by taking a d-th order difference. The model has the following form[2]:

$$\Delta^d Y_t = \sum_{i=1}^{p} \alpha_i \Delta^d Y_{t-i} + \sum_{j=1}^{q} \beta_j \varepsilon_{t-j} - \varepsilon_t,$$

where Y_t is the value of the time series at the t-th moment in time, Y_{t-i} are the values of the time series at previous moments in time, $\alpha_1, .., \alpha_p$ are autoregression coefficients, $\beta_1, .., \beta_q$ - model parameters, ε_t is the noise component, ε_{t-j} are the values of the noise component at previous points in time, Δ^d is the operator for taking the difference of the d-th order (differentiation).

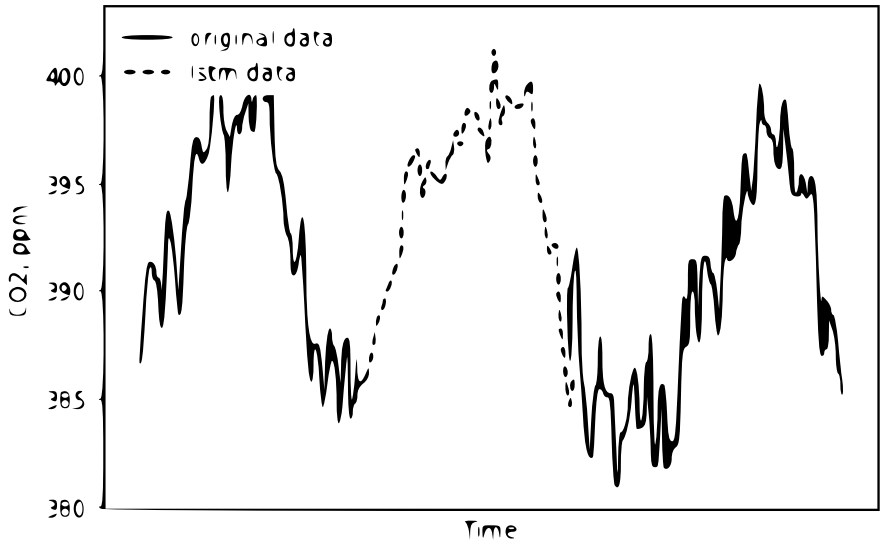


Fig. 1. The graph shows the dependence of the initial CO2 concentration data and the data constructed using the LSTM model on time.

In the case of recovery options based on machine learning methods, direct signal propagation networks, convolutional neural network (CNN), long short-term memory neural network (LSTM), bidirectional LSTM and hybrid ARIMA+LSTM time series recovery technique were used.

The results obtained using the proposed methods turned out to be comparable in quality with the results of classical forecasting methods[3], and some of the models contributed to their improvement.

[1] *V.A. Gazaryan, J.A. Kurbatova, T.A. Ovsyannikov, N.E. Shapkina. Contemporary climate changes in the southwest of the valdai hills: A statistical analysis of the long-term dynamics of the Air Temperature. Moscow University Physics Bulletin 2015. 15. N 5. P. 346-353*

[2] *Bianchi Marco. X-12 - ARIMA (Beta Version 1.1a) // The Economic Journal. Vol. 107. No.. 444. Sep. 1997. p. 1613 - 1620*

[3] *Kurbatova J., Tatarinov F., Molchanov A. et al. // Environ. Res. Lett. 2013. No.8.045028.*

Графовые модели для построения карты связности функциональных групп

Панченко Святослав Константинович[1]⋆ panchenko.sk@phystech.edu
Вареник Наталья Викторовна[1] varenik.nv@phystech.edu
Стрижов Вадим Викторович[2] strijov@phystech.edu
[1]Москва, МФТИ
[2]Москва, ФИЦ Информатика и управление РАН

Решается задача построения модели анализа активности головного мозга, учитывающей пространственную структуру сигнала. Данные об активности мозга представлены в виде многомерных временных рядов, считываемых электродами, расположенными на голове испытуемого одним из универсальных стандартов размещения. Из-за отсутствия регулярного определения окрестности на сферической поверхности мозга классические сверточные нейронные сети не могут быть эффективно применены для учета пространственной информации. Предлагается использовать графовое представление сигнала для выявления взаимосвязей различных областей активности в пространстве и провести нейробиологическую интерпретацию функциональных связей мозга. Исследуются различные методы построения матрицы связности, определяющей графовую структуру для ее последующего использования графовой моделью. Для определения матрицы связности рассматриваются детерминированные методы оценки линейной связи между временными рядами на основе корреляции, спектрального анализа, авторегрессионного подхода и нелинейный метод синхронизации фаз. В качестве модели для решения задачи декодирования предлагается использовать композицию графовой свертки для агрегации пространственной информации и рекуррентного блока для обработки временной последовательности. Также исследуется применимость диффузных уравнений на графах.

[1] *Исаченко Р.В., Стрижов В.В.* Quadratic programming feature selection for multicorrelated signal decoding with partial least squares // Expert Systems with Applications, Выпуск 207, 30 ноября 2022, 117967

Graph models for constructing the connectivity map of functional groups

Panchenko Sviatoslav[1]⋆ panchenko.sk@phystech.edu
Varenik Natalya[1] varenik.nv@phystech.edu
Strijov Vadim[2] strijov@phystech.edu
[1]Moscow, MIPT
[2]Moscow, FRC Computer Sciences RAS

The problem of constructing the model of human brain activity with respect to the spatial structure of the signal is considered. The brain activity is described by multidimensional temporal series extracted from electrodes, arranged on the patient's head in accordance with one of the universal arrangement schemes. Classical convolutional neural networks can not be utilised effectively to account for the spatial information due to absence of a regular definition of a neighbourhood on a spherical surface of the brain. A graph representation of the signal for identifying the interactions between different activity zones in space and for neurobiological interpretation of the functional connections of the brain is proposed. Different methods of constructing the connectivity matrix defining the graph structure are studied. Deterministic methods of estimating linear relationship between time series based on cirrelations, spectral analysis, autoregression approach and nonlinear phase synchronisation are employed for evaluating the connectivity matrix. A composition model of graph convolution for aggregating spatial information and recurrent block for time series processing is proposed for decoding. Applicability of graph neural diffusion equations is also studied.

[1] *Isachenko R.V., Strijov V.V.* Quadratic programming feature selection for multicorrelated signal decoding with partial least squares // Expert Systems with Applications, Volume 207, 30 November 2022, 117967

Метод анализа природных данных на основе вейвлет-фильтрации и нейронных сетей NARX

Мандрикова Оксана Викторовна[1] oksanam1@mail.ru
Полозов Юрий Александрович[1] up_agent@mail.ru
Мандрикова Богдана Сергеевна[1]⋆ 555bs5@mail.ru

[1]Паратунка, Институт космофизических исследований и распространения радиоволн ДВО РАН

Введение. Предложен метод анализа природных данных, основанный на совместном применении операций вейвлет-фильтрации и нейронных сетей NARX. Сети NARX выполняют аппроксимацию временных рядов на основе «моделей нелинейной авторегрессии с экзогенными входами». Процедура вейвлет-фильтрации в данной работе включает комбинацию конструкции кратномасштабного анализа [1] и пороговых функций. Предложен алгоритм вейвлет-фильтрации и способ оценки порогов, основанный на стохастическом подходе. Описаны операции реализации метода.

В работе рассматриваются временные ряды критической частоты ионосферного слоя F2 (foF2). Аномалии в ионосфере влияют на различные аспекты жизни, функционирование космических аппаратов и стабильную работу радиосвязи. Применяемые традиционные методы анализа ионосферных данных не достаточно эффективны для обнаружения ионосферных аномалий [2]. В работе показана эффективность предлагаемого метода для обнаружения ионосферных аномалий в периоды магнитных бурь. Показано, что применение процедуры вейвлет-фильтрации позволяет получить более точную нейросетевую модель NARX временного ряда параметров ионосферы. Также выполнено сравнение метода с непосредственным использованием нейронных сетей NARX, подтвердившее его эффективность.

Описание метода. Применение вейвлет-преобразования и пороговой функции. Для повышения качества процедуры анализа природных данных на основе нейронных сетей, следуя работам [3, 4], применены операции подавления шума. Алгоритм подавления шума следующий:

1. Вейвлет-разложение сигнала $f_0(t)$ на компоненты:

$$f_0(t) = \sum_{j=-1}^{-m} g_j(t) + f_{-m}(t),$$

где $f_{-m}(t) = \sum_k c_{-m,k}\varphi_{-m,k}(t)$ – сглаженная компонента, $c_{-m,k} = \langle f_0, \varphi_{-m,k}\rangle$, $\varphi_{-m,k}(t) = 2^{-m/2}\varphi(2^{-m}t - k)$ – скейлинг-функция, $g_j(t) = \sum_k d_{j,k}\Psi_{j,k}(t)$ – детализирующие компоненты, $d_{j,k} = \langle f_0, \Psi_{j,k}\rangle$, $\Psi_{j,k}(t) = 2^{j/2}\Psi(2^j t - k)$ – вейвлет, j – уровень разложения, для исходного сигнала предполагается уровень разложения $j = 0$.

2. Применение пороговой функции к коэффициентам компонент $g_j(t)$:

$$T(d_{j,k}) = \begin{cases} 0, \text{ если } |d_{j,k}| \leqslant T_j \\ d_{j,k}, \text{ если } |d_{j,k}| > T_j \end{cases},$$

где $T_j = t_{1-\frac{\alpha}{2},N-1}\hat{\sigma}_j$, $t_{\alpha,N}$ – α-квантили распределения Стьюдента, $\hat{\sigma}_j$ – выборочное стандартное отклонение, уровни разложения $j = \overline{-1,-m}$.

3. Вейвлет-восстанавление сигнала:

$$\widetilde{f}_0(t) = \sum_{j,k} T(d_{j,k})\Psi_{j,k}(t) + f_{-m}(t).$$

Определение порогов T_j как $T_j = t_{1-\frac{\alpha}{2},N-1}\hat{\sigma}_j$, где $t_{\alpha,N}$ – α-квантили распределения Стьюдента.

Применение нейронных сетей NARX. Использовались сети NARX с обратными связями [5]. Вход сети обозначен как $\widetilde{f}_0(t)$, а выход $\hat{f}_0(t+1)$. Входные векторы для скрытого слоя являются блоками временных линий задержки. Значение выхода нейронной сети $\hat{f}_0(t+1)$ имеет вид:

$$\hat{f}_0(t+1) = F[\widetilde{f}_0(t), \widetilde{f}_0(t-1), ..., \widetilde{f}_0(t-l_x), \hat{f}_0(t), \hat{f}_0(t-1), ..., \hat{f}_0(t-l_y)], \qquad (1)$$

где $F(\cdot)$ - функция отображения нейронной сети.

В выражении (1) представлен аналитический вид для архитектуры NARX PA. В NARX SPA на вход сети вместо выходов $\widetilde{f}_0(i)$ подаются прошлые значения $\hat{f}_0(i), i = \overline{t, t-l_y}$. Количество линий задержки входа и выхода $l_x = l_y$ позволяет регулировать глубину ретроспективного анализа.

Результаты применения метода для данных ионосферы. На рисунке 1 представлены результаты работы нейронных сетей в период магнитной бури, произошедшей 14 – 19 июля 2017 г.. Анализ данных foF2 показывает нарушение регулярного хода 16 – 18 июля (рис. 1a-e, g-k), что обусловлено возникновением ионосферных возмущений. Аномальные изменения во временном ходе данных foF2 привели к возрастанию ошибок нейронных сетей (рис. 1b-e, h-k).

Анализ результатов показывает, что применение вейвлет-фильтрации позволяет существенно повысить качество работы сетей, вариации ошибок близки к нулю для малой линии задержки сети $l_x = l_y = 2$. В период возникновения продолжительной ионосферной аномалии ошибки сетей существенно возрастают, что позволяет её детектировать. Анализ медианных значений (отмечены пунктиром на рисунке 1 a,g) подтверждает наличие аномалии в ионосфере в период магнитной бури. Сравнение результатов нейронных сетей NARX SPA с медианным методом показывает эффективность предлагаемого метода. Вследствие изменения временного хода данных foF2 во время магнитной бури, в расчетах медианных значений возникли погрешности в период после бури 18 июля 2017 г., которые отсутствуют в нейросетевой модели.

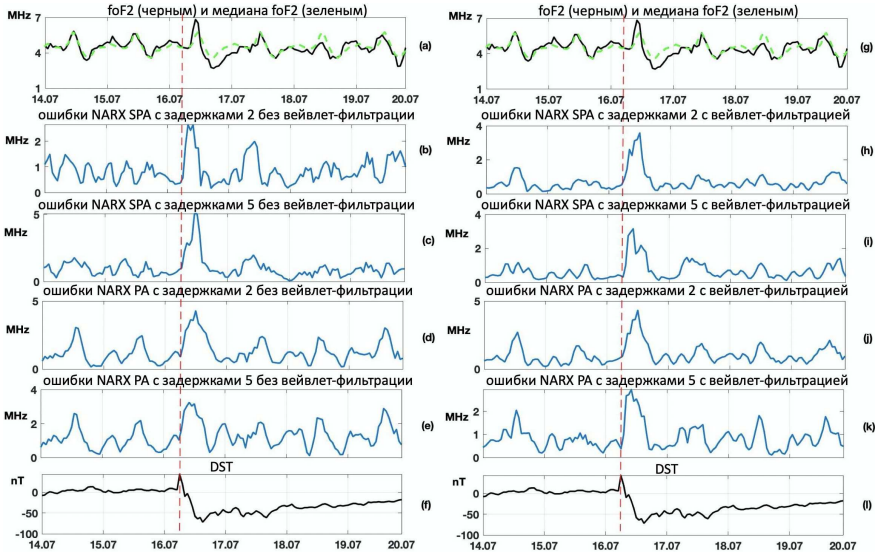


Рис. 1. Результаты обработки данных 14-19 июля, 2017. Красный пунктир – начало магнитной бури.

Заключение. Применение метода показало его эффективность в задаче анализа ионосферных данных и обнаружения аномалий. Операции вейвлет-фильтрации позволяют существенно повысить качество работы нейронных сетей NARX. Ошибки сетей близки к нулю для малой линии задержки $l_x = l_y = 2$.

Работа выполнена в рамках ГЗ «Физические процессы в системе ближнего космоса и геосфер при солнечных и литосферных воздействиях. Регистрационный номер: АААА-А21-121011290003-0».

[1] Mallat, S.G. (1999) *A wavelet tour of signal processing*. San Diego: Academic Press.

[2] Tebabal, A., Radicella, S.M., Nigussie, M., Damtie, B., Nava, B. and Yizengaw, E. (2018). Local TEC modelling and forecasting using neural networks. *Journal of Atmospheric and Solar-Terrestrial Physics*, 172, pp.143–151.

[3] Mandrikova, O. and Mandrikova, B. (2021). Method of wavelet-decomposition to research cosmic ray variations: Application in space weather. *Symmetry*, 13(12), p.2313.

[4] Mandrikova, O. and Mandrikova, B. (2022). Hybrid method for detecting anomalies in cosmic ray variations using neural networks autoencoder. *Symmetry*, 14(4), p.744.

[5] Haykin, S.S. (1999). *Neural networks: a comprehensive foundation*. 2nd ed. Upper Saddle River, N.J: Prentice Hall.

Natural data analysis method based on wavelet filtering and NARX neural networks

Mandrikova Oksana[1] oksanam1@mail.ru
Polozov Yuriy[1] up_agent@mail.ru
Mandrikova Bogdana[1]⋆ 555bs5@mail.ru
[1]Paratunka, Institute of Cosmophysical Research and Radio Wave Propagation FEB RAS

Introduction. A method for analyzing natural data is proposed. It is based on joint application of operations of wavelet filtering and NARX neural networks. NARX networks approximate time series based on the models of nonlinear autoregressive with exogenous inputs. Wavelet filtering procedure includes the combination of the structure of multiscale analysis [1] and threshold functions. A wavelet filtering algorithm and a threshold estimate technique based on a stochastic approach are proposed. Method realization operations are described.

The paper considers time series of the ionospheric layer F2 critical frequency (foF2). Anomalies in the ionosphere affect different aspects of life, space craft functioning and radio communication stable operation. The applied traditional method for ionospheric data analysis are not effective enough to detect ionospheric anomalies [2]. The paper shows the efficiency of the proposed method for detection of ionospheric anomalies during magnetic storm. It was shown that application of wavelet filtering procedure allows us to obtain a more accurate NARX model of ionospheric parameter time series. The method is also compared with direct application of NARX neural networks that confirmed its efficiency.

Method description. Application of wavelet transform and threshold function. To improve the quality of natural data analysis based on neural netowks, according to the papers [3, 4], we applied noise suppression operations. The algorithm for noise suppression is the following:

1. Signal $f_0(t)$ wavelet decomposition into the components:

$$f_0(t) = \sum_{j=-1}^{-m} g_j(t) + f_{-m}(t),$$

where $f_{-m}(t) = \sum_k c_{-m,k}\varphi_{-m,k}(t)$ is the smoothed component, $c_{-m,k} = \langle f_0, \varphi_{-m,k}\rangle$, $\varphi_{-m,k}(t) = 2^{-m/2}\varphi(2^{-m}t-k)$ is the scaling function, $g_j(t) = \sum_k d_{j,k}\Psi_{j,k}(t)$ are detailing components, $d_{j,k} = \langle f_0, \Psi_{j,k}\rangle$, $\Psi_{j,k}(t) = 2^{j/2}\Psi(2^j t-k)$ is the wavelet, j is the decomposition level, the decomposition level $j = 0$ is assumed for the initial signal.

2. Application of the threshold fucntion to the component coefficients $g_j(t)$:

$$T(d_{j,k}) = \begin{cases} 0, \ if \ |d_{j,k}| \leqslant T_j \\ d_{j,k}, \ if \ |d_{j,k}| > T_j \end{cases},$$

where $T_j = t_{1-\frac{\alpha}{2},N-1}\hat{\sigma}_j$, $t_{\alpha,N}$ are α-quantiles of Student's distribution, $\hat{\sigma}_j$ is the sample standard deviation, the decomposition levels $j = \overline{-1,-m}$.

3. Signal wavelet recovery:

$$\widetilde{f}_0(t) = \sum_{j,k} T(d_{j,k})\Psi_{j,k}(t) + f_{-m}(t).$$

Determination of thresholds T_j as $T_j = t_{1-\frac{\alpha}{2},N-1}\hat{\sigma}_j$, where $t_{\alpha,N}$ are the α-quantiles of Student's distribution.

Application of NARX neural networks. We applied feedback NARX networks [5]. Network input is denoted as $\widetilde{f}_0(t)$ and the output as $\hat{f}_0(t+1)$. Input vectors of the hidden layer are the delay time line blocks. The neural network input value $\hat{f}_0(t+1)$ has the form:

$$\hat{f}_0(t+1) = F[\widetilde{f}_0(t), \widetilde{f}_0(t-1), ..., \widetilde{f}_0(t-l_x), \hat{f}_0(t), \hat{f}_0(t-1), ..., \hat{f}_0(t-l_y)], \qquad (1)$$

where $F(\cdot)$ is the neural network display function.

Analytical form for the NARX PA architecture is represented in expression (1). In NARX SPA, previous values $\widetilde{f}_0(i)$ are applied to the input instead of the outputs $\hat{f}_0(i), i = \overline{t, t-l_y}$. The number of input and output delay lines $l_x = l_y$ makes it possible to regulate retrospective analysis depth.

Results of method application for ionospheric data. Fig. 1 shows the results of neural network operation during a magnetic storm, which occurred on July 14-19, 2017.

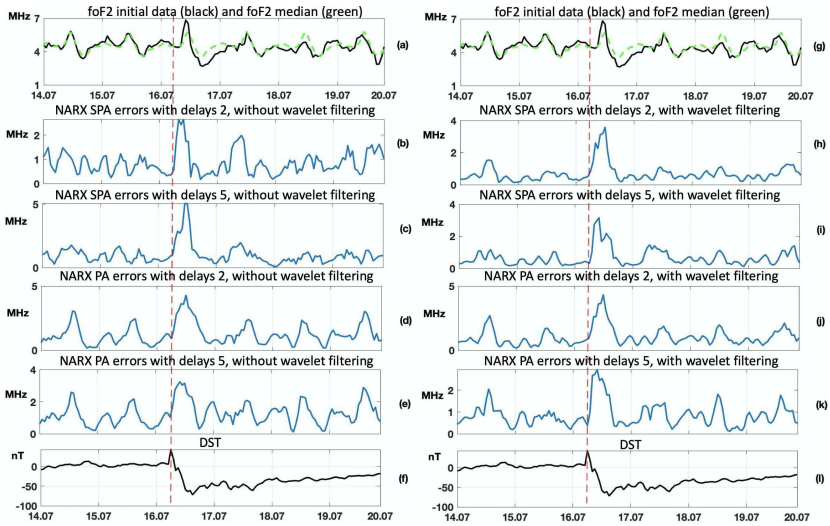


Fig. 1. Result of data processing for the period July 14-19, 2017. Red dashed line indicates the magnetic storm beginning .

Analysis of foF2 data shows regular variation deviations on July 16-18 (Fig. 1a-e, g-k) that is determined by ionospheric disturbance occurrences. Anomalous changes in foF2 data time variation caused increase in neural network errors (Fig. 1b-e, h-k).

Analysis of the results shows that application of the wavelet filtering makes it possible to improve significantly the network operation quality. Error variations are close to zero for the network delay small line $l_x = l_y = 2$. During continuous ionospheric anomaly occurrence, network errors increase significantly that allows us to detect it. Analysis of median values (dashed line in Fig. 1a,g) confirms the anomaly in the ionosphere during the magnetic storm. Comparison of the results of NARX SPA neural networks with the median method shows efficiency of the proposed method. Due to the time series change in foF2 data during the magnetic storm, errors appeared in median value estimates for the period after the storm on July 18, 2017, which are absent in the neural network model.

Conclusion. The application of the method has shown its effectiveness in the task of analyzing ionospheric data and detecting anomalies. Wavelet filtering operations can significantly improve the performance of NARX neural networks. Network errors are close to zero for a small delay line $l_x = l_y = 2$.

The work was carried out as a part of implementation of the State Task AAAA-A21-121011290003-0. The work was carried out by the means of the Common Use Center "North-Eastern Heliogeophysical Center".

[1] Mallat, S.G. (1999) *A wavelet tour of signal processing.* San Diego: Academic Press.

[2] Tebabal, A., Radicella, S.M., Nigussie, M., Damtie, B., Nava, B. and Yizengaw, E. (2018). Local TEC modelling and forecasting using neural networks. *Journal of Atmospheric and Solar-Terrestrial Physics*, 172, pp.143–151.

[3] Mandrikova, O. and Mandrikova, B. (2021). Method of wavelet-decomposition to research cosmic ray variations: Application in space weather. *Symmetry*, 13(12), p.2313.

[4] Mandrikova, O. and Mandrikova, B. (2022). Hybrid method for detecting anomalies in cosmic ray variations using neural networks autoencoder. *Symmetry*, 14(4), p.744.

[5] Haykin, S.S. (1999). *Neural networks: a comprehensive foundation.* 2nd ed. Upper Saddle River, N.J: Prentice Hall.

Применение принципов беспризнакового распознавания образов на основе базисной совокупности объектов в задаче детектирования падений человека

Сурков Егор Эдуардович[1]⋆ eg-su@mail.ru
Середин Олег Сергеевич[1] oseredin@yandex.ru
Копылов Андрей Валериевич[1] and.kopylov@gmail.com
[1]Тула, Тульский государственный университет

В работе рассматривается беспризнаковый подход к распознаванию образов [1] в задаче детектирования падений человека на основе скелетного представления его фигуры. Беспризнаковый подход предполагает представление объектов подходящей мерой их попарного сравнения. Использование функции попарного сходства скелетных моделей между собой предоставляет возможность сокрытия от внешнего наблюдателя представления скелетной модели (скелета) в координатном пространстве. В работе каждая скелетная модель представлена набором вещественных значений, отражающих меру несходства этой модели с некоторым фиксированным набором моделей базисной совокупности [2]. Таким образом, любая скелетная модель может быть представлена фиксированным набором расстояний до каждого объекта базисной совокупности, образуя вектор-столбец. Последовательность векторов формируют матрицу расстояний, которую предложено называть картой активности. В работе для решения задачи детектирования падений предлагается классификация карт активности методами глубокого обучения с применением сверточных нейронных сетей.

В работе определена функция попарного сравнения скелетных моделей, а также выделены аспекты, которые необходимо учитывать при ее вычислении: разный рост людей и смещение скелета в сцене относительно положения камеры (начала координат). Проведено исследование двух методов оценки роста для исключения влияния антропометрических свойств человека на длину сегментов скелетной модели; описаны процедуры устранения вертикального смещения каждой точки скелетной модели и совмещения скелетных моделей в точке начала координат по осям X и Z определена функция попарного несходства, которая для двух скелетных моделей принимает вид евклидовой нормы разности координат соответствующих точек.

В работе проведен анализ баз данных активностей людей, записанных при помощи 3D сенсора и содержащих скелетное описание человека для каждого кадра видеопоследовательности. Такими базами данных являются TST Fall Detection v2 [3] и NTU RGB+D 120 dataset [4]. Также описан набор данных из 136 скелетных моделей, собранный в процессе проведения лабораторных исследований. Объекты этого набора данных были выбраны для представления базисной совокупности.

На основе сформированной базисной совокупности проведены эксперименты и получены карты активностей для нескольких демонстративных видеопоследовательностей. Анализ изображений данных карт активностей выявил необходимость улучшения качества карты активности. Качество карты активности напрямую зависит от порядка расположения строк в ней, который изначально не определен.

Такое предположение позволяет выдвинуть гипотезу о том, что если каждый элемент базисной совокупности является элементом метрического пространства, то кратчайший незамкнутый путь (КНП) расставит элементы в таком порядке, что переход между строками матрицы расстояний на изображении карты активности будет более плавным, границы между отдельными активностями более явными и распознаваемыми, а само изображение более гладким и репрезентативным.

Для упорядочивания объектов базисной совокупности использовались алгоритмы поиска КНП между объектами предложенные в работе [5]. Упорядочивание базисной совокупности способствовало появлению более выраженных и контрастных переходов между активностями.

Проведены экспериментальные исследования применения сверточной нейронной сети ResNET50 для решения двухклассовой задачи классификации на картах активности с использованием бинарной кросс-энтропии в качестве функции потерь. Обучающий набор данных представлен набором карт активностей размера 136x32. Записаные карты активности по всем видеопоследовательностям базы данных TST Fall detection разделены на два класса: «Fall» и «ADL». К классу «Fall» относятся такие карты активностей, которые содержат внутри себя кадры с началом и концом падения или только кадры с падением, остальные карты активностей принадлежат классу «ADL». Карты активностей, в которых кадры с падениями частично присутствуют, исключены из обучающей выборки и используются только при тестировании. Тестирование так же, как и в предыдущей работе [6] выполнялось по процедуре Leave-One-Person-Out [7].

Предложенный подход к решению задачи детектирования падений человека позволил не только повысить оценку точности с 0.936 до 0.947 относительно предыдущих результатов [6], но и превзойти другие алгоритмы детектирования падений, указанные в статье [6].

Работа поддержана грантом ФСИ No.16406ГУ/2021

[1] *Mottl V., Seredin O., Dvoenko S., Kulikowski C., Muchnik I.* Featureless pattern recognition in an imaginary Hilbert space. // In 2002 International Conference on Pattern Recognition (Vol. 2), Quebec City, QC, Canada: IEEE, 2002. — p. 88–91.

[2] *Mottl V., Seredin O., Krasotkina O.* Compactness hypothesis, potential functions, and rectifying linear space in machine learning // Braverman Readings in Machine Learning. Key Ideas from Inception to Current State., Cham: Springer, 2018. — p. 52–102.

[3] *Gasparrini S., Cippitelli E., Gambi E., Spinsante S., Wahslen J., Orhan I., Lindh T.,* Proposal and Experimental Evaluation of Fall Detection Solution Based on Wearable

and Depth Data Fusion // Braverman Readings in Machine Learning. Key Ideas from Inception to Current State., Stockholm: Springer, 2016. — p. 99–108.

[4] *Liu S., Shahroudy E., Perez E., Wang L., Duan L.-Y.., Kot A.C.,* NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding // IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 10, p. 52–102.

[5] *Seredin O., Surkov E., Kopylov A., Dvoenko S.* Multidimensional Data Visualization Based on the Shortest Unclosed Path Search // Artificial Intelligence in Data and Big Data Processing. ICABDE 2021, Ho Chi Minh City: Springer, 2021. — p. 279–299.

[6] *Seredin O., Kopylov A., Surkov E.* The study of skeleton description reduction in the human fall-detection task // Computer Optics, 2021. — p. 279–299.

[7] *Seredin O., Kopylov A., Huang S., Rodionov D.* A Skeleton Features-Based Fall Detection Using Microsoft Kinect v2 with One Class-Classifier Outlier Removal // ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2019. — p. 189–195.

Featureless pattern recognition based on the object basic assembly applying to human fall detection problem

Surkov Egor[1]⋆ eg-su@mail.ru
Seredin Oleg[1] oseredin@yandex.ru
Kopylov Andrei[1] and.kopylov@gmail.com
[1]Tula, Tula State University

This paper considers featureless approach to pattern recognition [1] in the human fall detection problem based on figure skeletal representation. Featureless approach implies object representation by their pairwise comparison measure. The pairwise similarity function of skeletal models make it possible to hide skeletal representation in coordinate space from the external observation. Each skeletal model is represented by a vector of real values (column vector) reflecting the dissimilarity measure of this model with respect to a fixed set of skeletons named basic assembly [2]. Sequence of this vectors forms a distance matrix which we define as "activity map". We propose an activity map classification by CNN deep learning methods to solve the human fall detection problem.

The work determine the pairwise comparison function of skeletal models, and show aspects should be taken into account due to function compute: various people height and skeleton shifts relative to camera position (coordinates origin) in the scene. We research several height estimation methods to exclude an influence of human anthropometric characteristics on the length of the skeleton segments, describe procedures for eliminating the vertical bias of each point of skeletal model and combining skeletal models at the origin points along X and Z axes. And as a result we determine the pairwise function for two skeletal models. It takes a form of Euclidean norm of the coordinate difference between corresponding points.

In this work we performed databases containing activity monitoring data analysis: TST Fall Detection v2 [3] and NTU RGB+D 120 dataset [4]. Also we describe a set of 136 skeletal modes recorded during the laboratory research. This objects constitute the basic assembly.

We provide experiments based on formed basic assembly and receive activity maps for several demonstrative video sequences. Analysis of this activity maps visualization allow to reveal the necessity to improve the quality of activity maps. This quality directly depends on the order of rows in the distance matrix. Initially this order not strictly defined.

This assumption allow to set up the following hypothesis. If each element of the basic assembly is an element of the metric space, then the shortest unclosed (SUP) path will arrange the elements in such an order that: the transition between the boundaries will be smoother; the boundaries between the individual activities themselves will be more explicit and recognizable; the activity map will be smoother and more representative.

We apply the SUP search algorithms proposed in [5] to order the basic assembly objects. The ordering of basic assembly allows to obtain more explicit and contrast transitions between the activities on an activity map.

In the experimental part we apply the CNN ResNETv2 with the binary cross-entropy as loss function to solve the two-class problem on activity maps. The training dataset is presented by activity maps with 136x32 shape. The activity maps recorded on all video sequences from TST Fall detection database was divided into two classes: "ADL" and "Fall". "Fall" labeled activity maps contain frames with start and end of fall activity or only fall activity frames inside itself. The others activity maps is labeled as "ADL". Activity maps where frames with partially presented falls are excluded from the training sample and are used only for tests. Test as in the previous work [6], is performed according to the Leave-One-Person-Out procedure [7].

The proposed approach to solve the human fall detection problem allow to increase the accuracy estimate from 0.936 to 0.947 relative to previous results [6] and also exceed other fall detection algorithms specified in the article [6].

This research is funded by grant of the Innovation Promotion Fund No.16406GU/2021

[1] *Mottl V., Seredin O., Dvoenko S., Kulikowski C., Muchnik I.* Featureless pattern recognition in an imaginary Hilbert space. // In 2002 International Conference on Pattern Recognition (Vol. 2), Quebec City, QC, Canada: IEEE, 2002. — p. 88–91.

[2] *Mottl V., Seredin O., Krasotkina O.* Compactness hypothesis, potential functions, and rectifying linear space in machine learning // Braverman Readings in Machine Learning. Key Ideas from Inception to Current State., Cham: Springer, 2018. — p. 52–102.

[3] *Gasparrini S., Cippitelli E., Gambi E., Spinsante S., Wahslen J., Orhan I., Lindh T.,* Proposal and Experimental Evaluation of Fall Detection Solution Based on Wearable and Depth Data Fusion // Braverman Readings in Machine Learning. Key Ideas from Inception to Current State., Stockholm: Springer, 2016. — p. 99–108.

[4] *Liu S., Shahroudy E., Perez E., Wang L., Duan L.-Y.., Kot A.C.,* NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding // IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 10, p. 52–102.

[5] *Seredin O., Surkov E., Kopylov A., Dvoenko S.* Multidimensional Data Visualization Based on the Shortest Unclosed Path Search // Artificial Intelligence in Data and Big Data Processing. ICABDE 2021, Ho Chi Minh City: Springer, 2021. — p. 279–299.

[6] *Seredin O., Kopylov A., Surkov E.* The study of skeleton description reduction in the human fall-detection task // Computer Optics, 2021. — p. 279–299.

[7] *Seredin O., Kopylov A., Huang S., Rodionov D.* A Skeleton Features-Based Fall Detection Using Microsoft Kinect v2 with One Class-Classifier Outlier Removal // ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2019. — p. 189–195.

Анализ ключевых точек для задачи оценки позы человека

Каприелова Мариам Семеновна[1,2]⋆ kaprielova.ms@phystech.edu
Тихонова Анастасия Дмитриевна[2] tihonova.ad@phystech.edu
Нейчев Радослав Георгиев[2] neychev@phystech.edu
[1]Москва, ФИЦ "Информатика и Управление"РАН
[2]Москва, Московский Физико-Технический Институт

Оценка позы человека – активно исследуемая задача в компьютерном зрении. Интерес к решению данной задачи обусловлен большим количеством областей прикладных применений, таких как безопасность, медицина, виртуальная реальность и другие. Решения задачи оценки позы человека должны быть устойчивы к частичному или полному перекрытию ключевых точек, различиям в телосложении людей и силуэтов одежды и изменению количества людей в кадре. Одним из методов повышения качества и стабильности решений является использование априорной информации[1]. В частности, такой подход применяется и в задачах компьютерного зрения[2]. В качестве априорной информации можно использовать пропорций тела человека. Один из способов учитывать априорные знания – добавление регуляризационного слагаемого в функцию потерь. В качестве априорной может рассматриваться информация о различных пропорциях человека. Предыдущие исследования показали, что использование некоторых пропорций в регуляризационном слагаемом более эффективно, чем использование других. В данной работе производится анализ различных ключевых точек тела человека и исследуется изменение динамики процесса обучения фиксированной модели при использовании различных пропорций в регуляризационном слагаемом. Работа выполнена на датасете Human3.6m[3].

[1] *Vapnik V., Izmailov R.,* Learning using privileged information: similarity control and knowledge transfer, The Journal of Machine Learning Research, vol. 16, pp. 2023–2049.

[2] *Lehrmann A. M., Gehler P. V. and Nowozin S.,* A Non-parametric Bayesian Network Prior of Human Pose, 2013 IEEE International Conference on Computer Vision, 2013, pp. 1281-1288.

[3] *Ionescu C., Papava D., Olaru V. and Sminchisescu C.,* Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 7, pp. 1325-1339.

Keypoint analysis in human pose estimation task

Kaprielova Mariam[1,2]⋆ kaprielova.ms@phystech.edu
Tikhonova Anastasia[2] tihonova.ad@phystech.edu
Neychev Radoslav[2] neychev@phystech.edu
[1]Moscow, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences
[2]Moscow, Moscow Institute of Physics and Technology

Human Pose Estimation is a fundamental problem in computer vision. High interest of pose estimation task is due to its application in a wide range of fields such as security, healthcare, virtual reality etc. Human pose estimation methods should be stable in cases of partly or full keypoints occlusion, various body types, oversize clothes, different number of people in the frame. Most popular approach of improvement models quality assumes using privileged information [1]. It is also useful in computer vision tasks [2]. Human body proportions can be employed as informative prior. One way to consider prior knowledge is to complement loss function with regularization term. Resent research has shown that usage of certain proportions in a regularization term are more efficient in comparison to others. In this work, we analyze different keypoints of human body and investigate the dynamic changes in model leaning process depending on various proportion in regularization term. The research is made using Human3.6m dataset[3].

[1] *Vapnik V., Izmailov R.,* Learning using privileged information: similarity control and knowledge transfer, The Journal of Machine Learning Research, vol. 16, pp. 2023–2049.

[2] *Lehrmann A. M., Gehler P. V. and Nowozin S.,* A Non-parametric Bayesian Network Prior of Human Pose, 2013 IEEE International Conference on Computer Vision, 2013, pp. 1281-1288.

[3] *Ionescu C., Papava D., Olaru V. and Sminchisescu C.,* Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 7, pp. 1325-1339.

Непрерывное представление сигналов головного мозга

Самохина Алина Максимовна[1]⋆ alina.samokhina@phystech.edu
Стрижов Вадим Викторович[1] strijov@phystech.edu

[1]Москва, Московский физико-технический институт (национальный исследовательский университет)

Исследуется проблема построения непрерывного по времени представления сигнала. В рамках задачи декодирования сигнала встречаются ряды с нерегулярной по времени сеткой или низким качеством записи выборки, что затрудняет решение задачи декодирования и применение привычных методов работы с сигналами, например, построение фазовых траекторий сигналов.

Решается задача построения нейроинтерфейса, работающего с многомерными временными рядами как с непрерывными по времени. Для решения поставленной задачи используются управляемые нейронные дифференциальные уравнения. Предполагается, что скрытое состояние управляемого нейронного дифференциального уравнения является непрерывным представлением сигнала.

В вычислительном эксперименте проведено сравнение моделей, использующих непрерывное и дискретное представление сигнала. В качестве реальных данных взята выборка вызванных потенциалов, связанных с событием P300 в задаче определения объекта внимания пользователя. Также показано, что скрытое состояние нейронного управляемого дифференциального уравнения сохраняет информацию о сигнале и снижает зашумленность. В качестве примера приводится сравнение фазовых траекторий сигнала и его представления.

[1] *Самохина А. М., Нейчев Р. Г., Гончаренко В. В., Григорян Р. К., Стрижов В. В.* Модели классификации выборки вызванных потенциалов P300 // Системы и средства информатики, 2022.

[2] *Kidger P., Morrill J., Foster J., Lyons T.* Neural controlled differential equations for irregular time series // Advances in Neural Information Processing Systems, 2020.

Continuous-in-time representation of brain signals

Samokhina Alina[1]⋆ alina.samokhina@phystech.edu
Strijov Vadim[1] strijov@phystech.edu
[1]Moscow, Moscow Institute of Physics and Technology

In this research we investigate the problem of building a continuous-in-time representation of a signal. While solving a signal decoding task we handle some irregular or low-quality time series. These time series make the decoding task and some of research techniques, like investigating signal phase trajectories, difficult.

This work addresses the task of building a brain-computer interface which works with the signal as if it were continuous in time. We suggest using the latent space of neural controlled differential equations as a continuous-in-time representation of a signal.

In this work, we compared both discrete and continuous time models on the binary classification task using original dataset of P300 potentials. We also obtained the continuous-in-time representation of a signal. This representation saves all information from the original signal and reduces its noize. As an example of representation usage we compared phase trajectories of an original signal and its representation.

[1] *Samokhina A. M., Neychev R. G., Goncharenko V. V., Grigoryan R. K., Strijov V. V.* Classification models for evoked P300 potentials dataset // Systems and Means of Informatics, 2022.

[2] *Kidger P., Morrill J., Foster J., Lyons T.* Neural controlled differential equations for irregular time series // Advances in Neural Information Processing Systems, 2020.

Референтный текстовый корпус и оценивание близости коротких текстов смысловому эталону

Михайлов Дмитрий Владимирович[1]⋆ mdv74@list.ru
Емельянов Геннадий Мартинович[1] Gennady.Emelyanov@novsu.ru
[1]Великий Новгород, Россия, НовГУ

На практике достаточно часто требуется сформировать подборку публикаций по заданной теме. Помимо научных исследований, такая задача актуальна при подготовке электронного учебного материала. При этом наибольшую значимость, как правило, имеют публикации, для которых при максимально полном раскрытии интересующей пользователя темы характерен максимум среднего числа наиболее значимых терминов в расчёте на одно простое распространённое предложение (фразу) при минимуме его длины (в словах). Содержательно это соответствует максимально краткому и ёмкому изложению, отвечающему эталонному варианту передачи смысла средствами заданного естественного языка. Данное требование можно переформулировать следующим образом: тексты в составе подборки (коллекции) должны быть максимально релевантны заданной предметной области с точки зрения эксперта как по лексическому составу, так и по внутритекстовым связям (синтаксическим, семантическим и т. п.). Саму текстовую коллекцию при этом называют представительной (референтной). Если тексты внутри коллекции размечены по определённым правилам (синтаксически, с применением базы ролевых зависимостей и т. п.), то мы имеем дело с референтным корпусом, [1].

Вместе с тем, актуальной остаётся задача минимизации ручного труда эксперта при формировании подобного рода коллекций. Наиболее целесообразно здесь использование экспертом коротких текстов, которые сопоставлялись бы по словарному составу и (возможно) по связям между словами с документами, добавляемыми в референтную коллекцию. В роли таких текстов вполне могут быть аннотации научных статей либо другие тексты, резюмирующие значимые (с точки зрения эксперта) факты заданной предметной области. При этом имеем задачу, обратную абстрактивной суммаризации: найти текст, в котором описанные в аннотации (коллекции аннотаций) общие идеи отражены наиболее полно. В настоящей работе указанная задача решается на основе долей ненулевых значений частоты слова в анализируемом документе, рассчитываемых по фразам аннотаций. Будем для каждой фразы в составе каждой аннотации вычислять долю ненулевых значений TF-меры (отношения числа вхождений слова к общему числу слов документа, *term frequency*) для входящих во фразу слов относительно анализируемого документа. Одна фраза здесь соответствует простому распространённому предложению (в терминологии теории языка как преобразователя «Смысл⇔Текст»). Поскольку в реальных аннотациях доля сложных предложений минимальна, то применять данный термин к предложениям в составе аннотаций вполне допустимо. В целях максимального отражения содержа-

ния статей их аннотации будем рассматривать вместе с заголовками. При этом допускается, что одна и та же фраза может встречаться в нескольких аннотациях коллекции (например, если это статьи одного и того же автора). В любом случае каждая фраза принимается к рассмотрению только один раз.

В задаче оценки когнитивной сложности текста [1] предполагалось, что 95% токенов в референтном корпусе на каждом из языковых уровней не превышают своей фиксированной частоты. В нашей же задаче речь идёт о минимально необходимой представленности слов (терминов) из аннотаций в документе, анализируемом на предмет включения в референтный корпус. Логично предположить, что здесь следует рассматривать 5-й перцентиль частотной характеристики слова относительно заданного документа. Основное требование к самой частотной характеристике — независимость от числа слов документа. Для интуитивной наглядности далее мы ограничимся рассмотрением лексического уровня, для других уровней языка (фонетического, морфологического, синтаксического, дискурсивного) рассуждения проводятся аналогично.

Пусть d — документ-кандидат на включение в референтный корпус. Для каждого слова w в составе каждой фразы Ts каждой аннотации из сформированной экспертом коллекции вычисляется значение TF-меры относительно документа d, $\mathrm{tf}\,(w, d)$. При этом доля ненулевых значений TF по отдельной фразе Ts формально определяется как

$$c(Ts, d) = \frac{\big|w\colon\ (w \in Ts) \wedge (\mathrm{tf}\,(w, d) > 0)\big|}{\big|w\colon w \in Ts\big|}. \tag{1}$$

Обозначим 5%-й квантиль эмпирического распределения величины (1) по документу d для заданной коллекции аннотаций $\mathbb{T}\mathrm{s}$ как $C_5\,(\mathbb{T}\mathrm{s}, d)$. Сама $\mathbb{T}\mathrm{s}$ при этом есть объединение множеств фраз для отдельных аннотаций. Отсортируем документы-кандидаты на включение в референтный корпус по убыванию $C_5\,(\mathbb{T}\mathrm{s}, d)$. Пусть $d_{\max}$ — документ с максимальным по множеству документов-кандидатов D значением C_5 для фраз аннотаций из $\mathbb{T}\mathrm{s}$. Введём для каждого документа $d \in D$ вектор значений квантилей

$$\overline{V}\,(\mathbb{T}\mathrm{s}, d) = \Big(C_\gamma\,(\mathbb{T}\mathrm{s}, d)\Big)_{\gamma\in[5,10,20,25,30,40,50,60,70,75,80,90,95]}, \tag{2}$$

куда помимо упомянутых выше 5-го и 95-го перцентилей войдут децили, а также первый и третий квартили. Пусть $\overline{V}\,(\mathbb{T}\mathrm{s}, d_{\max})$ — вектор вида (2) для документа $d_{\max}$. Обозначим последовательность указанных векторов по коллекции $\mathbb{T}\mathrm{s}$ для документов $d_j \in D\colon d_j \neq d_{\max}$, отсортированную по убыванию Евклидова расстояния до $\overline{V}\,(\mathbb{T}\mathrm{s}, d_{\max})$, как $\mathbb{V}(\mathbb{T}\mathrm{s}, D)$. Разобьём последовательность $\mathbb{V}(\mathbb{T}\mathrm{s}, D)$ на кластеры $H_1, \ldots, H_r$ с применением алгоритма, содержательно близкого алгоритмам класса FOREL. Далее в настоящей работе применительно к разбиению последовательности на кластеры будем подразумевать именно этот алгоритм. При этом кластер H_r по определению будет отвечать документам с наименьшим расстоянием до документа $d_{\max}$.

Утверждение 1. *Наибольшую значимость для референтного корпуса, оцениваемую величиной C_5, имеют $d \in D$, отнесённые к кластеру H_r, плюс сам $d_{\max}$.*

Пусть $\mathbb{T}\mathrm{s}_i \subset \mathbb{T}\mathrm{s}$ — множество фраз i-й аннотации в составе $\mathbb{T}\mathrm{s}$, $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right)$ — 5%-й квантиль эмпирического распределения величины (1) по документу $d_{\max}$ относительно фраз этой аннотации. Обозначим документ с максимальным по множеству D значением величины C_5 для фраз в составе $\mathbb{T}\mathrm{s}_i$ как $d_{\max(i)}$.

Утверждение 2. *По значимости для вычисления $C_5\left(\mathbb{T}\mathrm{s}, d_{\max}\right)$ среди аннотаций из $\mathbb{T}\mathrm{s}$ можно выделить следующие пять групп:*

- *аннотации, где $d_{\max} = d_{\max(i)}$, а $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right) > C_5\left(\mathbb{T}\mathrm{s}, d_{\max}\right)$;*
- *аннотации, где $d_{\max} \neq d_{\max(i)}$, но $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right) > C_5\left(\mathbb{T}\mathrm{s}, d_{\max}\right)$, а $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max(i)}\right)$ и $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right)$ могут быть отнесены к одному кластеру;*
- *аннотации, где $d_{\max} \neq d_{\max(i)}$, но $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right) > C_5\left(\mathbb{T}\mathrm{s}, d_{\max}\right)$, а $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max(i)}\right)$ и $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right)$ лежат в разных кластерах;*
- *аннотации, где $d_{\max} \neq d_{\max(i)}$ и $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right) < C_5\left(\mathbb{T}\mathrm{s}, d_{\max}\right)$, а $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max(i)}\right)$ и $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right)$ могут быть отнесены к одному кластеру;*
- *аннотации, где $d_{\max} \neq d_{\max(i)}$ и $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right) < C_5\left(\mathbb{T}\mathrm{s}, d_{\max}\right)$, а $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max(i)}\right)$ и $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right)$ лежат в разных кластерах.*

При этом наибольшую точность поиска значимых для референтного корпуса документов дают аннотации из первой (содержательно — максимально близкие смысловому эталону), второй и третьей группы. В совокупности предложенные решения (результаты экспериментов представлены в [2]) дают минимум пятикратное сокращение числа документов из минимально релевантных формируемому референтному корпусу. Результаты экспериментов в [2] здесь показали полноту поиска (отношение числа документов, отвечающих условию *Утверждения 1* и признанных экспертом значимыми, к общему числу документов $d \in D$ из признанных значимыми), приблизительно равную 5/6. В целях повышения полноты поиска описанную классификацию документов следует провести независимо по нескольким коллекциям аннотаций статей близких тематических направлений. При этом более высокую оценку значимости получают те $d \in D$, которые при большем числе фраз будут иметь большее среднее число наиболее важных терминов в расчёте на одну фразу при минимуме её длины. Содержательно это отвечает более краткому и ёмкому изложению — правилу «хорошего тона» изданий по физико-математическим и техническим наукам.

Работа поддержана грантом РФФИ No. 19-01-00006.

[1] *Eremeev M., Vorontsov K.* Lexical Quantile-Based Text Complexity Measure // Proceedings of Recent Advances in Natural Language Processing, 2019.

[2] *Михайлов Д. В., Емельянов Г. М.* Формирование референтного текстового корпуса для оценивания близости тематических текстов смысловому эталону // Pattern Recognition and Image Analysis, 2022. Т. 32, №4.

Reference text corpus and estimating the closeness of short texts to the semantic standard

Mikhaylov Dmitry[1]⋆ mdv74@list.ru
Emelyanov Gennady[1] Gennady.Emelyanov@novsu.ru
[1]Russia, Veliky Novgorod, Yaroslav-the-Wise Novgorod State University

In practice, it is quite often required to collect publications on some topic. In addition to research work, such problem is actual in the preparation of e-learning material. The most significant here, as a rule, are publications, for which at maximal disclosure of the topic of the interest to the end user, a maximum of an average number of most significant terms per a one simple spread sentence (i. e. phrase) at the minimum of its length measured in words, is typical. Substantially, this corresponds to the most brief, but succinct narration, that satisfy to the standard variant of sense transfer in a given natural language. This requirement can be reformulated as follows: texts in a collection should be relevant as much as possible to a given topical area from the point of expert view both in vocabulary, and in internal text relations (syntactic, semantic, etc.). The text collection itself here is called representative (or reference). In a case when collection texts are labeled according to some determined rules (syntactically, by application of a base of role dependencies, etc.) we have a reference corpus, [1].

However, the problem of minimization of the handwork of expert in the formation of such collections is actual here. The most advisable here is to use short texts by an expert that would be compared in terms of vocabulary and (possibly) relationships between words with documents being added to the reference collection. In the role of such texts abstracts of scientific articles or other texts that are resume significant facts of a given topical area can entirely be. Here we have the problem inverse to the well-known problem of abstractive summarization: to find a text, in which general ideas described in the abstract (or in the collection of abstracts) are reflected the most fully. In current paper the mentioned problem is solved by using of shares of non-zero values of word frequency in an analyzed document, calculated by phrases of abstracts. Let's calculate for each phrase in each abstract the share of non-zero values of the *term frequency* (TF-measure, the ratio of the number of occurrences of a word in a document to the total number of document words) for phrase words relatively to the analyzed document. Here, one phrase corresponds to a simple common (spread) sentence (in the terminology of "Meaning⇔Text" linguistic theory). Since the share of complex sentences in real abstracts is minimal, it is quite acceptable to apply this term to sentences in abstracts. In order to reflect the content of articles as much as possible, their abstracts will be considered together with the titles. At the same time, it is permitted that the same phrase can occur in several abstracts of the collection (for example, if these are articles by the same author). In any case, each phrase is accepted for consideration only once.

To solve the problem of measuring the cognitive complexity of text [1] the assumption was make about that 95% of tokens in a reference corpus at each of language level did not exceed their fixed occurrence frequency. In our task, we are dealing with the minimum necessary representation level of words (terms) from abstracts in a document being analyzed for adding to the target reference corpus. It's reasonable to assume that the 5th percentile of frequency characteristic of word relatively to the given document here we should consider. The main requirement to the used frequency characteristic itself is independence from the number of words in the document. For intuitive clarity, we will restrict ourselves to the lexical level in further reasoning; for other levels of the natural language (phonetic, morphological, syntactic, discursive levels), reasoning is carried out similarly.

Let document d be a candidate for adding to the reference corpus. We'll designate further the set of such documents as D. For each word w of each phrase Ts of each abstract from the collection formed by an expert the value of TF-measure is calculated relative to the document d, $\mathrm{tf}\,(w, d)$. Herewith the share of non-zero TF values for the separate phrase Ts is formally defined as

$$c(Ts, d) = \frac{\left|w\colon\ (w \in Ts) \wedge (\mathrm{tf}\,(w, d) > 0)\right|}{\left|w\colon w \in Ts\right|}. \tag{1}$$

Let's designate as $C_5\,(\mathbb{T}\mathrm{s}, d)$ the 5th percentile of the empirical distribution of estimation (1) value concerning the document d for the given collection of abstracts $\mathbb{T}\mathrm{s}$. We'll associate $\mathbb{T}\mathrm{s}$ with the combining of sets of phrases for separate abstracts. Let's sort documents that are candidates for adding to the reference corpus, by decreasing of $C_5\,(\mathbb{T}\mathrm{s}, d)$. Let $d_{\max}$ be a document with the maximal value of C_5 among the documents $d \in D$ for phrases of abstracts from $\mathbb{T}\mathrm{s}$. Let's enter into consideration for each $d \in D$ the vector of quantiles values

$$\overline{V}\,(\mathbb{T}\mathrm{s}, d) = \Big(C_\gamma\,(\mathbb{T}\mathrm{s}, d)\Big)_{\gamma \in [5,10,20,25,30,40,50,60,70,75,80,90,95]}, \tag{2}$$

into which deciles together with the first and third quartiles will be included in addition to the above-mentioned 5th and 95th percentiles. Let $\overline{V}\,(\mathbb{T}\mathrm{s}, d_{\max})$ be a vector of the form (2) for the document $d_{\max}$. Let's designate the sequence of mentioned vectors obtained for the collection $\mathbb{T}\mathrm{s}$ relatively to documents $d_j \in D\colon d_j \neq d_{\max}$ and sorted by descending the Euclidean distance to $\overline{V}\,(\mathbb{T}\mathrm{s}, d_{\max})$, as $\mathbb{V}(\mathbb{T}\mathrm{s}, D)$. Let us split the sequence $\mathbb{V}(\mathbb{T}\mathrm{s}, D)$ into clusters $H_1, \ldots, H_r$ using an algorithm close in meaning to the FOREL class of algorithms. Further in current paper, relatively to the partition of some sequence into clusters, we mean namely this algorithm. In this case, the cluster H_r will by definition correspond to documents with the shortest distance to the document $d_{\max}$.

Statement 1. *The highest significance for the reference corpus, which is estimated by the value of C_5, will be processed by documents $d \in D$ related to the cluster H_r, and the document $d_{\max}$ itself.*

Let $\mathbb{T}\mathrm{s}_i \subset \mathbb{T}\mathrm{s}$ be the set of phrases of the ith abstract from $\mathbb{T}\mathrm{s}$, and $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right)$ be the 5th percentile of the empirical distribution of the estimation (1) value concerning the document $d_{\max}$ relatively to phrases of this abstract. Let's designate the document with the maximal value of C_5 among the documents $d \in D$ for phrases of $\mathbb{T}\mathrm{s}_i$, as $d_{\max(i)}$.

Statement 2. *According to the degree of significance for precise calculating the value of* $C_5\left(\mathbb{T}\mathrm{s}, d_{\max}\right)$ *among abstracts related to* $\mathbb{T}\mathrm{s}$ *the following five groups can be distinguished:*

— *abstracts, where* $d_{\max} = d_{\max(i)}$ *and* $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right) > C_5\left(\mathbb{T}\mathrm{s}, d_{\max}\right)$*;*
— *abstracts, where* $d_{\max} \neq d_{\max(i)}$*, but* $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right) > C_5\left(\mathbb{T}\mathrm{s}, d_{\max}\right)$*, moreover, values of* $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max(i)}\right)$ *and* $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right)$ *can be assigned to the same cluster;*
— *abstracts, where* $d_{\max} \neq d_{\max(i)}$*, but* $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right) > C_5\left(\mathbb{T}\mathrm{s}, d_{\max}\right)$*, moreover, values of* $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max(i)}\right)$ *and* $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right)$ *belong to different clusters;*
— *abstracts, where* $d_{\max} \neq d_{\max(i)}$ *and* $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right) < C_5\left(\mathbb{T}\mathrm{s}, d_{\max}\right)$*, moreover, values of* $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max(i)}\right)$ *and* $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right)$ *can be assigned to the same cluster;*
— *abstracts, where* $d_{\max} \neq d_{\max(i)}$ *and* $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right) < C_5\left(\mathbb{T}\mathrm{s}, d_{\max}\right)$*, moreover, values of* $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max(i)}\right)$ *and* $C_5\left(\mathbb{T}\mathrm{s}_i, d_{\max}\right)$ *belong to different clusters.*

In this case, the highest search precision for documents that are important for the reference corpus is reached with abstracts of the first, second and third groups. In meaning, first group abstracts are closest to the semantic standard. As the recall estimation we used the ratio of the number of documents that meet the condition of *Statement 1* and recognized by the expert as significant for the reference corpus, to the total number of documents $d \in D$ from recognized as significant by the expert. Experimental results presented in [2] demonstrate here the recall value approximately equal to 5/6. Also experiments show that the proposed solutions jointly gives at least a fivefold reduction in the number of documents of those that are minimally relevant to the reference corpus being formed. In order to improving the search's recall the described classification of documents must be carried out independently for several collections of abstracts of articles in similar topical areas. It should be noted that a higher significance estimation is obtained by those documents $d \in D$ that will have a larger average number of the most significant terms per phrase at a minimum phrase length and a greater number of phrases themselves. Substantially, this corresponds to the most brief, but succinct narration, that satisfy to the "good manner" rule of publications in Physics, Mathematics and Technical Sciences.

This research is funded by RFBR, grant 19-01-00006.

[1] *Eremeev M., Vorontsov K.* 2019. Lexical Quantile-Based Text Complexity Measure. *Proceedings of Recent Advances in Natural Language Processing.*

[2] *Mikhaylov D. V., Emelyanov G. M.* 2022. Reference corpus formation for estimating the closeness of topical texts to the semantic standard. *Pattern Recognition and Image Analysis.* 32(4).

Использование генеративных моделей в вопросно-ответных системах

Сурков Вячеслав Олегович surokpro2@gmail.com
Евсеев Дмитрий Андреевич dmitrij.euseew@yandex.ru
Москва, МФТИ (НИУ)

Генерация ответов на вопросы — неотъемлемая часть работы множества диалоговых систем и чат-ботов. Для того, чтобы речь системы была живой и яркой, она должна производить развернутые ответы на вопросы.

К сожалению, на данный момент вопросно-ответные наборы обучающих данных на русском языке (например, SberQUAD), содержат либо короткие ответы, либо подстроки текста, содержащие ответы, что усложняет обучение модели. В то же время, сбор нового датасета —- это затратная активность по времени и финансам. Также есть проблема оценки качества модели — с одной стороны, она должна генерировать длинный грамматически корректный текст, а с другой стороны, в этом тексте должен содержаться ответ на вопрос, поэтому стандартные метрики, оценивающие сходство сгенерированного ответа с эталоном не дадут полной картины.

В данной работе предлагается алгоритм построения длинного ответа, использующий деревья синтаксического разбора вопроса и короткого ответа. С помощью этого алгоритма были сгенерированы развернутые ответы для примеров SberQuad, на модифицированном датасете были обучены несколько вариантов T5 — моделей архитектуры Трансформер, использующих архитектуру encoder-decoder.

Были исследованы RuT5 (модель от Сбербанка, предобученная на русскоязычном корпусе текстов) и mT5 (мультиязычная модель от Google). Также было изучено влияние трансферного обучения — обучения на обширном англоязычном наборе данных с развернутыми ответами перед обучением на русскоязычном — на качество генерируемых ответов. В качестве англоязычного датасета использовался MS-MARCO — вопросно-ответный набор основанный на запросах в систему поиска Bing.

Результаты экспериментов показали, что трансферное обучение даёт значительное улучшение метрик при few-shot дообучении модели mT5. Это означает, что обученная на обширном англоязычном наборе (MS-MARCO) модель быстрее (то есть, за меньшее количество примеров из тренировочной выборки) подстраивается под русскоязычную задачу, чем необученная. Также, при сравнении такой модели с русскоязычной RuT5 оказалось, что они имеют схожее качество генерируемого текста (по метрике sacrebleu), однако mT5 несколько точнее находит в правильные ответы в тексте.

Utilizing generative models in question-and-answer systems

Surkov Vyacheslav⋆ surokpro2@gmail.com
Evseev Dmitry dmitrij.euseew@yandex.ru
Moscow, MIPT

Generative question answering —an integral part of the operation of many dialog systems and chatbots. In order for the system's speech to be vivid, it must produce detailed answers to questions.

At the moment, question-answer training datasets in Russian (e.g., SberQUAD), contain either short answers or substrings of text containing answers; this makes model training difficult. However, collecting a new dataset —- is a time-consuming and costly activity. There is also the problem of evaluating the quality of the model — it must generate a long grammatically correct text, and this text must contain the answer to the question. Therefore, standard metrics evaluating the similarity of the generated answer to the ground truth are not completely informative.

This paper proposes an algorithm for generating a long answer using the parsing trees of the question and the short answer. Using this algorithm, we generated extended answers for SberQuad, and fine-tuned several variants of T5 — transformer models using encoder-decoder architecture.

We investigated RuT5 (Russian-language model from Sberbank) and mT5 (multilingual model from Google). The effect of transfer learning — training on the extensive English-language dataset with expanded answers before fine-tuning on the Russian-language one — on the quality of the generated answers was also studied. MS-MARCO was used as an English-language dataset – a question-answer set based on queries to the Bing search engine.

Experimental results have shown that transfer learning gives a significant improvement in metrics on mT5 few-shot training. This means that the model trained on the extensive English-language set adjusts to the problem on Russian language than untrained one. Also, comparing this model with the Russian-language RuT5, they have similar quality of the generated text (according to sacrebleu metric), but mT5 is somewhat more accurate in finding the correct answers in the text.

Анализ на внутренние заимствования как способ отбора высокооригинальных документов

Сафин Камиль Фанисович[1]⋆ kamil.safin@phystech.edu
Чехович Юрий Викторович[2,3] chehovich@ap-team.ru

[1]Москва, Московский физико-технический институт (национальный исследовательский университет)
[2]Москва, Антиплагиат
[3]Москва, Федеральный исследовательский центр «Информатика и управление» РАН

Рассматривается задача обнаружения некорретных текстовых заимствований. Поиск заимствований в текстовых документах является сложной, но в то же время востребованной задачей, особенно в академической и студенческой средах [1].

Можно выделить два глобальных подхода к задаче поиска некорректных заимствований в тексте: поиск внешних заимствований (external plagiarism detection) и поиск внутренних заимствований (intrinsic plagiarism detection). Поиск внешних заимствований представляет собой поиск по внешней коллекции документов, которые могли быть использованы в качестве источника заимствования. Поиск внутренних заимствований же, наоборот, не использует внешнюю коллекцию потенциальных источников, а анализирует текст изолированно.

Методы поиска заимствований по внешней коллекции являются точными, так как обнаруживают точные совпадения в анализируемом тексте и в тексте источнике. Однако они являются ресурсоемкими, так как размеры коллекций для поиска как правило очень большие. Методы поиска внутренних заимствований, напротив, являются гораздо менее точными, так как выявляют нерегулярности в стиле письма автора, которые не обязательно могут оказаться заимствованиями.

Методы поиска текстовых заимствований, используемые в промышленных системах (таких как «Антиплагиат» [2]) постоянно совершенствуются, так как сами методы заимствований тоже усложняются. Например, появились методы обнаружения перефразирований, поиска переводных заимствований [3] или обнаружения скрытых заимствований [4]. Развитие таких методов ведет к увеличению сложности и объемов вычислительных ресурсов, необходимых для осуществления проверок. При этом пратика показывает, что применение всего спектра методов обнаружения некорректных заимствований далеко не всегда оправданно.

Предлагается использовать подход по поиску внутренних заимствований. Как было сказано, в качестве самостоятельного инструмента, такой подход имеет очень низкое качество работы. Но его можно использовать как грубый фильтр перед более точной проверкой, который будет отсеивать документы, которым не нужна детальная экспертиза.

В работе рассматривается алгоритм фильтрации высокооригинальных текстов [5], основанный на анализе частот употребления символьных и словесных n-грам. На основе данного алгоритма реализован программный комплекс, предназначенный для внедрения в систему выявления некорректных текстовых заимствований с использованием внешних текстовых коллекций. Снижение нагрузки происходит путём отбора документов, не требующих детальной проверки. Документы же, требующие детальной проверки, проходят полную проверку.

Описывается вычислительный эксперимент, демонстрирующий работоспособность данного метода, а также объем сэкономленных вычислительных ресурсов. Показывается, что на размеченных и синтетических данных подход позволяет сократить поток документов, которым не требуется детальная проверка, почти на треть. При этом важно подчеркнуть, что это не только ускоряет время обработки отдельных документов, а позволяет использовать вычислительные ресурсы более целенаправленно, то есть детально анализировать именно те документы, которые нуждаются в такой проверке.

[1] *Никитов А. В., Орчаков О. А, Чехович Ю.В.* Плагиат в работах студентов и аспирантов: проблема и методы противодействия // Университетское управление: практика и анализ, 2012. — (5):61-68.

[2] *Журавлев Ю. И, Рудаков К. В., Инякин А. С., Кирсанов А. А., Лисица А. В., Никитов Г. В., Песков Н. В., Романов М. Ю., Яминов Р. И., Чехович Ю.В.* Система распознавания интеллектуальных заимствований «Антиплагиат» // Математические методы распознавания образов, 2005. — 12(1):329-332.

[3] *Кузнецова Р. В., Бахтеев О. Ю., Чехович Ю.В.* Методы обнаружения переводных заимствований в больших текстовых коллекциях // Информ. и её примен., 2021. — 15:1 (2021), 30–41.

[4] *Chekhovich Y. V., Khazov A. V.* Analysis of duplicated publications in Russian journals // Journal of Informetrics, 2022. — 16(1):101246.

[5] *Сафин К. Ф., Чехович Ю.В.* О комбинированном алгоритме обнаружения заимствований в текстовых документах // Труды Института системного программирования РАН, 2022. — 34(1):151-160.

Intrinsic plagiarism analysis for highly original documents selection

Safin Kamil[1]⋆ kamil.safin@phystech.edu
Chekhovich Yury[2,3] chehovich@ap-team.ru
[1]Moscow, Moscow Institute of Physics and Technology
[2]Moscow, Antiplagiat
[3]Moscow, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences

Text plagiarism detection is a complex, but at the same time a demanded task, especially in academic and student [1] areas.

There are two global approaches to the task of plagiarism detection in a text: external plagiarism detection and intrinsic plagiarism detection. External plagiarism detection is a search through an external collection of documents that may have been used as a source of borrowing. Intrinsic plagiarism detection, on the other hand, does not use an external collection of potential sources, but analyzes the text by itself.

External plagiarism detection search methods are accurate because they detect exact matches in the analyzed text and in the source text. However, they are resource-intensive, since the size of collections to search is usually very large. Methods of searching for internal plagiarism, on the contrary, are less accurate, as they reveal irregularities in the author's writing style, which may not necessarily turn out to be plagiarism.

The methods of searching for textual borrowings used in industrial systems (such as ≪Antiplagiat≫) are constantly improving, as the borrowing methods themselves are also becoming more complex. For example, methods of detecting paraphrases, searching for translated borrowings [2] or detecting hidden borrowings [3] have appeared. The development of such methods leads to an increase in the complexity and the amount of computational resources needed to perform the checks. At the same time, practice shows that the use of the full range of methods for the detection of incorrect borrowings is far from always justified.

For this purpose, we propose to use the approach to search for internal plagiarism. As mentioned, as a standalone tool, this approach has a very low quality of work. But it can be used as a filter before a more accurate check, which will filter out documents that do not need a detailed examination.

In this paper we consider an algorithm for filtering highly-original texts [4], based on the analysis of char and word n-gram frequencies. On the basis of this algorithm, a software package designed to implement the system to detect incorrect textual borrowings using external text collections is implemented. The load is reduced by selecting the documents that do not require detailed verification. Documents requiring detailed verification, on the other hand, undergo full verification.

A computational experiment demonstrating the efficiency of this method as well as the amount of computational resources saved is described. It is shown that on

marked and synthetic data, the approach reduces the flow of documents that do not require detailed verification by almost a third. It is important to emphasize that this not only accelerates the processing time of individual documents, but also allows us to use computing resources more purposefully, that is, to analyze in detail exactly those documents that need such a verification.

[1] *Nikitov A. V., Orchakov O. A, Chehovich Yu. V.* Plagiarism in works of undergraduate and graduate students: problem and methods of counteraction // University Management: Practice and Analysis, 2012. — (5):61-68.

[2] *Kuznetsova R. V., Bakhteev O. Yu., Chehovich Yu. V.* Methods Of Cross-lingual Text Reuse Detection In Large Textual Collections // Informatics and Applications, 2021. — 15:1 (2021), 30–41.

[3] *Chekhovich Y. V., Khazov A. V.* Analysis of duplicated publications in Russian journals // Journal of Informetrics, 2022. — 16(1):101246.

[4] *Safin K., Chehovich Yu.* Combined method for plagiarism detection in text documents // Proceedings of the Institute for System Programming of the RAS (Proceedings of ISP RAS), 2022. — 34(1):151-160.

Архитектура системы концептуального моделирования на основе метаграфового подхода

Тодосиев Никита Дмитриевич[1]⋆ todosievnd@student.bmstu.ru
Янковский Владислав[1] vlyankov@mail.ru
Гапанюк Юрий Евгеньевич[1] gapyu@bmstu.ru
[1]Москва, МГТУ им. Н.Э.Баумана

Человечество накопило огромное количество текстовых документов. Задача извлечения смысла текста из большого количества документов сложна для пользователя и может потребовать значительных временных затрат. Задача усложняется, когда пользователь является лицом, принимающим решения, и должен различать смысл входящих текстов и принимать решения за ограниченное время. Одним из распространенных способов "концептуального сжатия" текстовой информации является использование концептуальных моделей. К таким моделям относятся диаграммы ментальных карт, концептуальные карты и, частично, онтологические модели. В настоящее время ведутся исследования по разработке концептуальных моделей, представленных в виде сложных графовых структур, таких как гиперграфы, гиперсети и метаграфы. Использование сложных графовых структур обеспечивает значительную степень "концептуального сжатия". Использование концептуальных моделей в задаче принятия решений предполагает последовательное выполнение трех укрупненных этапов: синтез концептуальной модели на основе текстового описания; концептуальное моделирование, в результате которого формируются новые концептуальные модели; анализ результатов моделирования, принятие решений и формирование отчетов на основе принятых решений.

В данной работе применяется подход объединения концепций метаграфической модели и фреймовых концептов для концептуального сжатия. Для английского языка за основу был взят датасет FrameNet [1], для русского — FrameBank [2].

Архитектура системы концептуального моделирования состоит из трех больших модулей: "модуль анализа текста", "модуль генерации текста" и "модуль моделирования понятий метаграфа".

Работа системы состоит из девяти основных этапов. На "Шаге I" читается исходный текстовый документ. На "Шаге II" "модуль анализа текста" анализирует документ, извлекает понятия и отношения и создает структуру метаграфа. На "Шаге III" сгенерированная структура метаграфа записывается в "хранилище понятий метаграфа". На "Шаге IV" "модуль моделирования понятий метаграфа" получает исходные понятия для моделирования из "хранилища понятий метаграфа". На "Шаге V" выполняется концептуальное моделирование. Исходные понятия в форме метаграфа переводятся в целевые понятия. На "Шаге VI" результаты концептуального моделирования записываются в "хранилище метаграфических понятий". На "Шаге VII" "модуль генерации текста" получает

целевые понятия из “хранилища понятий метаграфа”. На “Шаге VIII” “модуль генерации текста” преобразует концепции назначения в текстовую форму. На “Шаге IX” генерируется выходной текстовый документ.

Архитектура системы также включает в себя “хранилище метаграфовых понятий”. Основные идеи хранения метаграфовой модели в реляционных, документно-ориентированных и графовых базах данных обсуждаются в [3].

В “модуле разбора текста” изначально исходный текст подвергается разрешению кореферентности, что позволяет удалить вершины, не несущие полезной информации, а также сделать будущий граф более связным. Далее текст попадает в модуль преобразования текста во фреймы [4]. После этого полученные фреймы связываются целевыми словами. Таким образом, получается граф, каждая вершина которого является словом. Атрибуты фрейма превращаются в аналогичные узлы. Связи между основным узлом и узлом атрибута помечаются соответствующим тегом атрибута фрейма. Дальнейшее обогащение фреймов приводит к формированию концептов высокого уровня на основе концептов низкого уровня. Но в рамках структуры плоского графа это невозможно, поэтому обратимся к модели метаграфа. Обогащенные фреймы можно рассматривать как метавершины метаграфа. Это связано с вложенной структурой результирующих фреймов — фреймы могут быть атомарными или содержать другие фреймы. В этой модели метавершины можно рассматривать как композиции низкоуровневых вершин.

В “модуле генерации текста” последовательно выполняются несколько шагов. Во-первых, выполняется обоснование запроса, который выделяет контекст запроса и определяет цель формирования текста. Затем на основе запроса и метаграфа собранных знаний извлекается подграф, называемый метаграфом ответа. Далее метаграф ответа разбивается на компоненты, которые преобразуются в формат словаря. Затем эти словари передаются модели T5 [5] для перевода из словарей в текстовое представление. Наконец, полученные предложения сортируются с помощью преобразователя BERT, настроенного на сортировку предложений в ответе.

Для оценки правильности предложений, полученных в ходе модуля разбора текста, необходимо учитывать смысловое сходство предложений, а не сходство предложений по сравнению с “золотым стандартом”, как это делается в существующих оценках, таких как BLEU, ROUGE, METEOR и другие. Для этого мы разрабатываем бенчмарк STS [6] для русского языка и строим оценку на основе трансформеров.

Проведенный эксперимент выполняется на существующем наборе примеров из датасета FrameBank. В дальнейшем планируется провести эксперимент с использованием новой оценки, а также с увеличением набора данных за счет использования других источников, в том числе текстов, преобразованных в метаграфы.

Проанализировав существующие подходы к представлению концептуальных карт, можно сделать вывод, что основной проблемой существующих подходов является использование плоского графа в качестве модели для представления концептуальной карты. Использование метавершин для описания концептуальных карт позволяет отказаться от представления концептуальной карты в виде плоского графа и перейти к холоническому пространственному описанию концептуальной карты в виде метаграфа.

[1] Baker, C.F., Fillmore, C.J., Lowe, J.B.: The berkeley FrameNet project. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics – Volume 1. pp. 86–90. ACL '98/COLING '98, Association for Computational Linguistics, USA, 1998.

[2] Lyashevskaya, O., Kashkin, E.: FrameBank: A database of russian lexical constructions. In: Analysis of Images, Social Networks and Texts. Springer International Publishing, 2015 — С. 350—360.

[3] Chernenkiy, V., Gapanyuk, Y., Kaganov, Y., Dunin, I., Lyaskovsky, M., Larionov, V.: Storing metagraph model in relational, document-oriented, and graph databases. In: Selected Papers of the XX International Conference on Data Analytics and Management in Data Intensive Domains, 2018. — С. 82—89.

[4] Swayamdipta, S., Thomson, S., Dyer, C., Smith, N.A.: Frame-semantic parsing with softmax-margin segmental RNNs and a syntactic scaffold, 2017.

[5] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer, 2019.

[6] Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., Specia, L.: SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Cross-lingual Focused Evaluation. In: Proceedings of the 10th International Workshop on Semantic Evaluation, 2017.

The Architecture of a Conceptual Modeling System Based on Metagraph Approach

Todosiev Nikita[1]⋆ todosievnd@student.bmstu.ru
Yankovskiy Vladislav[1] vlyankov@mail.ru
Gapanyuk Yuriy[1] gapyu@bmstu.ru
[1] Moscow, Bauman Moscow State Technical University

Humanity has accumulated a huge number of text documents. The task of extracting the meaning of a text from a large number of documents is difficult for the user and may require significant time costs. The task becomes more complicated when the user is a decision-maker and must distinguish the meaning of incoming texts and make decisions in a limited time. One of the common ways to "conceptually compress" text information is to use conceptual models. Such models include mindmap diagrams, concept maps, and, in part, ontological models. Research is currently underway to develop conceptual models presented in complex graph structures, such as hypergraphs, hypernetworks, and metagraphs. The use of complex graph structures provides a significant degree of "conceptual compression". The use of conceptual models in the decision-making task involves the sequential implementation of three enlarged steps: synthesis of a conceptual model based on a text description; conceptual modeling, as a result of which new conceptual models are formed; analysis of the results of modeling, decision-making, and the formation of reports based on the decisions made.

In this work, the approach of combining the metagraph model and frame concepts for conceptual transformation is applied. For the English language, FrameNet [1] was taken as the basis for the frame dataset, for Russian, the FrameBank [2].

The architecture of a conceptual modeling system consists of three large modules: the "text parsing module", the "text generation module" and the "metagraph concepts modeling module".

The operation of the system consists of nine main steps. In "Step I", a source text document is read. In "Step II", "the text parsing module" parses the document, extracts concepts and relationships, and creates a metagraph structure. In "Step III", the generated metagraph structure is recorded into "the metagraph concepts storage". In "Step IV", "the metagraph concepts modeling module" receives the source concepts for modeling from "the metagraph concepts storage". In "Step V", the conceptual modeling is performed. The source concepts in the form of metagraph are translated to the destination concepts. In "Step VI", the results of conceptual modeling are recorded into "the metagraph concepts storage". In "Step VII", "the text generation module" receives the destination concepts from "the metagraph concepts storage". In "Step VIII", "the text generation module" transforms destination concepts into text form. In "Step IX," the output text document is generated.

The system architecture also includes "the metagraph concepts storage". The main ideas of storing a metagraph model in relational, document-oriented, and graph databases are discussed in [3].

In the "the text parsing module", initially, the source text is subjected to the coreference resolution, which allows removing vertices that do not carry useful information, and also make the future graph more connected. Then the text gets into the module for converting text to frames [4]. After that, the received frames are linked by target words. Thus, a graph is obtained, each node of which is a word. The frame attributes are turned into similar nodes. Links between the main node and the attribute node are marked with the appropriate tag of the frame attribute. Further enrichment of frames leads to the formation of high-level concepts based on low-level concepts. But it is impossible within the framework of the flat graph structure, so we turn to the metagraph model. Enriched frames can be considered as metavertices of a metagraph. This is due to the nested structure of the resulting frames – frames can be atomic or contain other frames. In this model, metavertices may be considered as compositions of low-level vertices.

In the "the text generation module", several steps are performed sequentially. Firstly, query reasoning is carried out, which highlights the context of the request and determines the purpose of generating text. Then, a subgraph is extracted based on the query and the collected knowledge metagraph, called the response metagraph. Next, the response metagraph is divided into components, which are converted into a dictionary format. This dictionaries is then passed to the T5 [5] model for translation from dictionary to text representation. Finally, the received sentences are sorted using a BERT transformer configured to sort the sentences in the response.

To evaluate the correctness of sentences received during the work of the text parsing module, it is necessary to take into account the semantic similarity of sentences, and not the similarity of sentences compared to the "gold standard", as is done in existing evaluations, such as BLEU, ROUGE, METEOR and others. To do this, we are developing the STS benchmark [6] for Russian and building an evaluation based on transformers.

The conducted experiment is performed on existing dataset of FrameBank examples. The results proves the idea of this system. In the future, it is planned to conduct an experiment with usage of a new evaluation, as well as with increase of the dataset with other sources, including texts converted into a graph.

Having analyzed the existing approaches to the representation of conceptual maps, we can conclude that the main problem of the existing approaches is the use of a flat graph as a model for the representation of a conceptual map. The use of metavertices to describe conceptual maps allows us to abandon the representation of a conceptual map in the form of a flat graph and switch to a holonic spatial description of a conceptual map in the form of a metagraph.

[1] Baker, C.F., Fillmore, C.J., Lowe, J.B.: The berkeley FrameNet project. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and

17th International Conference on Computational Linguistics – Volume 1. pp. 86–90. ACL '98/COLING '98, Association for Computational Linguistics, USA, 1998.

[2] Lyashevskaya, O., Kashkin, E.: FrameBank: A database of russian lexical constructions. In: Analysis of Images, Social Networks and Texts. Springer International Publishing, 2015 — p. 350—360.

[3] Chernenkiy, V., Gapanyuk, Y., Kaganov, Y., Dunin, I., Lyaskovsky, M., Larionov, V.: Storing metagraph model in relational, document-oriented, and graph databases. In: Selected Papers of the XX International Conference on Data Analytics and Management in Data Intensive Domains, 2018. — p. 82—89.

[4] Swayamdipta, S., Thomson, S., Dyer, C., Smith, N.A.: Frame-semantic parsing with softmax-margin segmental RNNs and a syntactic scaffold, 2017.

[5] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer, 2019.

[6] Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., Specia, L.: SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Cross-lingual Focused Evaluation. In: Proceedings of the 10th International Workshop on Semantic Evaluation, 2017.

Использование методов кластеризации для анализа судебной практики арбитражных судов

Таран Мария Олеговна[1]⋆ mariyaot@list.ru
Гапанюк Юрий Евгеньевич[1] gapyu@bmstu.ru
[1]Москва, МГТУ им. Н.Э.Баумана

Судебной практикой можно назвать совокупность нескольких судебных актов, связанных какой-то общей темой. Причем эта тема может быть явно задана в документе, например, спор о взыскании денежных средств по договору поставки. Или документ может относиться к ней лишь косвенно, например, какие-то обстоятельства относятся к договору поставки, остальные касаются других тем.

В этой работе к судебным актам относятся результирующие документы (решения, определения, постановления), которые составляются по итогам рассмотрения судебного спора в арбитражных судах первой, апелляционной, кассационной инстанций.

К анализу судебной практики относится не только сбор статистических показателей, но и рассмотрение общих обстоятельств и выводов суда. В свою очередь, эта задача максимально близка к задаче экстрактивного реферирования нескольких документов, где на вход подают набор документов, а на выходе получают полезные абзацы из исходного текста.

Наиболее подходящими для реферирования судебного акта и судебной практики являются методы экстракции. В первую очередь это связано с тем, что точность передачи исходной информации является критичной для этой предметной области. Малейшее искажение может повлиять на позицию стороны в споре и результат его рассмотрения. Как указано в [1], использование методов абстракции может приводить к генерации ложных фактов, ошибкам и другим искажениям, что является недопустимым даже в единичных случаях.

Хотя набор документов для анализа подбирает специалист, основную сложность представляют вышеуказанные моменты с темами документов. Зачастую отбор судебных актов происходит не по теме документа, а по соответствию даже части документа неявному вопросу. На практике заинтересовать специалиста может один или два абзаца в многостраничном акте, что позволит ему добавить этот документ, как часть судебной практики.

В результате проведенной работы [2] были сделаны выводы о невозможности реферирования набора документов без дополнительной предварительной группировки, ввиду сложности применения на практике извлеченных текстовых блоков. Для минимизации влияния этого фактора были рассмотрены разные алгоритмы кластеризации. Их эффективность оценивалась на заранее подготовленном наборе документов, который применялся как “золотой стандарт”. В качестве основных признаков использовались как частотные, так и признаки, основанные на изучении предметной области. Также в работе предложен

метод кластеризации судебной практики, который показал лучший результат по сравнению с другими рассмотренными методами.

Так как кластеризация судебной практики являлась частью более крупной задачи по анализу судебной практики, то в работе был рассмотрен вариант отчета с извлечением выводов суда (реферат нескольких судебных актов). При этом использовался не один исходный набор документов, а несколько групп документов, сформированных по итогам кластеризации. Реферирование проводилось для каждой группы в отдельности. Документы, которые не попали в сформированные группы, реферировались как одиночные документы с использованием метода реферирования судебного акта описанного в [3].

Предложенный метод кластеризации судебной практики реализован в модуле кластеризации, который входит в гибридную интеллектуальную информационную систему реферирования и анализа судебной практики арбитражных судов.

В настоящей работе выполнены исследования по интеллектуальному анализу судебной практики с использованием методов кластеризации, предложены метод кластеризации судебной практики и метод реферирования судебной практики.

[1] *Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li* Faithful to the original: fact-aware neural abstractive summarization // In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence (AAAI'18/IAAI'18/EAAI'18). AAAI Press, Article 586, 4784–4791. (2021).

[2] *Taran, M.O., Revunkov, G.I., Gapanyuk, Y.E.* Creating a Brief Review of Judicial Practice Using Clustering Methods. In: Kryzhanovsky, B., Dunin-Barkowski, W., Redko, V., Tiumentsev, Y. (eds) Advances in Neural Computation, Machine Learning, and Cognitive Research VI. NEUROINFORMATICS 2022. Studies in Computational Intelligence, vol 1064. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-19032-2_48

[3] *Taran, M.O., Revunkov, G.I., Gapanyuk, Y.E.* Generating a Summary of a Court Act Based on an Improved Text Fragment Extraction Module. In: Kryzhanovsky, B., Dunin-Barkowski, W., Redko, V., Tiumentsev, Y., Klimov, V.V. (eds) Advances in Neural Computation, Machine Learning, and Cognitive Research V. NEUROINFORMATICS 2021. Studies in Computational Intelligence, vol 1008. Springer, Cham (2022).

Using Clustering Methods for Analysis Judicial Practice of Arbitration Courts

Taran Maria[1]⋆ mariyaot@list.ru
Gapanyuk Yuriy[1] gapyu@bmstu.ru
[1] Moscow, Bauman Moscow State Technical University

Judicial practice can be called a set of several judicial acts related to some common theme. Moreover, this topic can be explicitly set in the document, for example, a dispute about the recovery of funds under a supply agreement. Or the document may relate to it only indirectly, for example, some circumstances relate to the supply contract, the rest relate to other topics.

In this work, judicial acts include the resulting documents (decisions, rulings), which are drawn up following the consideration of a judicial dispute in arbitration courts of the first, appeal, and cassation instances.

The analysis of judicial practice includes not only the collection of statistical indicators, but also consideration of the general circumstances and conclusions of the court. In turn, this problem is as close as possible to the problem of extractive abstracting of several documents, where a set of documents is given as input, and useful paragraphs from the source text are received as output.

Extraction methods are the most suitable for summarizing a judicial act and judicial practice. First of all, this is due to the fact that the accuracy of the transmission of the original information is critical for this subject area. The slightest distortion can affect the position of the party in the dispute and the result of its consideration. As noted in [1], the use of abstraction methods can lead to the generation of false facts, errors and other distortions, which is unacceptable even in isolated cases.

Although the set of documents for analysis is selected by a specialist, the main difficulty is presented by the above points with the topics of documents. Often, the selection of judicial acts is not based on the topic of the document, but on the correspondence of even a part of the document to an implicit issue. In practice, a specialist may be interested in one or two paragraphs in a multi-page act, which will allow him to add this document as part of judicial practice.

As a result of the [2] work carried out, conclusions were drawn about the impossibility of abstracting a set of documents without additional preliminary grouping, due to the complexity of applying the extracted text blocks in practice. To minimize the influence of this factor, different clustering algorithms were considered. Their effectiveness was evaluated on a pre-prepared set of documents, which was used as a "gold standard". As the main features, both frequency and features based on the study of the subject area were used. Also, the paper proposes a method for clustering judicial practice, which showed the best result compared to other methods considered.

Since the clustering of judicial practice was part of a larger task of analyzing judicial practice, the paper considered a version of the report extracting the conclu-

sions of the court (an abstract of several judicial acts). In this case, not one initial set of documents was used, but several groups of documents formed as a result of clustering. Referencing was carried out for each group separately. Documents that did not fall into the formed groups were abstracted as single documents using the judicial act abstracting method described in [3].

The proposed method of judicial practice clustering is implemented in the clustering module, which is included in the hybrid intellectual information system of referencing and analyzing the judicial practice of arbitration courts.

In this work, research on the intellectual analysis of judicial practice using clustering methods is carried out, the method of clustering judicial practice and the method of referencing judicial practice are proposed.

[1] *Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li* Faithful to the original: fact-aware neural abstractive summarization // In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence (AAAI'18/IAAI'18/EAAI'18). AAAI Press, Article 586, 4784–4791. (2021).

[2] *Taran, M.O., Revunkov, G.I., Gapanyuk, Y.E.* Creating a Brief Review of Judicial Practice Using Clustering Methods. In: Kryzhanovsky, B., Dunin-Barkowski, W., Redko, V., Tiumentsev, Y. (eds) Advances in Neural Computation, Machine Learning, and Cognitive Research VI. NEUROINFORMATICS 2022. Studies in Computational Intelligence, vol 1064. Springer, Cham (2023).

[3] *Taran, M.O., Revunkov, G.I., Gapanyuk, Y.E.* Generating a Summary of a Court Act Based on an Improved Text Fragment Extraction Module. In: Kryzhanovsky, B., Dunin-Barkowski, W., Redko, V., Tiumentsev, Y., Klimov, V.V. (eds) Advances in Neural Computation, Machine Learning, and Cognitive Research V. NEUROINFORMATICS 2021. Studies in Computational Intelligence, vol 1008. Springer, Cham (2022).

Генерация вопросов на естественном языке с применением подхода Гибридных Интеллектуальных Информационных Систем

Белянова Марина Александровна[1]⋆ flerchy@gmail.com
Гапанюк Юрий Евгеньевич[1] gapyu@bmstu.ru
[1]Москва, МГТУ им. Н.Э.Баумана

Задавать вопросы – это ключевая функция человеческого сознания, которая ведёт к более эффективному процессу обучения и улучшению качества полученной информации. В книге [1] вопросы описаны как “рабочие лошадки в строительстве фундаментального знания”.

Вопросы также могут быть использованы в качестве инструмента обучения, с помощью которого преподаватели управляют и направляют процесс обучения, проверяют, насколько хорошо обучающиеся поняли материал, и определяют темы, с которыми могут быть проблемы. Умелое использование вопросов может улучшить обучение и увеличить продуктивность обучающихся [2].

Как утверждается в исследовании [3], преподаватель чаще использует конвергентные вопросы по сравнению с процедурными (13,3%) или дивергентными (26,7%), потому что это в большей степени воодушевляет учащихся отвечать на вопросы по материалу, из которого получены знания.

Исследование [4] показывает, что инструкции с дополнительной секцией ответов на вопросы более эффективны, чем инструкции без неё, а вопросы, которые задаёт учитель кроме того улучшают обучение учащегося, развивая навыки критического мышления, подкрепляя правильное, и исправляя неверное понимание концептов и связей, а также предоставляя обратную связь. Вопросы, сформулированные учителем, позволяют ему лучше управлять учебным процессом [5]. Также учитель воодушевляет обучающихся лучше понимать материал, задавая вопросы, которые требуют обобщения пройденного материала.

В настоящее время ручной труд учителя по созданию вопросов можно автоматизировать на основе методов интеллектуального анализа текстов, которые применяются к текстовым учебным материалам и автоматически генерируют вопросы.

В [6] мы предлагаем использовать подход на основе Гибридных Интеллектуальных Информационных Систем (ГИИС) как основу для архитектуры системы генерации вопросов. Три основных компонента ГИИС это среда, модуль подсознания, и модуль сознания. В [6] также приводятся результаты экспериментов, проведенных на системе, построенной по архитектуре ГИИС.

“Средой” в предлагаемой системе являются текстовые документы и графы знаний. Словари и тезаурусы, которые дополнительно используются для смыслового обогащения как текстов, так и графов знаний, также могут рассматриваться как элементы среды.

Модуль подсознания системы включает в себя следующие модули.

“Модуль извлечения текста” используется для выделения текстовых фрагментов, предназначенных для формирования вопросов и смыслового обогащения текстов на основе словарей и тезаурусов.

“Модуль извлечения графов знаний” реализует извлечение и хранение графов знаний в репозитории в различных форматах. Извлечение графа знаний не используется, если исходные данные явно представлены в формате графа знаний (например, в формате RDF).

Следует отметить, что графы знаний всегда неявно присутствуют в исходном тексте. Поэтому “модуль извлечения понятий и связей” позволяет строить графы знаний на основе исходного текста, выделяя понятия и связи между ними.

Если для генерации вопросов предполагается использовать только логические методы, то никаких дополнительных модулей не требуется. Отобранные и обогащенные фрагменты текстов и графов знаний сохраняются в репозитории и в дальнейшем используются для логической генерации вопросов. Но если для генерации вопросов предполагается использовать методы, основанные на машинном обучении, то для текстов и графов знаний необходимо формировать векторные представления (эмбеддинги). Для решения этой задачи используются “модуль векторизации текста” и “модуль векторизации графа знаний”.

Для хранения гибридных исходных данных необходим репозиторий, обеспечивающий хранение извлеченных и обогащенных текстов и графов знаний как в исходном, так и в векторном представлении. Для реализации хранилища рекомендуется использовать модель данных на основе метаграфа, позволяющую создавать сложные связи между элементами хранилища, представленными в текстовом, графическом и векторизованном виде.

Модуль сознания системы включает в себя следующие модули.

Модуль “логических методов” используется для генерации вопросов на основе извлеченных фрагментов графов знаний, исходных текстов, а также понятий и связей между ними, извлеченных из исходных текстов. Для генерации вопросов данный модуль использует методы, основанные на правилах.

Модуль “машинного обучения” использует методы, основанные на машинном обучении. Входными данными для этого модуля являются векторные представления текстов и графов знаний.

“Модуль логической коррекции вопросов” может дополнительно обрабатывать вопросы, сгенерированные модулем “машинного обучения”. Этот модуль также основан на правилах и предназначен для исправления возможных ошибок, допущенных при использовании методов машинного обучения. В качестве исправления могут быть использованы перестановки слов, изменение частей слов с целью согласования родов, падежей, склонений, спряжений и т.д., а также проверка фактов в тексте и замена неверных фактографических данных (таких, как имена, географические наименования, даты).

“Гибридный модуль генерации вопросов” может использовать либо данные из репозитория, либо выходные данные модулей генерации вопросов в качестве входных данных. В случае использования результатов работы предыдущих модулей, этот модуль реализует ансамблевую модель.

“Модуль оценки качества” сравнивает качество сгенерированных вопросов на основе метрик качества. Пользователю представляются варианты сгенерированных вопросов с метриками качества.

Таким образом, предлагаемый подход на основе ГИИС позволяет экспериментировать с различными вариантами архитектур интеллектуальной системы генерации вопросов. Это могут быть варианты архитектуры, реализованные как на основе правил и на основе методов машинного обучения, так и на основе гибридного варианта, включающего оба подхода.

[1] *Walsh, Jackie Acree, and Beth Dankert Sattes.* Quality questioning: Research-based practice to engage every learner. Corwin Press, 2016.

[2] *Caram, Chris A., and Patsy B. Davis.* Inviting student engagement with questioning // Kappa Delta Pi Record 42.1 (2005): 19-23.

[3] *Andana, Yona.* The types of teacher's questions in english teaching-learning process at MAN Mojokerto. Diss. UIN Sunan Ampel Surabaya, 2018.

[4] *Marzano, Robert J., Debra Pickering, and Jane E. Pollock* Classroom instruction that works: Research-based strategies for increasing student achievement // Ascd, 2001.

[5] *Bowker, Matthew H.* Teaching students to ask questions instead of answering them // Thought & Action 26 (2010): 127-134.

[6] *Belyanova, Marina A., Ark M. Andreev, and Yuriy E. Gapanyuk.* Neural Text Question Generation for Russian Language Using Hybrid Intelligent Information Systems Approach // International Conference on Neuroinformatics. Springer, Cham, 2021.

[7] *Musingafi, Maxwell Constantine Chando, and Kwaedza Enety Muranda.* Students and questioning: A review of the role played by students generated questions in the teaching and learning process // Studies in Social Sciences and Humanities 1.3 (2014): 101-107.

Text Question Generation based on Hybrid Intelligent Information Systems Approach

Belyanova Marina[1]⋆ flerchy@gmail.com
Gapanyuk Yuriy[1] gapyu@bmstu.ru
[1] Moscow, Bauman Moscow State Technical University

Asking questions is an essential function of the human mind that leads to an improvement of the learning process and the enhancing of the received information quality. The book [1] stated that "questions are 'workhorses' in building foundational knowledge".

Questions can also be used as a teaching tool by which instructors manage and direct learning, test student understanding, and diagnose problem areas. The skillful use of questioning can enhance learning and increase student performance. [2]

As stated in the research [3], the teacher uses convergent questions sufficiently more frequent, than procedural (13.3%) or divergent (26.7%) questions, and this is used to encourage students to answer based on the material.

In general, research [4] states that instructions with questioning are more effective than those without it, and teacher-initiated questions enhance student learning by developing critical thinking skills, reinforcing student understanding, correcting student misunderstanding and providing feedback for students. Teacher generated questions are questions that keep control of the learning process in the hands of the teacher [5]. Teachers engage students by asking questions that require generative answers.

Currently, the manual work of the teacher in creating questions can be automated based on text mining methods that are applied to text-based educational materials and automatically generate questions.

In [6] we propose Hybrid Intelligent Information Systems (HIIS) approach to be used as a basis of the new architecture for question generation. Three main components of the HIIS are the "environment", the subconsciousness module, and the consciousness module.

In this research we additionally provide results of the experiments, conducted on the system, that was constucted following this architecture.

The "environment" in the proposed system are text documents and knowledge graphs. Dictionaries and thesauri, which are additionally used for semantic enrichment of both texts and knowledge graphs, can also be considered as elements of the environment.

The subconsciousness module of the system includes the following modules.

The "text extraction module" is used to highlight text fragments intended for the formation of questions and the semantic enrichment of texts based on dictionaries and thesauri.

The "knowledge graph extraction module" implements the extraction and storage of knowledge graphs in the repository in various formats. Any knowledge graph

extraction is not required if the original data is presented in an explicit knowledge graph format (for example, in the RDF format).

It should be noted that knowledge graphs are implicitly present in the source text. Therefore, the "concepts and links extraction module" allows you to build knowledge graphs based on the source text, highlighting concepts and links between them.

If only logical methods are supposed to be used to generate questions, then no additional modules are required. The selected and enriched fragments of texts and knowledge graphs are stored in the repository and are subsequently used for the logical generation of questions. But if it is supposed to use methods based on machine learning to generate questions, then for texts and knowledge graphs it is necessary to form vector representations (embeddings). To solve this problem, the "text vectorization module" and the "Knowledge graph vectorization module" are used.

Hybrid source data corresponds to a repository that provides storage of extracted and enriched texts and knowledge graphs in both the original and vectorized representation. To implement the storage, it is recommended to use the metagraph data model, which allows creating complex links between the storage elements presented in text, graph, and vectorized forms.

The consciousness module of the system includes the following modules for question generation.

The "logical methods" module is used to generate questions based on the extracted fragments of knowledge graphs, source texts, as well as concept and links between them, that are extracted from source texts. It uses rule-based methods to generate questions.

The "machine learning" module uses machine learning-based methods to generate question texts. The input data for this module are vectorized representations of texts and knowledge graphs.

The "module of logical correction of questions" can additionally process questions generated by the "machine learning" module. This module is also rule-based and is designed to correct possible errors made by machine learning methods. Permutations of words, changing parts of words in order to harmonize genders, cases, declensions, conjugations, etc., as well as fact-checking in the text and replacing incorrect factual data (such as names or dates) can be used as the correction methods.

The "hybrid question generation module" can use either data from the repository or the output of the generating questions modules as its input data. In the case of using the results of the work of the previous modules, this module implements the ensemble model.

The "quality assessment module" compares the quality of the generated questions based on quality metrics. Variants of generated questions with quality ratings are presented to the user.

Thus, the proposed HIIS-based approach allows experiments with various architectures of an intelligent question generation system. These can be architecture options implemented both on the basis of rules and based on machine learning methods, as well as on the basis of a hybrid option that includes both approaches.

[1] *Walsh, Jackie Acree, and Beth Dankert Sattes.* Quality questioning: Research-based practice to engage every learner. Corwin Press, 2016.

[2] *Caram, Chris A., and Patsy B. Davis.* Inviting student engagement with questioning // Kappa Delta Pi Record 42.1 (2005): 19-23.

[3] *Andana, Yona.* The types of teacher's questions in english teaching-learning process at MAN Mojokerto. Diss. UIN Sunan Ampel Surabaya, 2018.

[4] *Marzano, Robert J., Debra Pickering, and Jane E. Pollock* Classroom instruction that works: Research-based strategies for increasing student achievement // Ascd, 2001.

[5] *Bowker, Matthew H.* Teaching students to ask questions instead of answering them // Thought & Action 26 (2010): 127-134.

[6] *Belyanova, Marina A., Ark M. Andreev, and Yuriy E. Gapanyuk.* Neural Text Question Generation for Russian Language Using Hybrid Intelligent Information Systems Approach // International Conference on Neuroinformatics. Springer, Cham, 2021.

[7] *Musingafi, Maxwell Constantine Chando, and Kwaedza Enety Muranda.* Students and questioning: A review of the role played by students generated questions in the teaching and learning process // Studies in Social Sciences and Humanities 1.3 (2014): 101-107.

Упорядочивание гипотез в моделях перевода с использованием человеческой разметки

Скачков Николай Андреевич[1]⋆ nikolaj-skachkov@yandex.ru
Воронцов Константин Вячеславович[1] vokov@forecsys.ru
[1]Москва, ВЦ ФИЦ ИУ РАН

Создание систем автоматического перевода является одной из сложных задач анализа текстов естественного языка. Обучение алгоритмов, лежащих в основе таких систем, требует большого количества параллельных текстов на разных языках и существенно зависит от качества этих данных и степени их выравненности. Ввиду высокой стоимости работы профессиональных переводчиков данные для обучения алгоритмов перевода собираются автоматически с помощью эвристических алгоритмов. При этом высокая степень выравненности отдельных обучающих примеров не гарантируется, что позволяет собирать большие объёмы параллельных текстов.

Алгоритмы перевода при обучении воспроизводят данные на вход обучающие примеры [1]. Плохо выравненные примеры приводят к появлению таких систематических ошибок перевода, как недопереводы и избыточные переводы, когда в переведённом тексте либо теряется часть исходного смысла, либо добавляется какой-то новый. Для борьбы с этими ошибками можно использовать обучение с негативными примерами [2]. При таком подходе выход алгоритма перевода рассматривается как некоторое естественное ранжирование гипотез, в котором сгенерированные гипотезы отсортированы по вероятности текста с точки зрения модели перевода. Улучшения качества перевода при таком подходе можно достичь за счёт увеличения вероятности некоторого перевода y_+ предложения x относительно более плохого перевода y_- того же предложения x с помощью максимальной целевой функции интервала:

$$L(x, y_+, y_-) = \max(0, \log P(y_-|x) - \log P(y_+|x) + \alpha).$$

В простом случае получить более плохой перевод y_- из перевода y_+ можно с помощью эвристических аугментаций. Так, с помощью удаления и дублирования случайных слов в переводах удаётся улучшить качество системы за счёт уменьшения количества ошибок недопереводов и избыточных переводов [2]. Также, с помощью случайных изменений грамматических форм слов экспериментально удалось добиться улучшения согласованности текстов с точки зрения норм языка.

Однако описанный подход с синтетическими примерами позволяет бороться только с определёнными видами ошибок перевода, заданными с помощью эвристики в аугментации. В то же время сложность генерируемых примеров также ограничена сложностью эвристики. Всё это ограничивает широкое при-

менение данного подхода и снижает эффект от улучшения ранжирования. Более общим решением могло бы быть использование человеческих исправлений машинно-переведённого текста. Такие данные позволяли бы исправлять любой вид систематических ошибок, встречающихся в переводах, но это решение всё еще трудно масштабируемо из-за сложности задания.

В данной работе предлагается более простой способ интеграции человека в процесс улучшения перевода. Вместо того, чтобы просить переводить текст или исправлять в нём ошибки, будем просить разметчика выбрать из двух переводов лучший. В соответствии с разметкой будем представлять модели в процессе обучения соответствующие y_+ и y_- для улучшения ранжирования.

Экспериментально подтверждается, что данный подход позволяет заметно улучшить качество перевода в среднем, а также позволяет значимо уменьшить количество недопереводов и избыточных переводов без использования синтетических примеров. Более того, обучение с улучшением ранжирования на англо-русском направлении в 16% случаев генерирует более хорошие переводы с точки зрения разметчиков, не участвовавших в создании обучающих данных.

Модель	Функция потерь	en-ru-wmt-19, BLEU
базовая модель	ПЭ	35.6
дообуч. на разметку	ПЭ	36.7
дообуч. на разметку	$L(x, y_+, y_-)$	**37.3**

Кроме того, улучшение качества удалось перенести на языковые направления отличные от англо-русского, для которых не собирались данные разметки. При обучении многоязыковых моделей типа Many2One [3] с предложенным методом удалось повысить качество перевода в 5-10% случаев на французско-русском и немецко-русском направлениях перевода за счёт переноса знаний между языками.

Также, благодаря предложенной процедуре улучшения с человеческой разметкой удалось добиться доменной адаптации модели под языковой домен, на котором не удалось собрать параллельных данных. Без предложенного подхода данная процедура потребовала бы существенной работы переводчиков.

Таким образом, благодаря описанным методам упорядочивания гипотез перевода удалось интегрировать человеческую разметку в процесс обучения моделей перевода. Данный подход показывает заметный прирост качества моделей перевода, уменьшение долей систематических ошибок перевода, а также открывает возможности для более быстрой доменной адаптации машинного перевода.

[1] *Felix Stahlberg* (2019) Neural Machine Translation: A Review

[2] *Yang, Zonghan and Cheng, Yong and Liu, Yang and Sun, Maosong* (2019) Reducing Word Omission Errors in Neural Machine Translation: A Contrastive Learning Approach, Association for Computational Linguistics, pp 6191–6196

[3] *Johnson, Melvin and Schuster, Mike and Le, Quoc V. and Krikun, et. al* Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation // CoRR, 2016.

Hypotheses re-ranking in translation models using human markup

Skachkov Nikolay[1]⋆ nikolaj-skachkov@yandex.ru
Vorontsov Konstantin[1] vokov@forecsys.ru
[1]Moscow, CC FRC CSC RAS

The creation of automatic translation systems is one of the complex tasks in NLP. Learning neural translation models requires a large amount of parallel texts in different languages and significantly depends on the quality of this data and the degree of the alignment. Due to the high cost of professional translators, data for training translation algorithms is collected automatically using heuristic algorithms. At the same time, quality of individual training examples is not guaranteed, which allows to collect large volumes of parallel texts.

Translation algorithms during training reproduce the input data [1]. Poorly aligned examples lead to such systematic translation errors as undertranslations and redundant translations, when a part of the original meaning is either lost in the translated text, or some new one is added. To fix these errors, one can use training with negative examples [2]. With this approach, the output of the translation algorithm is considered as ranking, in which the generated hypotheses are sorted by the probability of the text from the translation model's point of view. Improving the quality of translation with this approach can be achieved by increasing the probability of some translation y_+ of the sentence x relative to a poorer translation y_- of the same sentence x using the maximum margin loss:

$$L(x, y_+, y_-) = \max(0, \log P(y_-|x) - \log P(y_+|x) + \alpha).$$

In a simple case, one can get a worse translation of y_- from the translation of y_+ using heuristic augmentations. Thus, by removing and duplicating random words in translations, it is possible to improve the quality of the system by reducing the number of errors of undertranslations and redundant translations [2]. Also, with the help of random changes in grammatical forms of words, we improved the consistency of texts in terms of language fluency in our experiments.

However, the described approach with synthetic examples allows to deal only with certain types of translation errors specified using heuristics in augmentations. At the same time, the complexity of the generated examples is also limited by the complexity of the heuristics. All this limits the wide application of this approach and reduces the effect of improved ranking. A more general solution could be to use post-editing of machine-translated text. Such data would make it possible to correct any kind of systematic errors encountered in translations, but this solution is still difficult to scale due to the complexity of the task.

This paper presents a simpler way to integrate a person into the process of improving translation. Instead of asking to translate the text or correct errors in it,

we will ask the assessors to choose the best of the two translations. In accordance with the markup, we will feed the model with the corresponding y_+ and y_- to improve ranking.

It is experimentally confirmed that this approach makes it possible to significantly improve the quality of translation on average, and also significantly reduces the number of undertranslations and redundant translations without using synthetic examples. Moreover, training with improved ranking in the English-Russian direction in 16% of cases generates better translations from the assessor's point of view.

Model	Loss Function	en-ru-wmt-19, BLEU
base model	CE	35.6
training with markup	CE	36.7
training with markup	$L(x, y_+, y_-)$	**37.3**

In addition, the quality improvement was transferred to language directions other than English-Russian, for which markup data was not collected. French-Russian and German-Russian translation directions have been improved in 5-10% cases due to the transfer of knowledge between languages when teaching multilingual models of the Many2One type [3] with the proposed method of improving ranking.

Also, thanks to the proposed procedure for ranking improvement with human markup, it was possible to adapt the translation model to the language domain on which parallel data could not be collected. Without the proposed approach, this procedure would require substantial work of translators.

Thanks to the described methods of ranking improvements in translation models, it was possible to develop an approach for embedding human markup in the learning process of translation models. This approach has shown noticeable increases in the quality of translation models, a decrease of systematic translation errors share, and also opens up opportunities for faster domain adaptation of machine translation.

[1] *Felix Stahlberg* (2019) Neural Machine Translation: A Review

[2] *Yang, Zonghan and Cheng, Yong and Liu, Yang and Sun, Maosong* (2019) Reducing Word Omission Errors in Neural Machine Translation: A Contrastive Learning Approach, Association for Computational Linguistics, pp 6191–6196

[3] *Johnson, Melvin and Schuster, Mike and Le, Quoc V. and Krikun, et. al* Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation // CoRR, 2016.

Специфика теоретико-графовых моделей в задаче атрибуции фольклорных и литературных произведений

Москин Николай Дмитриевич[1]⋆ moskin@petrsu.ru
Рогов Александр Александрович[1] rogov@petrsu.ru
Лебедев Александр Александрович[1] perevodchik88@yandex.ru
[1]Петрозаводск, Петрозаводский государственный университет

Для решения различных вопросов, связанных с анализом текстов, активно используются компьютерные технологии и математические методы. Существует ряд способов, предусматривающих решение собственно лингвистических задач (таких, как машинный перевод, классификация текстов, генерация текстов, определение плагиата и т. п.) с применением методов машинного обучения и искусственного интеллекта. Преимущества подобного математического подхода очевидны — с его помощью можно отыскать различные скрытые закономерности, которые, скорее всего, не будут обнаружены специалистом-филологом. Например, нейросетевые технологии (такие, как recurrent neural network, convolutional neural network, Transformer) способны эффективно выделять скрытые языковые закономерности, но нуждаются при этом в тонкой настройке, а также в привлечении большого объема языковых данных. Еще один значимый минус подобного подхода состоит в том, что полученные результаты зачастую не имеют в явном виде сформулированного обоснования (что вызывает закономерные вопросы и возражения у филологов, а, следовательно, они не могут быть введены в научный оборот).

Стилеметрия – это направление исследований, которое активно задействуется в решении задач, связанных с атрибуцией текстов (в ходе такого анализа можно получить конкретные статистические параметры, которые позволят идентифицировать авторский стиль) [4]. Сложность подхода заключается в том, что текст обладает структурой, состоящей из множества уровней и планов содержания. Текст может быть раздроблен на структурные уровни (к примеру, фонетический, морфологический, синтаксический, семантический и т. п.) со своими компонентами, а также может предусматривать разные типы связей между такими компонентами – и, как следствие, может быть смоделирован по-разному.

Полученные модели текстов могут быть представлены как графы, состоящие из вершин-объектов и ребер — связей между этими объектами. Такой подход к формированию теоретико-графовых моделей обладает преимуществами в решении задач. Формирование обобщенной модели позволяет извлекать комплексную информацию из исследуемых текстов, что очень важно в анализе литературных и фольклорных источников (коллекции таких документов могут быть невелики по объему, а увеличить их не представляется возможным; также нередка ситуация несбалансированности выборок — объем текстов первого автора может быть значительно больше, чем у второго автора). Синтез различных моделей (в том числе тех, которые были апробированы ранее, как, напри-

мер, моделей, изложенных в работах [1, 5]) позволяет формировать гибридные структуры, которые потенциально являются более совершенными, чем выделяемые одиночные признаки. Сравнение результатов, полученных с использованием различных методов, позволит выделить те из них, которые будут наиболее эффективны. Все это подчеркивает важность единообразия в описании разных типов теоретико-графовых моделей, их систематизации и классификации.

В работе [2] представлена обобщенная контекстно-зависимая теоретико-графовая модель текста $G = (V, H, E, \alpha, \beta, \mu, \gamma)$, заданная рекурсивным образом. Минимальной структурной единицей модели является слово $w_k \in W$ ($k = 1, 2, \ldots, K$, где $K > 0$ – количество слов в тексте T). Подмножества слов объединяются в вершины, определяющие множество V, причем одно и то же слово может иметь отношение к разным вершинам. Графовые подструктуры из множества $H = \{\{v_i\}_{i=1}^{n} \cup \{G_j = (V_j, H_j, E_j)\}_{j=1}^{m}\}$ и ребра из множества $E \subset H \times H$ отражают лексические, синтаксические и семантические связи текста. Характеристиками модели являются ее нечеткость (заданная функцией $\mu : H \cup E \to [0, 1]$), иерархичность (определяется H) и темпоральность (k соответствует порядку появления слова w_k в тексте). Функции α и β задают атрибуты (метки) вершинам и ребрам соответственно. Соответствие между объектами теоретико-графовой модели $x_i \in H \cup E$ и подмножествами слов $W_l \subset W$ задается с помощью функции γ.

Методика построения теоретико-графовых моделей была апробирована на различных группах текстов, среди которых бесёдные песни; фольклорные материалы и тексты, стилизованные под фольклор; анонимные и псевдонимные публицистические статьи, написанные в середине XIX века. Некоторые примеры моделей представлены в [2, 3]. Для хранения и последующего анализа текстов в информационной системе «Фольклор» был реализован формат SNG. Этот формат представляет собой текстовый файл, который можно легко редактировать. Файл делится на пять частей: общие характеристики текста, слова, объекты, связи и матрица инцидентности (технически части разделены между собой одиночной строкой с комментариями, начинающимися с символов //) [2]. Описание единого подхода к построению различных теоретико-графовых моделей текстов позволяет находить расстояния между ними. Это позволит решать задачи классификации и кластеризации различных текстов или их фрагментов, что полезно не только при решении задачи атрибуции, но и при поиске неоднородных фрагментов в тексте.

[1] *Милов Л. В., Бородкин Л. И., Иванова Т. В., Неберекутина Е. В., Полянская И. В., Романкова Н. В., Саркисова Г. И.* От Нестора до Фонвизина: Новые методы определения авторства, Москва: Прогресс, 1994. — 445 с.

[2] *Москин Н. Д., Рогов А. А., Воронов Р. В.* Обобщенная контекстно-зависимая теоретико-графовая модель фольклорных и литературных текстов // Труды Института системного программирования РАН, Москва: ИСП РАН, 2022. — Т. 34. No. 1. — С. 73–86.

[3] *Москин Н. Д.* Теоретико-графовые модели фольклорных текстов и методы их анализа, Петрозаводск: Изд-во ПетрГУ, 2013. — 148 с.

[4] *Рогов А. А., Абрамов Р. В., Бучнева Д. Д., Захарова О. В., Кулаков К. А., Лебедев А. А., Москин Н. Д., Отливанчик А. В., Савинов Е. Д., Сидоров Ю. В.* Проблема атрибуции в журналах “Время”, “Эпоха” и еженедельнике “Гражданин”, Петрозаводск: Изд-во “Острова”, 2021. — 391 с.

[5] *Севбо И. П.* Графическое представление синтаксических структур и стилистическая диагностика, Киев: Наукова Думка, 1981. — 192 с.

The specifics of graph-theoretic models in the task of attribution of folklore and literary works

Moskin Nikolai[1]⋆ moskin@petrsu.ru
Rogov Alexander[1] rogov@petrsu.ru
Lebedev Alexander[1] perevodchik88@yandex.ru
[1]Petrozavodsk, Petrozavodsk State University

Computer technologies and mathematical methods are actively used to solve various issues related to text analysis . There are a number of ways to solve linguistic problems themselves (such as machine translation, text classification, text generation, plagiarism detection, etc.) using machine learning and artificial intelligence methods. The advantages of such a mathematical approach are obvious – with its help, you can find various hidden regularities that, most likely, will not be discovered by a specialist philologist. For example, neural network technologies (such as recurrent neural network, convolutional neural network, Transformer) are able to effectively identify hidden language regularities, but at the same time they need fine adjustment, as well as attracting a large amount of language data. Another significant disadvantage of this approach is that the obtained results often do not have an explicitly formulated justification (which raises legitimate questions and objections from philologists and, therefore, they cannot be introduced into scientific circulation).

Stylometry is the direction of research that is actively used in solving problems related with the text attribution (during such analysis, it is possible to obtain specific statistical parameters that will allow identifying the author's style) [4]. The complexity of the approach lies in the fact that the text has a structure consisting of many levels and content plans. The text can be divided into structural levels (for example, phonetic, morphological, syntactic, semantic, etc.) with its own components, and can provide different types of connections between such components – and, as a result, can be modeled in different ways.

The resulting text models can be represented as graphs consisting of vertices-objects and edges – connections between these objects. This approach to the formation of graph-theoretic models has advantages in solving problems. The formation of a generalized model allows you to extract comprehensive information from the studied texts, which is very important in the analysis of literary and folklore sources (collections of such documents may be small in volume, and it is not possible to increase them; the situation of unbalanced samples is also not uncommon – the volume of texts of the first author may be significantly larger than the second author). The synthesis of various models (including those that have been tested earlier, such as models described in the works [1, 5]) allows us to further form hybrid structures that are potentially more perfect than isolated single features. Comparing the results obtained using different methods will allow you to identify those that will be

most effective. All this underlines the importance of uniformity in the description of different types of graph-theoretic models, their systematization and classification.

At work [2] a generalized context-dependent graph-theoretic model of text $G = (V, H, E, \alpha, \beta, \mu, \gamma)$ is presented, given recursively. The minimum structural unit of the model is the word $w_k \in W$ ($k = 1, 2, \ldots, K$, where $K > 0$ is the number of words in the text T). Subsets of words are combined to the vertices defining the set V, and the same word can relate to different vertices. Graph substructures from the set $H = \{\{v_i\}_{i=1}^{n} \cup \{G_j = (V_j, H_j, E_j)\}_{j=1}^{m}\}$ and edges from the set $E \subset H \times H$ reflect the lexical, syntactic and semantic connections of the text. The characteristics of the model are its fuzziness (defined by the function $\mu : H \cup E \to [0, 1]$), hierarchy (defined by H) and temporality (k corresponds to the order of appearance of the word w_k in the text). The functions α and β set attributes (labels) to vertices and edges , respectively. The correspondence between the objects of the graph-theoretic model $x_i \in H \cup E$ and subsets of words $W_l \subset W$ is set using the function γ.

The method of constructing graph-theoretic models was tested on various groups of texts, including conversational songs; folklore materials and texts stylized as folklore; anonymous and pseudonymous publicistic articles written in the middle of the XIX century. Some examples of models are presented in [2, 3]. For the storage and subsequent analysis of texts in the information system "Folklore", the SNG format was implemented. This format is a text file that can be easily edited. The file is divided into five parts: general characteristics of the text, words, objects, connections and the incidence matrix (technically the parts are separated by a single line with comments starting with characters //) [2]. Description of a unified approach to the construction of various graph-theoretic models of texts makes it possible to find the distances between them. It will allow solving the problems of classification and clustering of various texts or their fragments. It is useful not only when solving the attribution problem, but also when searching for nonuniform fragments in the text.

[1] *Milov L. V., Borodkin L. I., Ivanova T. V., Neberekutina E. V., Polyanskaya I. V., Romankova N. V., Sarkisova G. I.* From Nestor to Fonvizin: New methods for determining authorship, Moscow: Progress Publishing House, 1994. — 445 p.

[2] *Moskin N. D., Rogov A. A., Voronov R. V.* Generalized context-dependent graph-theoretic model of folklore and literary texts // Proceedings of ISP RAS, Moscow: ISP RAS, 2022. — Vol. 34. No.1. — P. 73–86.

[3] *Moskin N. D.* Graph-theoretic models of folklore texts and methods of their analysis, Petrozavodsk: PetrSU Publishing House, 2013. — 148 p.

[4] *Rogov A. A., Abramov R. V., Buchneva D. D., Zakharova O. V., Kulakov K. A., Lebedev A. A., Moskin N. D., Otlivanchik A. V., Savinov E. D., Sidorov Y. V.* The problem of attribution in the magazines "Time", "Epoch" and the weekly "Citizen", Petrozavodsk: Islands Publishing House, 2021. — 391 p.

[5] *Sevbo I. P.* Graphical representation of syntactic structures and stylistic diagnostics, Kiev: Naukova Dumka, 1981. — 192 p.

Распознавание таблиц в форматированных документах

Копаничук Илья Владимирович[1]⋆ kopanichuk@ap-team.ru
Очнева Инга Михайловна[1] ochneva@ap-team.ru
Огальцов Александр Владимирович[1] ogaltsov@ap-team.ru
Каприелова Мариам Семеновна[1,3] kaprielova@ap-team.ru
Финогеев Евгений Леонидович[1] finogeev@ap-team.ru
Кильдяков Александр Сергеевич[1] kildyakov@ap-team.ru
Чехович Юрий Викторович[1,3] chehovich@ap-team.ru

[1]Москва, Антиплагиат
[2]Москва, Московский физико-технический институт (национальный исследовательский университет)
[3]Москва, Федеральный исследовательский центр «Информатика и управление» РАН

При поиске заимствований в письменных работах значительной проблемой является распознавание отдельных структурных элементов документа: таблиц, иллюстраций, оглавления, библиографии. Без точного определения структурного элемента, к которому относится конкретный фрагмент текста, невозможен учет особенностей этого структурного элемента при выявлении заимствований. Растет количество ложноположительных и ложноотрицательных ошибок, как и количество методов маскировки заимствований для недобросовестных пользователей [1]. В докладе предлагается мультиязычный метод распознавания таблиц.

Мы свели проблему распознавания таблиц к задаче классификации отдельных слов документа по параметрам описывающего их прямоугольника (далее токена): высоте, ширине и положению на странице. Такое решение автоматически будет мультиязычным и позволит обрабатывать все страницы многостраничного документа независимо. Далее токены, относящиеся к таблицам, будут обозначены как таблицы, а токены, не относящиеся к таблицам как текст. Координаты всех точек на странице нормированы таким образом, чтобы быть в пределах [0, 1]. Тогда сама страница всегда будет размером (1, 1).

Для обучения модели мы создали 10000 синтетических документов размером в 1 страницу, в которых рандомизировали все необходимые параметры таблиц. 90% синтетических документов содержат и текст, и таблицы. Мы разметили токены по размерам шрифта. Для каждого документа токены таблицы, текста и подписей к таблицам имеют разные размеры (a_1, a_2, a_3) в рамках одной страницы, но между страницами набор чисел (a_1, a_2, a_3) выбирается случайно. Кроме синтетических данных мы использовали 2008 размеченных вручную страниц из реальных документов. Из имеющихся данных мы собрали 3 выборки таким образом, чтобы доля токенов таблиц была примерно равна такой у реальных данных, то есть около 7%. Обучающая выборка содержит реальные и синтетические данные в соотношении 1:1, валидационная выборка 2:1, а тестовая выборка содержит только реальные данные.

	LogReg	DecTree	SVM	CatBoost	RandForest
Precision	0.73	0.80	0.59	0.87	0.88
Recall	0.55	0.81	0.69	0.84	0.86
F_1	0.63	0.80	0.64	0.85	0.87

Таблица 1. Сравнение моделей на классификации линий

Поскольку реальные форматированные документы состоят из строк, для каждого токена мы установили его принадлежность к строке документа (далее линии) и тем самым свели исходную задачу к классификации линий. На основании имеющихся данных мы использовали набор признаков для каждой линии, по которым ее можно было бы классифицировать: число токенов на странице, к которой принадлежит линия, число токенов в линии, отступ до левого края страницы, отступ до правого края страницы, положение линии по оси ординат, расстояние до предыдущей линии, расстояние до следующей линии, сумма ширин токенов в линии, ширина линии от левого до правого края, доля линии, заполненная токенами.

В качестве модели мы выбрали случайный лес, который показал себя лучше других классических моделей на полученных признаках на тестовой выборке при классификации линий до настройки параметров.

Постпроцессинг включал в себя подсчет взвешенного среднего вероятности p_i с соседними линиями с учетом расстояния до них по оси ординат d:

$$p_i = \frac{\sum_{k=i-2}^{i+2} p_k w_k}{\sum_{k=i-2}^{i+2} w_k}, w_k = a^{||k-i|-2|}(1 - \frac{d_k}{\sqrt{2}})^b, a = 1.12, b = 19.4$$

Для первых двух линий и последних двух линий формула не применяется. Цель усреднения вероятностей с соседями в том, чтобы снизить количество ложноотрицательных и ложноположительных одиночных ошибок, связанных с недостатками алгоритма принадлежности токена в линии. Например, если после токена в таблице стоит перенос строки, то следующий токен уже не будет считаться стоящим с ним в одной линии, даже если их координаты по оси ординат совпадают. Но очевидно, что соседние линии должны оказывать меньшее влияние на конечную вероятность, чем вероятность самой линии. Множитель $w^{||k-i|-2|}$ как раз решает эту проблему: вектор вероятностей линии и ее соседей p усредняется с весами $(1, a, a^2, a, 1)$. Второй множитель $(1 - \frac{d_k}{\sqrt{2}})^b$ позволяет учесть расстояния между линиями: более далеко отстоящие линии должны вносить значительно меньший вклад. Множитель сконструирован таким образом, чтобы решить проблему чтения документа не по порядку: не всегда подряд идущие линии в списке действительно находятся рядом друг с другом на странице. Функция от d должна достаточно быстро убывать, но экономить машинное время, поэтому была выбрана степенная функция с подбираемым показателем.

Расстояние нормировано на $\sqrt{2}$, так как это наибольшее возможное расстояние между точками на странице размером (1, 1).

Порог классификации равен 0.387. Так же все линии дополнительно классифицировались с учетом отношений расстояния до предыдущей и последующей: текущей линии i присваивалась метка от предыдущей линии, пока выполняется критерий: $\frac{d_{i+1}}{d_{i-1}} > 0.97$. Все параметры модели и постпроцессинга подбирались с помощью библиотеки hyperopt.

max_depth	**min_samples_leaf**	**n_estimators**
20	2	325

Таблица 2. Параметры случайного леса после тюнинга

Мы провели сравнение с другими решениями: PDF Plumber [2] и CascadeNet [3]. Скорость работы всех решений, кроме CascadeNet, проверялась на одной и той же машине, CascadeNet дополнительно потребовал подключения GPU. Предложенный метод показал лучшее качество и скорость работы, чем аналогичные методы распознавания таблиц.

	PDF Plumber	**CascadeNet**	**Предложенный метод**
Precision	0.29	0.88	0.83
Recall	0.62	0.85	0.93
F_1	0.39	0.87	0.88
t_{av}, с/стр	0.16	0.88	0.02

Таблица 3. Метрики качества и среднее время работы t_{av}

[1] *Bakhteev O., Khazov A.* Author masking using sequence-to-sequence models: Notebook for PAN at CLEF 2017

[2] *Singer-Vine J., Jain S.* https://github.com/jsvine/pdfplumber

[3] *Prasad D., Gadpal A., Kapadni K., Visave M., Sultanpure K.* CascadeTabNet: An approach for end to end table detection and structure recognition from image-based documents, CoRR, 2020

Table extraction from formatted documents

Kopanichuk Ilia[1]⋆ kopanichuk@ap-team.ru
Ochneva Inga[1] ochneva@ap-team.ru
Ogaltsov Aleksadr[1] ogaltsov@ap-team.ru
Kaprielova Mariam[1,3] kaprielova@ap-team.ru
Kildyakov Alexander[1] kildyakov@ap-team.ru
Finogeev Evgeny[1] finogeev@ap-team.ru
Chekhovich Yury[1,3] chehovich@ap-team.ru
[1]Moscow, Antiplagiat
[2]Moscow, Moscow Institute of Physics and Technology
[3]Moscow, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences

When searching for a plagiarism in written works, a significant problem is the recognition of individual structural elements of the document: tables, illustrations, contents, and bibliography. Without a precise definition of the structural element to which a particular text fragment belongs, a differentiated approach to identifying plagiarism is impossible. The number of false positives and false negatives increases, as well as the number of ways to mask plagiarism by unscrupulous users [1]. The method of table extraction will be presented in this report.

We have reduced the problem of table extraction to the problem of classifying individual words of a document according to the parameters of the rectangle describing them (hereinafter referred to as a token): height, width and position on the page. This solution would automatically be multilingual and allow to process all pages of a multipage document independently. Further, tokens related to tables will be labeled as tables, and tokens not related to tables as text. The coordinates of all points on the page are normalized to be within [0, 1]. Then the page itself will always be of size (1, 1).

To train the model, we created 10,000 one-page synthetic documents in which we randomized all the required table parameters. 90% of the synthetic documents contain both text and tables. We marked the tokens by font size. For each document, table tokens, text, and table captions have different sizes (a_1, a_2, a_3) within one page, but for different pages a set of numbers (a_1, a_2, a_3) is selected at random. In addition to synthetic data, we used 2008 manually labeled pages from real documents. From the available data, we collected 3 datasets so that the share of table tokens was

	LogReg	**DecTree**	**SVM**	**CatBoost**	**RandForest**
Precision	0.73	0.80	0.59	0.87	0.88
Recall	0.55	0.81	0.69	0.84	0.86
F_1	0.63	0.80	0.64	0.85	0.87

Table 1. Model comparison by a classification of lines

max_depth	min_samples_leaf	n_estimators
20	2	325

Table 2. Random forest parameters after tuning

about the same as the real data, that is, about 7%. The training set contains real and synthetic data in a 1:1 ratio, the validation set 2:1, and the test set contains only real data.

We chose a random forest as a model, which showed itself better than other classical models on the generated features on the test sample when classifying lines before any tuning of the parameters.

We also applied postprocessing, which includes calculating a weighted average probability p_i with neighboring lines, taking into account the distance to them along the ordinate axis d:

$$p_i = \frac{\sum_{k=i-2}^{i+2} p_k w_k}{\sum_{k=i-2}^{i+2} w_k}, w_k = a^{||k-i|-2|}(1 - \frac{d_k}{\sqrt{2}})^b, a = 1.12, b = 19.4$$

For the first two lines and the last two lines, the formula does not apply. The purpose of averaging probabilities with neighbors is to reduce the number of false negatives and false positives of single errors, associated with the shortcomings of the token belonging algorithm in the line. For example, if there is a line break after a token in the table, the next token will no longer be considered to be in the same line with it, even if their coordinates on the ordinate axis coincide. But obviously, neighboring lines should have less effect on the final probability than the probability of the line itself. Multiplier $w^{||k-i|-2|}$ just solves this problem: the probability vector of the line and its neighbors p is averaged with weights $(1, a, a^2, a, 1)$. The second multiplier $(1 - \frac{d_k}{\sqrt{2}})^b$ allows you to account for distances between lines: lines that are farther apart should contribute much less. This multiplier is designed to solve the problem of reading the document out of order: not always the neighboring lines in a list are actually next to each other on the page. The function of d must decrease fast enough, but save machine time, so a power function with a selectable power was chosen. The distance is normalized to $\sqrt{2}$, since this is the largest possible distance between points on a page of size (1, 1).

	PDF Plumber	CascadeNet	Presented solution
Precision	0.29	0.88	0.83
Recall	0.62	0.85	0.93
F_1	0.39	0.87	0.88
t_{av}, s/page	0.16	0.88	0.02

Table 3. Metrics and average time per page t_{av}

The classification threshold is 0.387. Also, all lines were additionally classified taking into account the distance ratio between the previous and the next line: the current line i was assigned a label from the previous line, as long as the criterion: $\frac{d_{i+1}}{d_{i-1}} > 0.97$ is fulfilled. All parameters of the model and postprocessing were chosen using the hyperopt library.

We made a comparison with another solutions: PDF Plumber [2] and CascadeNet [3]. The working time of all solutions has been tested on the same hardware, except CascadeNet. We have to use GPU to launch CascadeNet. The presented solution showed better quality and speed than similar methods of table recognition.

[1] *Bakhteev O., Khazov A.* Author masking using sequence-to-sequence models: Notebook for PAN at CLEF 2017

[2] *Singer-Vine J., Jain S.* https://github.com/jsvine/pdfplumber

[3] *Prasad D., Gadpal A., Kapadni K., Visave M., Sultanpure K.* CascadeTabNet: An approach for end to end table detection and structure recognition from image-based documents, CoRR, 2020

Методы поиска почти-дубликатов рукописных документов в больших коллекциях текстов

Бахтеев Олег Юрьевич[1,2] bakhteev@ap-team.ru
Грабовой Андрей Валериевич[1,3]⋆ grabovoy@ap-team.ru
Каприелова Мариам Семеновна[1,2] kaprielova@ap-team.ru
Кильдяков Александр Сергеевич[1] kildyakov@ap-team.ru
Сейил Темирлан Батырбекулы[1] seilov@ap-team.ru
Финогеев Евгений Леонидович[1] finogeev@ap-team.ru
Чехович Юрий Викторович[1,2] chehovich@ap-team.ru

[1]Москва, Антиплагиат
[2]Москва, Федеральный исследовательский центр «Информатика и управление» РАН
[3]Москва, Московский физико-технический институт (национальный исследовательский университет)

В работе рассматривается задача поиска почти-дубликатов текстов школьных сочинений в больших коллекциях данных. Предпосылками к решению данной задачи является возможность школьников применять для написания выпускных сочинений заранее заготовленные тексты, в том числе полученные из открытых коллекций школьных сочинений. Актуальность задачи подтверждается работами [1, 2], посвященными анализу нарушений при написании академических испытаний, а также частичному переходу школьного образования на удаленный режим.

Задача поиска почти-дубликатов рассматривается как задача информационного поиска, где сочинению ставится в соответствие заимствованный текст из коллекции. В рамках рассматриваемой задачи сочинение представляется набором изображений рукописного текста, написанного автором, в то время как документы из коллекции представимы в виде машиночитаемых текстов.

В данной работе сравниваются два подхода к поиску почти-дубликатов: поиск на основе методов глубокого обучения и поиск на основе анализа последовательностей длин извлекаемого текста. Поиск на основе методов глубокого обучения использует нейросетевую модель, оптимизация которой производится в режиме обучения с учителем. Работа подхода на основе глубокого обучения состоит из двух этапов. На первом этапе производится распознавание рукописного текста. На втором этапе производится разбиение полученного текста на биграммы и их поиск в индексе коллекции. Второй метод [3, 4] предполагает рассмотрение текста, находящегося в сканах школьных сочинений, как последовательности однородных характеристик текста, например, длин обнаруженных в тексте слов. Производится выделение слов из изображения без дальнейшего его распознавания. По выделенным словам строится последовательность нормированных длин слов, которая является инвариантной для рукописных и машиночитаемых вариантов написания текста. В работе сравнивается качество поиска на примере двух выборок: выборки сочинений [4], написанных на бланк-

ках специального вида, соответствующих бланкам государственного экзамена, и выборки сочинений, написанных на различных видах бланков и тетрадных листах. В качестве коллекций для поиска почти-дубликатов выступает подвыборка открытой коллекции текстов Тайга [5].

Результаты эксперимента демонстрируют применимость обоих методов к рассмотренной задаче. Показано, что нейросетевая модель является более устойчивой к неоднородности данных, и в частности лучше справляется с неоднородностью подложки сочинения, а также освещенности и качества сканов. В то же время, при хорошем качестве сканирования изображений и использовании стандартизированных бланков сочинений, оба метода показывают приемлемое качество поиска.

Работа поддержана грантом РФФИ No 19-29-14100.

[1] *Ma H. J., Wan G., Lu E. Y.* Digital cheating and plagiarism in schools // Theory Into Practice, 2008. Vol. 47. No 3. Pp. 197–203.

[2] *Wrigley S.* Avoiding “de-plagiarism”: Exploring the affordances of handwriting in the essay-writing process // Active Learning in Higher Education, 2019. Vol. 20. No 2. Pp.167–179

[3] *Бахтеев О. Ю., Кузнецова Р. В., Хазов А. В., Огальцов А. В., Сафин К. Ф., Горленко Т. А., Суворова М. А., Ивахненко А. А., Чехович Ю. В., Моттль В. В.* Поиск почти-дубликатов в рукописных текстах школьных сочинений // Интеллектуализация обработки информации: Тезисы докладов 13-й Международной конференции, Москва, 2020.

[4] *Bakhteev O., Kuznetsova R., Khazov A., Ogaltsov A., Safin K., Gorlenko T., Suvorova M., Ivahnenko A., Botov P., Chekhovich Y., Mottl V.* Near-duplicate handwritten document detection without text recognition // Computational Linguistics and Intellectual Technologies, is. 20, 2021.

[5] *Shavrina T., Shapovalova O.* To the methodology of corpus construction for machine learning: “taiga” syntax tree corpus and parser //Proceedings of the “Corpora”: 78-84, 2017.

Methods of near-duplicate handwritten document search in large collections of texts

Bakhteev Oleg[1,2] bakhteev@ap-team.ru
Grabovoy Andrey[1,3]⋆ grabovoy@ap-team.ru
Kaprielova Mariam[1,2] kaprielova@ap-team.ru
Kildyakov Aleksandr[1] kildyakov@ap-team.ru
Seyil Temirlan[1] seilov@ap-team.ru
Finogeev Evgeny[1] finogeev@ap-team.ru
Chekhovich Yury[1,2] chehovich@ap-team.ru

[1]Moscow, Antiplagiat
[2]Moscow, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences
[3]Moscow, Moscow Institute of Physics and Technology

This paper considers the problem for searching of near-duplicate of school essays in large data collections. The problem is conditioned by the availability of school essay collections, which can be used for writing essays and essays by school students. The relevance of the task is confirmed by the works [1, 2] devoted to the analysis of violations in writing academic tests, as well as the partial transition of school education to remote mode.

The near-duplicate search problem is considered as an information search problem, where an essay is matched with a reused text from the collection. In the framework of the problem, an essay is represented as a set of scanned images of handwritten text, while a collection of essays is represented by machine-readable texts.

This paper compares two approaches to near-duplicate search: a deep learning-based search and an approach based on text length sequence analysis. The deep learning-based search uses a neural network model that is optimized in a supervised manner. This approach consists of two stages. The first stage performs Optical Character Recognition (OCR). The second step is bigram search in the collection. The second method [3, 4] involves treating the text in scans of school essays as a sequence of features extracted from the text. The method extracts words from the image without further recognition. Based on the extracted words, a sequence of normalized word lengths is constructed, which is invariant for handwritten and machine-readable variants of the text. This paper compares the quality of the search with two datasets: a dataset of essays [4] written on special forms corresponding to the forms of the state exam, and a dataset of essays written on various types of forms and notebook sheets. A subsample of the Taiga open corpus [5] is used as a search collection.

The results of the experiment demonstrate the applicability of both methods to the considered problem. It is shown that the deep learning-based approach is more robust to data heterogeneity, and in particular, it copes better with the heterogeneity

of the essay layout, as well as the general quality of scans. At the same time, with good image scanning quality and the use of standardized essay forms, both methods show an acceptable search quality.

This research is funded by RFBR, grant 19-29-14100.

[1] *Ma H. J., Wan G., Lu E. Y.* Digital cheating and plagiarism in schools // Theory Into Practice, 2008. Vol. 47. No 3. Pp. 197–203.

[2] *Wrigley S.* Avoiding "de-plagiarism": Exploring the affordances of handwriting in the essay-writing process // Active Learning in Higher Education, 2019. Vol. 20. No 2. Pp.167–179

[3] *Bakhteev O., Kuznetsova R., Khazov A., Ogaltsov A., Safin K., Gorlenko T., Suvorova M., Ivakhnenko A., Chekhovich Y., Mottl V.* Near-duplicate detection in handwritten school essays //ntelligent Data Processing: Theory and Applications: Book of abstract of the 13th International Conference, Moscow, 2020.

[4] *Bakhteev O., Kuznetsova R., Khazov A., Ogaltsov A., Safin K., Gorlenko T., Suvorova M., Ivahnenko A., Botov P., Chekhovich Y., Mottl V.* Near-duplicate handwritten document detection without text recognition // Computational Linguistics and Intellectual Technologies, is. 20, 2021.

[5] *Shavrina T., Shapovalova O.* To the methodology of corpus construction for machine learning: "taiga" syntax tree corpus and parser //Proceedings of the "Corpora": 78-84, 2017.

Инкрементное обучение тематических моделей с аддитивной регуляризацией для выявления трендовых научных тем

Герасименко Николай Александрович[1,2]⋆ nikgerasimenko@gmail.com
Чернявский Александр Сергеевич[2,3] alschernyavskiy@gmail.com
Никифорова Мария Андреевна[2,3] labenzom@gmail.com
Никитин Максим Дмитриевич[2] mdnikitin@sberbank.ru
Воронцов Константин Вячеславович[1] vokov@forecsys.ru
[1]Москва, ФИЦ ИУ РАН
[2]Москва, ПАО Сбербанк
[3]Москва, НИУ «Высшая школа экономики»

Стремительный рост числа научных публикаций, интенсивное появление новых направлений и подходов ставит перед научным сообществом задачу своевременного выявления трендов. Под трендом понимается семантически однородная тема, которая характеризуется устойчивым во времени лексическим ядром и резким, зачастую экспоненциальным, ростом числа публикаций [1]. Примерами трендов в машинном обучении являются «LSTM», «deep learning», «word2vec», «BERT», «fake news detection».

Для выделения трендовых тем в потоке научных публикаций в реальном времени мы используем инкрементные методы вероятностного тематического моделирования. При помощи подхода, основанного на аддитивной регуляризации тематических моделей (ARTM) [2], удалось превзойти результаты ряда классических и нейросетевых моделей, использовавшихся ранее для выявления новых трендов. Для оценки качества выявления трендов мы вручную сформировали и сделали общедоступным датасет AITD, состоящий из 91 тренда по тематике машинного обучения.

Эксперименты по выявлению трендов производились на коллекции из 73959 статей, опубликованных с 2000 по 2021 год на конференциях по машинному обучению с h-индексом, превышающим 100. Валидационный датасет охватывает тренды в области машинного обучения и искусственного интеллекта 2009-2021 годов, каждый из которых характеризуются набором из не менее, чем 10 ключевых статей и 5 ключевых терминов. Обучение моделей производилось без учителя, а валидационная разметка использовалась только для финальной оценки качества.

Вероятностная тематическая модель коллекции текстовых документов задаётся двумя матрицами, Φ размера $W \times T$ и Θ размера $T \times D$, где W — конечное множество (словарь) токенов (слов), D — конечное множество (коллекция) документов, T — множество тем. Каждый столбец матрицы Φ описывает тему $t \in T$ дискретным распределением вероятностей слов $p(w|t)$. Каждый столбец матрицы Θ описывает документ $d \in D$ дискретным распределением вероятностей тем $p(t|d)$.

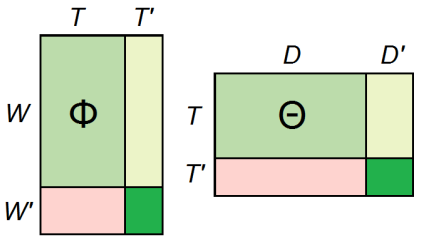


Рис. 1. Блочная структура $W \times T$-матрицы Φ и $T \times D$-матрицы Θ в инкрементной тематической модели. Нулевые блоки: новые токены в старых темах $W' \times T$, новые темы в старых документах $T' \times D$. Ненулевые блоки: новые токены в новых темах $W' \times T'$, новые темы в новых документах $T' \times D'$.

Чтобы отслеживать появление новых тем, предлагается обучать отдельные модели для каждого временного интервала. При поступлении новой порции документов D' матрица Φ предыдущей модели используется в качестве начального приближения; словарь пополняется новыми терминами W' и могут образоваться новые темы T'. Предполагается, что новая лексика, появившаяся в новых документах, относится преимущественно к новым темам, Рис. 1. Дополнительные ограничения на тематическую модель накладываются в рамках подхода аддитивной регуляризации ARTM с использованием библиотеки BigARTM [3] с открытым исходным кодом. В частности, используются регуляризатор декорреляции тем как столбцов матрицы Φ и регуляризатор разреживания тематических векторных представлений документов как столбцов матрицы Θ.

Для определения числа новых тем на каждом временном интервале оценивается относительное приращение числа токенов в словаре на текущем временном шаге. При этом предполагается, что число уникальных токенов в каждой теме примерно одинаково.

На выходе модели каждой выделенной теме соответствуют ранжированные списки ключевых для темы документов D_{topic} и слов W_{topic}. Предлагаемый нами валидационный датасет AITD также состоит из множества трендов, которым соответствуют ранжированные списки ключевых документов D_{trend} и терминов W_{trend}, а также названия трендов S_{trend}. Чтобы сопоставить результаты модели реальным трендам, вычисляются три метрики Recall@k:

$$\text{XRecall@k} = \frac{|X_{\text{topic}}[:\text{k}] \cap X_{\text{trend}}|}{k} \tag{1}$$

где $X[:\text{k}]$ — первые k элементов списка X, а X принимает значения W, D, S. Для подсчета метрики для документов, слов и названий используются разные значения k, которые обозначаются как k_D, k_W и $k_S \leqslant k_W$ соответственно.

Мы провели серию экспериментов, где рассмотрели вероятностные тематические модели PLSA, LDA и ARTM, а также нейросетевой подход BERTopic.

Мы сравнили наше решение с вышеперечисленными на основе трех конфигураций, включающих в себя следующие параметры:

— *Config1*: DRecall@k > 0.1
— *Config2*: WRecall@k > 0.3 и SRecall@k > 0
— *Config3*: DRecall@k > 0.1 , WRecall@k > 0.3 и SRecall@k > 0

Здесь *Config1* соответствует сопоставлению извлеченных тем и трендов по документам, *Config2* — только по ключевым словам, а *Config3* объединяет в себе две предыдущие опции.

Наилучшие результаты показала модель ARTMi – модель ARTM с регуляризатором декоррелирования Φ и регуляризатором разреживания Θ, обучающаяся инкрементно. ARTMi детектирует правильные темы достаточно полно и быстро даже в наиболее сложной конфигурации *Config3*, хотя в некоторых случаях может извлекать суммарно меньше трендов. В конфигурации *Config1*, основной целью которой является правильное разделение документов по темам, ARTMi извлекает почти половину трендов за первые два месяца.

Таким образом, модель подходит для раннего выявления научных трендов. Также стоит отметить, что модель ARTMi способна выявлять тренды непосредственно в интервале их возникновения для *Config1*. Это связано с тем, что некоторые из трендов, относящиеся к типу «задача», не имеют конкретного первого документа.

[1] *Kontostathis A., Galitsky M. L., Pottenger M. W., Roy S., Phelps J. D.* A Survey of emerging trend detection in textual data mining // Springer New York, 2004. — Pp. 185–224.

[2] *Kochedykov D., Apishev M., Golitsyn L., Vorontsov K.* Fast and Modular Regularized Topic Modelling // Proceeding of The 21-st Conference Of FRUCT (Finnish-Russian University Cooperation in Telecommunications) Association. The seminar on Intelligence, Social Media and Web (ISMW). Helsinki, Finland, November 6–10, 2017. Pp.182–193.

[3] *Vorontsov K., Frei O., Apishev M., Romov P., and Dudarenko M.* BigARTM: Open source library for regularized multimodal topic modeling of large collections // AIST, Springer International Publishing, 2015. — Pp. 370–381.

Incremental Learning of Topic Models with Additive Regularization for Scientific Trend Topics Detection

Gerasimenko Nikolai[1,2]⋆ nikgerasimenko@gmail.com
Chernyavskiy Alexander[2,3] alschernyavskiy@gmail.com
Nikiforova Maria[2,3] labenzom@gmail.com
Nikitin Maxim[2] mdnikitin@sberbank.ru
Vorontsov Konstantin[1] vokov@forecsys.ru
[1]Moscow, FRC CSC RAS
[2]Moscow, Sberbank
[3]Moscow, Higher School of Economics

The rapid growth in the number of scientific publications, the intensive emergence of new directions and approaches poses a challenge to the scientific community to identify trends in a timely and automatic manner. By trend we mean a semantically homogeneous theme that is characterized by a steadily time lexical kernel and a sharp, often exponential, increase in the number of publications [1]. In addition, a trend is often characterized by a main term, such as the name of a problem, a theory, or a method. Examples of trends in machine learning are: "LSTM", "deep learning" "word2vec", "BERT", "fake news detection".

In order to extract trend themes in a stream of scientific publications, we use incremental methods of probabilistic topic modelling. To assess quality, we manually formed and made publicly available a dataset of 91 trends named Artificial Intelligence Trends Dataset (AITD), in which each trend is characterized by a set of at least 10 key articles and 5 key terms.

Trend extraction experiments were carried out on a collection of 73959 articles published between 2000 and 2021 at an h-index over 100 machine-learning conferences. The publication date of each article is known. The models were taught entirely unsupervised, and validation dataset were used only for final quality assessments.

The probabilistic topic model of a text documents collection outputs two matrices: Φ matrix of size $W \times T$ and Θ matrix of size $T \times D$, where W is a finite set (dictionary) of tokens (words), D is a finite set (collection) of documents, T is a set of topics. Each column of the Φ matrix describes a topic $t \in T$ by a discrete probability distribution over words $p(w|t)$. Each column of the Θ matrix describes a document $d \in D$ by a discrete probability distribution over topics $p(t|d)$. Additional restrictions on the topic model are imposed within the additive regularization (ARTM) approach [2] using the open source BigARTM [3] library. In this work, we apply the topic decorrelation regularizer to the columns of the Φ matrix and also we apply the sparse topical representations of documents regularizer to the columns of a Θ matrix.

To obtain real-time predictions and reduce training time, we suggest incremental training of the topic model. After a new batch of documents D' appears, the model

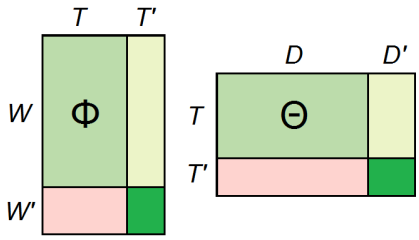


Fig. 1. Block structure of Φ matrix of size $W \times T$ and Θ matrix of size $T \times D$ in an incremental topic model. Zero blocks: new tokens in old topics $W' \times T$, new topics in old documents $T' \times D$. Non-zero blocks: new tokens in new topics $W' \times T'$, new topics in new documents $T' \times D'$.

considers a set of emerging new words W' and updates current topics T by adding new topics T', Figure 1. Generally, topic modeling approaches operate with matrices Φ and Θ representing word-topic and topic-document distributions respectively. We suggest an incremental update to each of them. So, we initialize the matrices Φ_{n+1} and Θ_{n+1} in the current step using the matrices Φ_n and Θ_n from the previous step.

To select the number of new topics, we propose a criterion based on the number of terms that have increased significantly since the last update of the model.

Let D_{trend} and W_{trend} be the labeled sets of documents and words associated with the given trend respectively. Apart from that, we consider "golden" set of topic names S_{trend} . Here, S_{trend} contains from one to three synonymous collocations, each of which can be used as the name of the trend. At the output of the model, each topic is represented by two ranked lists denoted as D_{topic} and W_{topic}. Also, we define $S_{\text{topic}} := W_{\text{topic}}$.

To perform matching, we firstly calculate three Recall@k based metrics:

$$\text{XRecall@k} = \frac{|X_{\text{topic}}[:\text{k}] \cap X_{\text{trend}}|}{k} \tag{1}$$

Here, $X[:\text{k}]$ denotes first k elements of the list X, and X is replaced with W, D or S. We use three different values of the parameter k for documents, words and topic names, which are denoted as k_D, k_W and $k_S \leqslant k_W$ respectively.

We matched the extracted topics with the labeled trend topics using three combinations of thresholds:

— *Config1*: DRecall@k > 0.1
— *Config2*: WRecall@k > 0.3 and SRecall@k > 0
— *Config3*: DRecall@k > 0.1 , WRecall@k > 0.3 and SRecall@k > 0

Here, the *Config1* option matches trends based on documents only, *Config2* is based on keywords only, and *Config3* is the joined option.

Our experiments demonstrate that the suggested ARTM-based approach named ARTMi outperforms the classic PLSA, LDA models and a neural approach based on BERT representations.

The best results was achived by ARTMi model is the ARTM model learning incrementally with a decorrelation Φ and sparse Θ regularization. ARTMi outperforms both BERTopic and LDA model. The ARTMi model generates the correct topics quickly enough even with the rigid configuration *Config3* compared to other topic models, although it can sometimes extract fewer trends in total. In the configuration *Config1*, when the main goal is to correctly divide documents by topic, ARTMi extracts almost half of the trends in the first two months. Thus, it is well suited for qualitative identification of trends in the problem of early detection.

[1] *Kontostathis A., Galitsky M. L., Pottenger M. W., Roy S., Phelps J. D.* A Survey of emerging trend detection in textual data mining // Springer New York, 2004. — Pp. 185–224.

[2] *Kochedykov D., Apishev M., Golitsyn L., Vorontsov K.* Fast and Modular Regularized Topic Modelling // Proceeding of The 21-st Conference Of FRUCT (Finnish-Russian University Cooperation in Telecommunications) Association. The seminar on Intelligence, Social Media and Web (ISMW). Helsinki, Finland, November 6–10, 2017. Pp.182–193.

[3] *Vorontsov K., Frei O., Apishev M., Romov P., and Dudarenko M.* BigARTM: Open source library for regularized multimodal topic modeling of large collections // AIST, Springer International Publishing, 2015. — Pp. 370–381.

Технология полуавтоматической суммаризации тематических подборок научных статей

Крыжановская Светлана Юрьевна[1]⋆ skryzhanovskaya@yandex.ru
Воронцов Константин Вячеславович[1,2] voron@forecsys.ru
[1]Москва, МГУ им. М.В. Ломоносова
[2]Москва, МФТИ

Целью автоматической суммаризации обычно является синтез краткого изложения подборки текстовых документов для быстрого понимания ключевых идей. Мы рассматриваем полуавтоматическую суммаризацию подборок научных публикаций, которая имеет другую цель — помочь пользователю реализовать свой авторский замысел. Обзоры, написанные по одной и той же подборке разными авторами и/или для разных целей (диссертация, отчёт, учебник) могут различаться существенно. В таких случаях пользователю нужен не готовый текст, а рекомендательный сервис, выполняющий рутинные операции информационного поиска, например, подбирающий релевантные фразы. Такие системы называются *автоматизированной авторской суммаризацией текстов* (machine aided human summarization, MAHS).

В данной работе предлагается процесс MAHS, основанный на решении серии задач машинного обучения [1].

Задача 1: *Построение сценария обзора.* Дана подборка текстов, требуется ранжировать их в том порядке, в котором они будут упоминаться в обзоре. Обучающая выборка может быть составлена автоматически по большой коллекции научных статей. Каждая статья порождает обучающий объект, в котором роль подборки выполняет список литературы, а обучающим ранжированием является последовательность ссылок в обзорной части статьи. Информативными признаками являются год публикации, её цитируемость, цитируемость её авторов, семантическая близость публикации к обзору, к его локальному контексту, и т. д.

Задача 2. *Ранжирование фраз-подсказок.* Для каждого документа из подборки (в порядке, определённом сценарием) система предлагает пользователю ранжированный список фраз-подсказок, из которых он может выбирать фразы для продолжения своего обзора. Существует множество разумных способов отбора и ранжирования фраз. Они реализуются в системе как функции ранжирования, называемые *суфлёрами.* В пользовательском интерфейсе суфлёр — это кнопка, нажатие которой перестраивает ранжированный список фраз. Например, *суфлёр аннотации (abstract prompter)* — это простейшая необучаемая функция, которая выдаёт фразы из аннотации статьи в их исходном порядке.

Задача 2.1: *Обучение экстрактивного суфлёра (extractive prompter).* Классические методы экстрактивной суммаризации основаны на выделении наиболее важных предложений в документе. Важность может определяться по-разному, в том числе относительно аспектов «актуальности», «новизны», «методов», «вы-

водов», «преимуществ», «недостатков», и т. д. Для обучения суфлёра по каждому из аспектов строится обучающая выборка документов, в каждом из которых разметчики выделяют фразы, релевантные аспекту.

Задача 2.2. *Обучение цитирующего суфлёра (citance prompter).* Особенностью научных публикаций является наличие цитирующих статей, которые отражают влияние исследуемой статьи на научное сообщество. Предлагается извлекать наиболее релевантные фрагменты, содержащие цитаты, и ранжировать их для составления реферата. Назовём такие фрагменты — *цитирующими фрагментами*. Задача локализации цитирующих фрагментов требует размеченной обучающей выборки вида «цитата, цитирующий фрагмент». Такую задачу можно переформулировать в терминах вопросно-ответного поиска в заданном фрагменте текста. При этом в качестве запроса рассматривать цитату, а в качестве текста для поиска ответа — текстовое окно, содержащее предложение со ссылкой. Для решения такого типа задач отлично подходят предобученные на большом текстовом корпусе модели на основе трансформера.

Задача 2.3. *Обучение ссылочного суфлёра (reference prompter).* Альтернативный подход заключается в том, чтобы выбрать из текста цитируемой статьи фразы, семантически близкие к локальным контекстам цитат в упоминающих статьях. Назовём такие фрагменты — *ссылочными фрагментами*. При этом точное определение границ цитирующих фрагментов уже не требуется, а цитирующий суфлёр становится разновидностью экстрактивного. Для обучения такого суфлёра требуется размеченная обучающая выборки вида «локальный контекст цитаты, ссылочный фрагмент». Задачу обучения можно сформулировать, как задачу бинарной классификации вхождения каждого предложения статьи в какой-либо ссылочный фрагмент. Для оценки семантической схожести фрагментов подходят модели градиентного бустинга с фиксированным набором текстовых признаков, а также предобученные текстовые модели на основе трансформера.

Задача 3: *Оценивание качества системы суфлёров.* Обучающая и валидационная выборка составляются автоматически по большой коллекции статей, аналогично первой задаче. В роли подборки выступает список литературы, в роли «идеального» обзора — объединение обзорных разделов статьи. Для каждого предложения подбирается суфлёр, дающий самую близкую фразу среди первых k позиций списка поисковой выдачи по метрике качества в сравнении с эталонным рефератом. Легко проводится аналогия с пользователем, последовательно просматривающим выдачу для поиска наилучшей фразы для продолжения. Оценкой качества работы суфлёров является средняя позиция наиболее релевантных фраз в их списках выдачи. Заодно оценивается полезность каждого суфлёра как число его фраз, которые вошли в обзор.

Для проведения экспериментов были построены обучающая и валидационная выборка из обзорных разделов коллекции S2ORC [3]. В качестве модели построения обзорной части быда выбрана модель на основе градиентного бустин-

га. Коэффициент корреляции Кенделла полученной модели $\tau = 0.48$ (при ранжировании по году публикации $\tau = 0.1$). Для обучения и оценивания цитирующего и ссылочного суфлёров использовалась коллекция CL-SciSumm 2018 [2]. Наилучшее качество показала предобученная на большой коллекции научных статей модель SciBERT [4], превзойдя наилучшие результаты при автоматическом составлении реферата на CL-SciSumm 2018. В качестве экстрактивного суфлёра использовалась модель TextRank.

Автоматическая оценка качества системы из всех суфлёров проводилась для каждого суфлёра по отдельности, а также для всех суфлёров в совокупности. Результаты показывают, что в совокупности суфлёры дают прирост по метрике ROUGE на 7 процентов (Таб. 1). При выборе из первых трёх позиций также удаётся добиться повышения качества ещё на 5 процентов.

	$top-1$	$top-3$
Abstract	0,11	0,13
Extractive	0,08	0,11
Citance	0,16	0,22
Reference	0,13	0,15
All	**0,21**	**0,26**

Таблица 1. $ROUGE_2(f_1)$ системы MAHS на подборке S2ORC.

В дальнейшем планируется оценивать качество ранжирования каждого суфлёра по логам поисково-рекомендательной системы SciSearch.ai как среднюю позицию тех его фраз, которые пользователи отбирали для обзора.

В заключении заметим, что предлагаемый способ полуавтоматической суммаризации может также рассматриваться как способ *нелинейного чтения*, когда перед пользователем стоит задача не только разобраться в мало знакомой для него области по обширной тематической коллекции, но и одновременно произвести информационный продукт в виде обзора.

Работа поддержана грантом РФФИ No. 20-07-00936.

[1] *А.В.Власов.* Методы полуавтоматической суммаризации подборок научных статей // Магистерская диссертация, МФТИ, 2020.
http://www.MachineLearning.ru/wiki/images/6/6d/Vlasov2020MSThesis.pdf

[2] *K.Jaidka, M.Yasunga, M.Chandrasekaran, D.Radev, M.-Y.Kan.* The CL-SciSumm Shared Task 2018: Results and Key Insights. 2018.

[3] *K.Lo, L.L.Wang, M.Neumann, R.M.Kinney, D.S.Weld.* S2ORC: The Semantic Scholar Open Research Corpus, ACL, 2020.

[4] *I.Beltagy, K.Lo, A.Cohan.* SciBERT: A Pretrained Language Model for Scientific Text, EMNLP, 2019.

Machine Aided Human Summarization of scientific articles collections

Kryzhanovskaya Svetlana[1]⋆ skryzhanovskaya@yandex.ru
Vorontsov Konstantin[1,2] voron@forecsys.ru
[1]Moscow, Lomonosov Moscow State University
[2]Moscow, Moscow Institute of Physics and Technology

The goal of automatic summarization is usually to synthesize a concise analogue of a set of documents for quickly understanding of key ideas. We are considering a semi-automatic summarization of a collection of scientific publications with another purpose: to help the user realize his authorial intent. Reviews written on the same selection by different authors and/or for different purposes (thesis, report, textbook) can differ significantly. In such cases, the user does not need a ready-made text, but rather a recommender service which performs routine information retrieval operations, such as selecting relevant phrases. Such systems are called *machine aided human summarization (MAHS)*.

This paper proposes a MAHS system based on a series of machine learning tasks [1].

Task 1: *Review scenario generation.* A collection of texts is given, and we need to rank them in the order to mention in the review. A training sample can be automatically generated from a large collection of scientific articles. Each article generates a training object in which the list of references acts as a collection, and the learning ranking is the sequence of references to these publications in the review part of the article. Informative features are the year of the publication, its citation, the citation of its authors, the semantic proximity of the publication to the review, to its local context, etc.

Task 2: *Clue phrases ranking.* For each document in the collection (in the order defined by the scenario), the system offers the user a ranked list of clue phrases from which to select phrases to continue the review. There are many reasonable ways of selecting and ranking phrases, which are implemented in the system as ranking functions, called *prompters*. In the user interface, a prompter is a button which, when clicked, rearranges the ranked list of phrases. For example, an *abstract prompter* is a simple, unlearnable function which rearranges the phrases from the abstract in their original order.

Task 2.1: *Training the extractive prompter.* Classical extractive summarization methods are based on extracting the most important sentences in a source document. Importance can be defined in various ways with respect to aspects of "relevance", "newness", "approach", "data", "experiments", "results", "conclusions", "advantages", "faults", etc. To train a prompter for each aspect, a training sample of documents is constructed, in which assessors select phrases relevant to the aspect.

Task 2.2: *Training the citation prompter.* The peculiarity of scientific publications is the presence of citation articles, which reflect the influence of the ar-

ticle on the scientific community. We propose to extract the most relevant spans containing citations — *citation spans*, and rank them to compose the summary. The problem of citation spans localization requires a marked training set of the samples ⟨citation, citation span⟩. This problem can be reformulated in terms of question answering search in a given text fragment, considering the citation as a query, and the text box containing the citation as the search text. The transformer-based models pre-trained on a large text corpus work excellent for solving this type of problem.

Task 2.3. *Training the reference prompter.* An alternative approach is to select phrases from the text of the cited article that are semantically close to the local contexts of the citations in the citing articles. Let us call such spans — *reference spans*. In this case, the exact definition of the boundaries of citation spans is no longer required, and the reference prompter becomes a kind of extractive prompter. A marked training set of the samples ⟨local citation context, reference span⟩ is required for training. The task can be formulated as a binary classification problem for the occurrence of each sentence of an article in some reference span. To evaluate the semantic similarity of the fragments, gradient boosting models with a fixed set of textual features, as well as pre-trained transformer-based models, are suitable.

Task 3: *Quality evaluation.* A validation sample is collected automatically from a large collection of scientific articles, similar to the first problem. Articles from the references act as the collection, and a union of the review sections as the "ideal"summary. For each summary sentence, a prompter that gives the closest phrase among the first k positions of the prompter ranked list is chosen. The quality score is the average position of the most relevant phrases from the prompters. At the same time, the usefulness of each prompter is evaluated as the number of its phrases that were included into the review.

For the experiments, the training and validation sets from the review sections of the S2ORC [3] collection were constructed. A gradient boosting model was chosen for review scenario generation. The Kendell correlation coefficient of the resulting model is $\tau = 0.48$ (ranked by year of publication is $\tau = 0.1$). The CL-SciSumm 2018 [2] collection was used to train and evaluate citation and reference prompters. The SciBERT [4] model pre-trained on the large collection of scientific articles shows the best quality, outperforming the best results for automatic abstract compilation on CL-SciSumm 2018. The TextRank model was used as an extractive prompter.

Automatic quality evaluation of the whole MAHS was performed for each prompter individually, as well as for all prompters in together. The results show that a combination of the prompters gives a 7 percent increase in ROUGE (Tab.1). When selecting from the first 3 positions, it is also possible to achieve an increase in quality by another 5 percent.

	$top-1$	$top-3$
Abstract	0,11	0,13
Extractive	0,08	0,11
Citance	0,16	0,22
Reference	0,13	0,15
All	**0,21**	**0,26**

Table 1. $ROUGE_2(f1)$ on the S2ORC validation subset of the MAHS.

In the future, the quality of each prompter's ranking will be evaluated in the SciSearch.ai search-recommendation system as the average position of those phrases which users selected for review.

In conclusion, it should be noticed that the proposed method can also be seen as a way of *nonlinear reading*, when the user has the task not only to understand a little familiar area to him, but also to produce an information product in the form of a review simultaneously.

This research is funded by RFBR grant 20-07-00936.

[1] *A.V.Vlasov.* Methods of machine aided human summarization of scientific articles // Master's thesis, MIPT, Moscow, 2020.
http://www.MachineLearning.ru/wiki/images/6/6d/Vlasov2020MSThesis.pdf

[2] *K.Jaidka, M.Yasunga, M.Chandrasekaran, D.Radev, M.-Y.Kan.* The CL-SciSumm Shared Task 2018: Results and Key Insights. 2018.

[3] *K.Lo, L.L.Wang, M.Neumann, R.M.Kinney, D.S.Weld.* S2ORC: The Semantic Scholar Open Research Corpus, ACL, 2020.

[4] *I.Beltagy, K.Lo, A.Cohan.* SciBERT: A Pretrained Language Model for Scientific Text, EMNLP, 2019.

Разработка алгоритма определения расстояния между радиоустройствами на основе информации о состоянии канала связи и искусственных нейронных сетей

Астафьев Александр Владимирович[1]⋆ Alexandr.Astafiev@mail.ru

[1]Владимир, Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых

Системы внутреннего позиционирования – это программно-аппаратные реализации системы для позиционирования внутри зданий и сооружений, где спутниковые технологии, такие как GPS и ГЛОНАСС, не достигают заявленной точности. В последние годы, для организации позиционирования внутри помещений, всё чаще стали использовать алгоритмы, основанные на анализа цифрового отпечатка сигнала, и более подробную информацию физического уровня радиоустройств - информацию о состоянии канала связи. Формат информации о состоянии канала связи в беспроводных сетях WiFi4 и WiFi5 регламентируется стандартами IEEE 802.11n и IEEE 802.11ac. Согласно этих стандартов, высокоскоростной поток разделяется на 56 и 114 параллельных потоков меньшей скорости, из которых можно извлечь информацию о фазе и амплитуде сигнала [1].

Исходя из этого целью исследования является Разработка алгоритма определения расстояния между радиоустройствами на основе информации о состоянии канала связи и искусственных нейронных сетей для построения системы внутреннего позиционирования.

Для обучения нейронной сети был сформирован набор данных, состоящий из замеров информации о состоянии канала связи на расстоянии от 0.5 до 3 метров с шагом 0.5 метра. В обучающую выборку было отобрано по 3000 значений фаз измерений, из которых 20% составили выборку валидации. В качестве модели нейронной сети была использована полносвязная нейронная сеть с входным слоем из 56 нейронов. Обработка информации производилась с использованием трёх полносвязных слоёв по 256 нейронов в каждом со слоем прореживания (Dropout). Количество нейронов на выходном слое соответствует количеству наблюдаемых позиций и равно 6 с функцией активации SoftMax.

Оценка результатов работы нейронной сети предлагается производить по количеству правильных ответов (True positive), количеству неправильных ответов (False positive) и доли точных ответов. Результат работы нейронной сети на обучающей выборке приведен в таблице 1.

Для проверки адекватности работы предложенной нейронной сети был сформирован тестовый набор данных, состоящий из 600 измерений в каждой наблюдаемой позиции. Результат работы нейронной сети на тестовой выборке приведен в таблице 2.

В ходе экспериментальных исследований предложенной модели нейронной сети была получена точность позиционирования на обучающей выборке в диа-

Таблица 1. Результат работы нейронной сети на обучающей выборке

Расстояние, м.	0.5	1	1.5	2	2.5	3
True Positive	2992	2747	2948	2920	2950	2998
False Positive	8	253	52	80	50	2
Точность	99.73%	91.57%	98.27%	97.33%	98.33%	99.93%

Таблица 2. Результат работы нейронной сети на тестовой выборке

Расстояние, м.	0.5	1	1.5	2	2.5	3
True Positive	593	340	595	539	589	595
False Positive	7	186	7	19	1	5
Точность	98.83%	69%	98.83%	96.83%	99.83%	99.17%

пазоне от 91.73% до 99.93%. Средняя точность нейронной сети составила 97.53%. Точность на тестовой выборке варьировалась от 69% до 99.83%. Средняя точность составила 93.75%. Исходя из полученных результатов можно сделать вывод, что полносвязные нейронные сети способны измерять расстояния между радиоустройствами в наблюдаемых позициях с использованием информации о состоянии канала связи.

Работа поддержана грантом РНФ No. 21-71-00133.

[1] *Астафьев А. В., Титов Д. В., Жизняков А. Л., Демидов А. А.* Метод позиционирования мобильного устройства с использованием сенсорной сети BLE-маяков, аппроксимации значений уровней сигналов RSSI и искусственных нейронных сетей // Компьютерная оптика, Самара, 2021. Т. 45, № 2. С. 277–285.

Development of an algorithm for determining the distance between radio devices based on channel state information and artificial neural networks

Astafiev Alexandr[1]⋆ Alexandr.Astafiev@mail.ru

[1]Vladimir, Vladimir State University

Indoor positioning systems are software and hardware implementations of a system for positioning inside buildings and structures where satellite technologies such as GPS and GLONASS do not achieve the declared accuracy. In recent years, for the organization of positioning indoors, increasingly began to use algorithms based on the analysis of the digital signal fingerprint, and more detailed information on the physical layer of radio devices - channel state information. The format of channel state information in WiFi4 and WiFi5 wireless networks is regulated by the IEEE 802.11n and IEEE 802.11ac standards. According to these standards, a high-speed stream is divided into 56 and 114 parallel streams of lower speed, from which information about the phase and amplitude of the signal can be extracted [1].

Based on this, the purpose of the study is to develop an algorithm for determining the distance between radio devices based on channel state information and artificial neural networks to build an internal positioning system.

To train the neural network, a data set was formed, consisting of measurements of channel state information at a distance of 0.5 to 3 meters with a step of 0.5 meters. 3000 values of the measurement phases were selected into the training sample, of which 20% constituted the validation sample. A fully connected neural network with an input layer of 56 neurons was used as a neural network model. Information processing was carried out using three fully connected layers of 256 neurons each with a thinning layer (Dropout). The number of neurons in the output layer corresponds to the number of observed positions and is equal to 6 with the SoftMax activation function.

It is proposed to evaluate the results of the neural network by the number of correct answers (True positive), the number of incorrect answers (False positive) and the proportion of accurate answers. The result of the neural network on the training set is shown in Table 1.

Table 1. The result of the neural network on the training set

Distance, m.	0.5	1	1.5	2	2.5	3
True Positive	2992	2747	2948	2920	2950	2998
False Positive	8	253	52	80	50	2
Accuracy	99.73%	91.57%	98.27%	97.33%	98.33%	99.93%

To test the adequacy of the proposed neural network, a test data set was formed, consisting of 600 measurements in each observed position. The result of the neural network on the test sample is shown in Table 2.

Table 2. The result of the neural network on the test set

Distance, m.	0.5	1	1.5	2	2.5	3
True Positive	593	340	595	539	589	595
False Positive	7	186	7	19	1	5
Accuracy	98.83%	69%	98.83%	96.83%	99.83%	99.17%

In the course of experimental studies of the proposed neural network model, the positioning accuracy on the training set was obtained in the range from 91.73% to 99.93%. The average accuracy of the neural network was 97.53%. The accuracy on the test set ranged from 69% to 99.83%. The average accuracy was 93.75%. Based on the obtained results, we can conclude that fully connected neural networks are able to measure the distances between radio devices in the observed positions using channel state information.

This research is funded by Russian Science Foundation, grant 21-71-00133.

[1] *Astafiev A, Titov D, Zhiznyakov A, Demidov A* A method for mobile device positioning using a sensor network of BLE beacons, approximation of the RSSI value and artificial neural networks. // Computer Optics, Samara, 2021; 45(2) p. 277–285.

Исследование механизма синтеза эвристических решений в рамках адаптивного управления мобильным роботом

Макаров Михаил Вячеславович* nauka-murom@yandex.ru
Семенов Иван Александрович 79601712352@yandex.ru
Трантина Наталья Сергеевна trantinanatalya@yandex.ru
Муром, Муромский институт Владимирского государственного университета

Прикладное использование мобильных роботов (МР) в динамической среде значительно усложняет процесс управления, где появляется множество трудноформализуемых задач принятия решения, включающих в себя результаты прогнозирования отдельных аспектов среды существования, планирования собственных действий и оценки последствий этих действий. Предполагается, что это будет способствовать исполнению системой управления рациональных поведенческих функций МР, основанных на квази-когнитивных принципах принятия решений, возможных благодаря использованию особых методов интеллектуальной обработки информации.

В качестве объекта проведенного экспериментального исследования выступала компьютерная модель разведывательного МР в динамической среде. Осуществлялась автономная навигация, способствующая перемещению МР по оптимальной с точки зрения объема получаемой информации траектории и при учёте ограничений, накладываемых динамическими объектами в оперативном поле пространства.

В ходе исследования создавался внутренний информационный субстрат для синтеза нового типа эвристических решений с помещением действий МР в контекст поведения. Таким образом, происходит устранение абсолютной зависимости синтезируемого решения от входных сенсорных данных и учитываются условия среды существования МР.

Структурная и параметрическая вариативность обработчика информации производилась за счет принципа информационной двойственности. Данный эффект был инкорпорирован в объект исследования с помощью модификации полносвязной нейросетевой ячейки и специальной функции активации нейронов. На выходе данного слоя ожидается нечёткое высказывание, интерпретируемое любой подходящей для этого системой нечеткого вывода. Под высказыванием подразумевается выбранный вариант следующего шага системы с обоснованием в виде одной из конечного множества стратегий.

Описываемая вычислительная ячейка предполагает как минимум два вычислительных слоя. Первый слой получает используемую для принятия решения входную информацию, но не в чистом виде, а на основе производимых матрицей целеполагания преобразований. Результатом активации первого слоя является не вектор скалярных значений степени активации каждого нейрона, а множество функций. Эти функции используются в качестве активационных для нейронов последующего слоя. Аналогичная количественная форма входной инфор-

мации подается и на второй вычислительный слой, где происходит другой этап преобразования образной информации для формирования нового пространства признаков. Такая архитектура ячейки позволяет воссоздать структурную вариативность и способствует устранению строгой зависимости синтезируемого решения от входных сенсорных данных. Отличительной особенностью процесса активации является тот факт, что при использовании привычных базовых нелинейных функций предполагается, что значение выхода нейрона является количественным эквивалентом степени его участия в общем результате решения задачи. Такой подход позволяет описывать степень участия в количественном измерении, но не в качественном. Для демонстрации качественных характеристик процесса обработки информации активацией нейрона должно быть не единичное скалярное значение, а функция. Это способствует формированию внутреннего состояния системы, которое включает в себя отражение внешних условий решения задачи. Такое состояние оказывает действие на активацию, что в свою очередь и определяет контекст синтезируемых решений.

Исследование состояло из группы экспериментов, где разработанная модель компонента принятия решения в составе объекта исследования симулировала функционирование этой системы в виртуальных пространствах различных конфигураций. В рамках подобных пространств осуществлялся набор различных динамических дестабилизирующих воздействий. Каждое такое воздействие изменяло общее правило преобразования информации и принятия решения. Это связано с тем, что менялась внутренняя поведенческая мотивация системы. Критерием, интерпретирующим результаты проведенного исследования можно считать установление факта отсутствия полной зависимости от изменения выходного вектора, интерпретирующего синтезированное решение. Из поученных результатов можно увидеть, что имела место существенная доля изменений решений вызванных внесенными в процесс обработки информации изменений.

Полученные результаты позволяют сделать вывод, что существует возможность формирования квази-когнитивных механизмов обработки информации, которые могли бы использоваться в качестве субстрата для синтеза нового типа эвристических решений внутри системы управления МР в динамической среде. В рассматриваемом случае внесение структурных изменений в компонент принятия решения привело к возникновению отклика в адаптационных способностях объекта исследования при дестабилизации внешних условий. Можно сказать, что подобный информационный субстрат для нового способа описания внешней среды привел к расширению поведенческого потенциала МР.

Исследование выполнено за счет гранта Российского научного фонда No 2221-20111, https://rscf.ru/project/22-21-20111/.

[1] *Макаров М В., Семенов И. А., Трантина Н. С.* Исследование нового типа эвристических решений для адаптивного управления мобильным роботом в динамической среде // Известия Юго-Западного государственного университета, Курск: Из-во ЮЗГУ, 2022.

Investigating the mechanism of synthesis of heuristic decisions for adaptive control of a mobile robot

Makarov Mikhail* nauka-murom@yandex.ru
Semenov Ivan 79601712352@yandex.ru
Trantina Natalya trantinanatalya@yandex.ru
Murom, Murom Institute of Vladimir State University

The applied use of mobile robots (MR) in a dynamic environment significantly complicates the control process, where many difficult-to-formalize decision-making tasks appear, including the results of predicting certain aspects of the environment of existence, planning one's own actions and assessing the consequences of these actions. It is assumed that this will contribute to the performance by the control system of rational behavioral functions of the MR, based on quasi-cognitive principles of decision-making, possible through the use of special methods of intellectual data processing.

The object of the experimental study was a computer model of an intelligence MR in a dynamic environment. Autonomous navigation was carried out, facilitating the movement of the MR along the trajectory optimal in terms of the amount of information received and taking into account the restrictions imposed by dynamic objects in the operational field of space. In the course of the study, an internal information substrate was created for the synthesis of a new type of heuristic decisions with the placement of MR actions in the context of behavior. Thus, the absolute dependence of the synthesized decision on the input sensory data is eliminated and the conditions of the MR environment are taken into account.

The structural and parametric variability of the information processor was carried out due to the principle of information duality. This effect was incorporated into the object of study by modifying a fully connected neural network cell and a special function for activating neurons. At the output of this layer, a fuzzy statement is expected, interpreted by any suitable fuzzy inference system. The statement means the chosen variant of the next step of the system with justification in the form of one of the finite set of strategies.

The described computational cell assumes at least two computational layers. The first layer receives the input data used for decision-making, but not in its pure form, but on the basis of transformations produced by the goal-setting matrix. The result of activation of the first layer is not a vector of scalar values of the degree of activation of each neuron, but a set of functions. These functions are used as activation functions for the neurons of the subsequent layer. A similar quantitative form of input information is also fed to the second computational layer, where another stage of transformation of figurative information takes place to form a new feature space. Such a cell architecture allows to recreate the structural variability and helps to eliminate the strict dependence of the synthesized solution on the input sensory data.

A distinctive feature of the activation process is the fact that when using the usual basic nonlinear functions, it is assumed that the value of the output of a neuron is the quantitative equivalent of the degree of its participation in the overall result of solving the problem. This approach makes it possible to describe the degree of participation in a quantitative measurement, but not in a qualitative one. To demonstrate the qualitative characteristics of the information processing process, the activation of a neuron should not be a single scalar value, but a function. This contributes to the formation of the internal state of the system, which includes a reflection of the external conditions for solving the problem. This state has an effect (a way of describing the process) on activation, which in turn determines the context of the synthesized decisions.

The study consisted of a group of experiments where the developed model of the decision-making component as part of the research object simulated the functioning of this system in virtual spaces of various configurations. Within the framework of such spaces, a set of various dynamic destabilizing influences was carried out. Each such impact changed the general rule of information transformation and decision-making. This is due to the fact that the internal behavioral motivation of the system was changing. The criterion interpreting the results of the study can be considered the establishment of the fact that there is no complete dependence on the change in the output vector interpreting the synthesized decision. From the studied results, it can be seen that there was a significant proportion of changes in decisions caused by changes made to the information processing process.

The results obtained allow us to conclude that there is a possibility of forming quasi-cognitive information processing mechanisms that could be used as a substrate for the synthesis of a new type of heuristic decisions inside the MR control system in a dynamic environment. In the case under consideration, the introduction of structural changes in the decision-making component led to the emergence of a response in the adaptive abilities of the object of study during the destabilization of external conditions. It can be said that such an information substrate for a new way of describing the external environment has led to an expansion of the behavioral potential of the MR.

This work has been supported by the grants the Russian Science Foundation, RSF 22-21-20111, https://rscf.ru/en/project/22-21-20111/.

[1] *Makarov M V., Semenov I. A., Trantina N. S.* Research of a new type of heuristic solutions for adaptive control of a mobile robot in a dynamic environment // Proceedings of the Southwestern State University, Kursk: Southwestern State University, 2022.

Исследование интеллектуальных элементов управления мобильным роботом и обеспечение информационной безопасности процесса его функционирования в динамической среде

Макаров Михаил Вячеславович[1]⋆ nauka-murom@yandex.ru
Семенов Иван Александрович[1] 79601712352@yandex.ru
[1]Муром, Муромский институт Владимирского государственного университета

Представленное исследование нацелено на установление и обоснование принципов эффективной инкорпорации интеллектуальных элементов системы управления мобильным роботом (МР), функционирующим в условиях динамической среды существования. В качестве предмета исследования использовалась входящая в состав управления процедура одновременной локализации и картографирования (SLAM). Критерием эффективности выступали показатели, связанные с обеспечением информационной безопасности процесса функционирования робота в реальных условиях эксплуатации.

Разработана и реализована методология экспериментального исследования программного исполнения процедуры SLAM в рамках задачи управления МР. Объектом исследования выступала компьютерная модель абстрактного МР, выполняющего разведывательные функции в виртуальной динамической среде существования. Инкорпорируемыми элементами интеллектуальной обработки информации в процедуру SLAM были сверточные и полносвязные нейросетевые слои, обеспечивающие фильтрацию динамических объектов.

При проведении экспериментального исследования выполнена симуляция процесса функционирования компьютерной модели разведывательного МР в виртуальной среде существования. Аналогичные эксперименты воспроизведены при различных структурно-функциональных конфигурациях процедуры SLAM. Получены количественные результаты, демонстрирующие точность позиционирования объекта исследования для каждого из способов организации данной процедуры. Проведен сравнительный анализ вариантов использования элементов интеллектуальной обработки информации внутри данной процедуры.

Установлено, что инкорпорация элементов интеллектуальной обработки информации в процедуру SLAM имеет влияние на точность позиционирования МР и надежность его функционирования в динамической среде существования. Это вносит вклад в соблюдение норм информационной безопасности при использовании МР в реальных условиях эксплуатации. Также определено, что существует и их избыточное употребление, которое приводит к отклонению от оптимальных качеств, необходимых для эффективного автономного управления МР и обеспечения информационной безопасности.

Исследование выполнено за счет гранта Российского научного фонда No.22-21-20111, https://rscf.ru/project/22-21-20111/.

[1] *Макаров М.В., Семенов И.А.* Исследование интеллектуальных элементов управления мобильным роботом и обеспечение информационной безопасности процесса его функционирования в динамической среде // Известия Юго-Западного государственного университета, Курск: Из-во ЮЗГУ, No.2, 2022.

The research of intelligent controls of a mobile robot and ensuring information security of its functioning in a dynamic environment

Makarov Mikhail[1]⋆ nauka-murom@yandex.ru
Semenov Ivan[1] 79601712352@yandex.ru
[1]Murom, Murom Institute of Vladimir State University

The presented research is aimed at establishing and substantiating the principles of effective incorporation of intelligent elements of the control system of a mobile robot (MR) operating in a dynamic environment of existence. The procedure of simultaneous localization and mapping (SLAM) included in the control was used as the subject of the study. The efficiency criterion was the indicators related to ensuring the information security of the MR functioning process in real operating conditions.

The methodology of experimental research of the program execution of the SLAM procedure within the framework of the task of controlling a MR has been developed and implemented. The object of the study was a computer model of an abstract MR performing intelligence functions in a virtual dynamic environment of existence. The incorporated elements of intelligent information processing in the SLAM procedure were convolutional and fully connected neural network layers that provide filtering of dynamic objects.

During the experimental study, a simulation of the process of functioning of a computer model of an intelligence MR in a virtual environment of existence was performed. Similar experiments were reproduced with different structural and functional configurations of the SLAM procedure. Quantitative results were obtained demonstrating the accuracy of the positioning of the object of study for each of the methods of organizing this procedure. A comparative analysis of the options for using elements of intelligent information processing within this procedure is carried out.

It is established that the incorporation of elements of intelligent information processing into the SLAM procedure has an impact on the accuracy of the positioning of the MR and the reliability of its functioning in a dynamic environment of existence. This contributes to compliance with information security standards when using MP in real operating conditions. It is also determined that there is also their excessive use, which leads to a deviation from the optimal qualities necessary for effective autonomous control of the MR and ensuring information security.

This work has been supported by the grants the Russian Science Foundation, RSF 22-21-20111, https://rscf.ru/en/project/22-21-20111/.

[1] *Makarov M.V., Semenov I.A.* The Research of Intelligent Controls of a Mobile Robot and Ensuring Information Security of Its Functioning in a Dynamic Environment // Proceedings of the Southwestern State University, Kursk: Southwestern State University, No.2, 2022.

Комплексирование данных нескольких физических методов при решении обратных задач спектроскопии растворов методами машинного обучения

Гуськов Артём Алексеевич[1]⋆ artemguskov99@mail.ru
Исаев Игорь Викторович[2,3] isaev_igor@mail.ru
Лаптинский Кирилл Андреевич[1,2] laptinskiy@physics.msu.ru
Бурикov Сергей Алексеевич[1,2] sergey.burikov@gmail.com
Сарманова Ольга Эдуардовна[1] oe.sarmanova@physics.msu.ru
Доленко Татьяна Альдефонсовна[1,2] tdolenko@lid.phys.msu.ru
Доленко Сергей Анатольевич[2] dolenko@srd.sinp.msu.ru

[1]Москва, МГУ имени М.В.Ломоносова, Физический факультет
[2]Москва, МГУ имени М.В.Ломоносова, Научно-исследовательский институт ядерной физики имени Д.В. Скобельцына
[3]Москва, Российская академия наук, Институт радиотехники и электроники имени В. А. Котельникова

Тяжелые металлы относятся к числу опасных загрязнителей воды, поэтому во многих областях промышленности и экологии существует потребность в определении концентрации растворенных в воде ионов тяжелых металлов. Их содержание в природе увеличивается как в результате природных процессов (выветривание горных пород, вулканическая активность), так и в ходе человеческой деятельности (выбросы промышленных предприятий, переработка нефти, добыча полезных ископаемых и т.д.). В водной среде тяжелые металлы обычно находятся в форме ионов. Поскольку они оказывают сильное негативное влияние на организм человека, то задача определения типа и концентрации ионов тяжелых металлов в воде является важной для экологического мониторинга.

Наиболее точными методами определения химического состава растворов являются традиционные химико-аналитические методы. Однако, такой подход требует длительного времени, хорошей подготовки образцов и расхода дорогостоящих реагентов. В тоже время, для решения большинства практических задач больший интерес представляют простые в применении, быстрые и бесконтактные методы. Поэтому в качестве альтернативы рассматриваются методы оптической спектроскопии, обладающие перечисленными преимуществами. В частности, в настоящей работе дистанционное экспресс-определение концентраций ионов тяжелых металлов в воде осуществляется с помощью лазерной спектроскопии комбинационного рассеяния света, спектроскопии поглощения и инфракрасной спектроскопии. К сожалению, на текущий момент для данных методов не существует аналитического решения задачи определения концентраций каждого компонента в многокомпонентных растворах по их спектрам, поэтому одним из немногих способов решения таких задач является применение методов машинного обучения, основанных на использовании экспериментальных данных.

Для повышения качества решения исследуемой обратной задачи было предложено совместное применение (комплексирование) физических методов. Основная идея такого подхода заключается в решении рассматриваемой задачи с одновременным использованием данных нескольких спектроскопических методов.

В настоящей работе сравнивались результаты определения концентраций ионов тяжёлых металлов в растворах, полученные путем комплексирования данных спектроскопии комбинационного рассеяния, спектроскопии поглощения и инфракрасной спектроскопии, с результатами использования каждого метода по отдельности.

Исследование выполнено за счёт гранта Российского Научного фонда, проект № 19-11-00333, https://rscf.ru/project/19-11-00333/.

Integration of data from various physical methods in solving inverse problems of spectroscopy of solutions by machine learning methods

Guskov Artem[1]⋆ artemguskov99@mail.ru
Isaev Igor[2,3] isaev_igor@mail.ru
Laptinskiy Kirill[1,2] laptinskiy@physics.msu.ru
Burikov Sergey[1,2] sergey.burikov@gmail.com
Sarmanova Olga[1] oe.sarmanova@physics.msu.ru
Dolenko Tatyana[1,2] tdolenko@lid.phys.msu.ru
Dolenko Sergey[2] dolenko@srd.sinp.msu.ru

[1]Moscow, M.V. Lomonosov Moscow State University, Faculty of Physics
[2]Moscow, M.V. Lomonosov Moscow State University, D.V. Skobeltsyn Institute of Nuclear Physics
[3]Moscow, Russian Academy of Sciences, Kotelnikov Institute of Radio Engineering and Electronics

Heavy metals are among the dangerous pollutants of water, therefore, in many areas of industry and ecology there is a need to determine the concentration of heavy metal ions dissolved in water. Their amount in nature increases both as a result of natural processes (weathering of rocks, volcanic activity) and as a result of human activity (emissions from industrial enterprises, oil refining, mining, etc.). In the aquatic environment, heavy metals are usually in the form of ions. Since they have a strong negative impact on the human body, the task of determining the type and concentration of heavy metal ions in water is important for environmental monitoring.

The most accurate methods for determining the chemical composition of solutions are traditional chemical-analytical methods. However, this approach requires a long time, good sample preparation and the consumption of expensive reagents. At the same time, easy-to-use, fast and non-contact methods are of greater interest for solving most practical problems. Therefore, optical spectroscopy methods with the above advantages are considered as an alternative. In particular, in this study, remote express determination of concentrations of heavy metal ions in water is carried out using Raman laser spectroscopy, absorption spectroscopy and infrared spectroscopy. Unfortunately, at present there is no analytical solution to the problem of determining the concentrations of each component in multicomponent solutions by their spectra, so one of the few ways to solve such problems is the application of machine learning methods based on the use of experimental data.

In order to improve the quality of the solution of the investigated inverse problem, the joint application of physical methods has been proposed. The main idea of this approach is to solve the problem under consideration with the simultaneous use of data from several spectroscopic methods.

In the present work, we compared the results of determining the concentrations of heavy metal ions in solutions obtained by combining data from Raman spectroscopy, absorption spectroscopy, and infrared spectroscopy with the results of using each method separately.

This study has been performed at the expense of the grant of the Russian Science Foundation (project no. 19-11-00333), https://rscf.ru/en/project/19-11-00333/.

Методика построения диагностических оценочных шкал и моделей

Гончарова Анастасия Борисовна[1]⋆ a.goncharova@spbu.ru
Аржаник Александра Алексеевна[1] arzh_sasha@mail.ru
Виль Мария Юрьевна[1] st054723@student.spbu.ru
[1]Санкт-Петербург, Санкт-Петербургский государственный университет

Специфическим инструментом, применяемым в медицинской практике, являются оценочные (прогностические) шкалы. Прогностические модели, построенные на статистическом анализе популяционных данных, наиболее ценны с точки зрения применения на общую популяцию. В 1952 г. на 27-м ежегодном конгрессе анестезиологов В. Апгар представила первую в мире и актуальную по сей день балловую шкалу диагностики асфиксии и оценки степени ее тяжести для новорожденного [1]. Для общей оценки тяжести состояния у взрослых применяются шкалы и их модификации SAPS, APACHE, SOFA, MODS и другие, у детей - PRISM, DORA, PIM и другие, для оценки хирургических больных в отделениях реанимации и интенсивной терапии EUROSCORE, MPM for cancer patients, Glasgow Coma Score, разработаны шкалы оценки тяжести травмы, оценки эффективности лечения и ухода, нозоспецифические шкалы [1]. Стандартной методики построения шкал не существует, так как параметры, входящие в шкалы разнообразны. Важно отметить зависимость шкал от популяций, на которых они построены, и от объема выборки, используемой для построения шкалы. Единственным важным требованием к шкалам является высокая чувствительность и универсальность применения.

Оценку прогноза на основании клинических рекомендаций Минздрава Российской Федерации по внебольничной пневмонии по всем пациентам рекомендуется делать по шкале CURB-65 [2], которая включает анализ 5 признаков, значения которых представленны в таблице 3 во втором столбце.

Наличие каждого признака оценивается в 1 балл, общая сумма может варьироваться от 0 до 5 баллов, риск летального исхода возрастает по мере увеличения суммы баллов. Объем выбороки исследования в случае шкалы CURB-65, разработанной в 2001 году по данным из Великобритании, Новой Зеландии и Нидерландов, составил 1068 пациентов [3].

В ходе исследования, проводимого ФГБОУ ВО СибГМУ Минздрава России составлена база по всем пациентам с внебольничной пневмонией, обратившимся в больницы г. Томска за 2017 год, она содержит данные о двухстах параметрах, измеренных у 1412 пациентов, а так же исход заболевания. Требовалось проанализировать параметры, установить какие влияют на исход заболевания и разработать систему прогнозирования исхода. Необходимо снизить размерность данных, то есть выделить те данные, которые влияют на летальность. Для построения шкалы прогнозирования исходная база разбивалась в отношении 2:1, верификация проводилась на 411 пациентах тестовой выборки.

Для определения различий между группами выживших и умерших пациентов использовались:

1. критерий Манна-Уитни для количественных переменных [4];
2. критерий Пирсона для номинальных переменных, при небольших выборках для переменных, принимающих только два значения при двух видах исходов применялся точный критерий Фишера [4].

Для уровня значимости p<0,05 из обучающей выборки исходной базы выделены 23 признака с количественными значениями и 16 признаков с номинальными значениями. Среди значимых переменных выделялись все с наибольшим уровнем значимости (p<0,0001), потом они ранжировались по величине значения критерия. Результаты представлены в таблицах 1 и 2.

Таблица 1. Результаты расчета критерия Манна-Уитни.

Количественный признак	**Значение критерия** U
Частота сердечных сокращений	25305,5
Возраст	23786
АЛТ	22312,5
Систолическое давление	18605,5
Диастолическое давление	18000
Креатинин	17289,5
Частота дыхания	15115
Температура	14319,5
Сегментоядерные	10976
Сатурация	9144,5
Азот мочевины	9096
Общий белок	8371,5

Таблица 2. Результаты расчета критерия χ^2.

Номинальный признак	**Значение критерия** χ^2
Частота сердечных сокращений	340,32
Неврологические нарушения	86,64

С использованием точек разделения - индекса Йодена [5], впервые модифицирована шкала CURB-65 для нашей популяции (третий столбец Таблицы 3).

Следует отметить, что параметры, входящие в шкалу CURB-65 легко измеряемы в учреждениях первичной помощи. В таблице 4 представлены характеристики моделей, построенных на параметрах шкалы CURB-65, на тестовой выборке. Модели Баейесовского классификатора и логистической регрессии строились и на всех значимых параметрах, однако, лучший результат показывают модели, построенные на основе параметров, входящих в шкалу CURB-65. Такой результат объясняется тем, что отбор параметров модели обуславливается недопустимостью ошибки первого рода.

Таблица 3. Сравнение признаков классической и модифицированной шкалы CURB-65.

Признак	**CURB-65**	**Модифицированная шкала CURB-65**
нарушение сознания, обусловленное пневмонией	да	да
повышение уровня азота мочевины	> 7 ммоль/л	$> 9{,}5$ ммоль/л
частота дыхания	$\geqslant 30$ /мин	> 21/мин
давление	снижение систолического артериального давления < 90 мм рт.ст. или диастолического $\leqslant 60$ мм рт.ст.	снижение систолического артериального давления $\leqslant 105$ мм рт.ст. или диастолического $\leqslant 65$ мм рт.ст.
возраст больного	$\geqslant 65$ лет	> 72 лет

Таблица 4. Показатели эффективности построенных моделей.

Модель	**Чувствительность**	**Специфичность**	**AUC**
CURB-65	0,875	0,763	0,868
Модификация CURB-65	0,875	0,863	0,906
Байесовский классификатор	0,875	0,891	0,961
Логистическая регрессия	1,000	0,741	0,945

[1] *Александрович Ю. С., Гордееев В. И.* Оценочные и прогностические шкалы в медицине критических состояний. Издательство «Сотис», 2007. 140 С.

[2] *Министрерство здравоохранения РФ.* Клинические рекомендации. Внебольничная пневмония у взрослых. Издательство Минздрава РФ, 2021. 117 С.

[3] *Lim W., van der Eerden M., Laing R. et al.* Defining community acquired pneumonia severity on presentation to hospital: an international derivation and validation study. // Thorax. 2003. 58 (5). p. 377–382.

[4] *Arzhanik A. A., Goncharova A. B. et al.* Detection of the Community-Acquired Pneumonia Factors Leading to Death. // SCP 2020. Lecture Notes in Control and Information Sciences. Springer Nature, Stability and Control Processes. 2022. p. 545–550.

[5] *Файнзильберг Л. С., Жук Т. Н.* Гарантированная оценка эффективности диагностических тестов на основе усиленного ROC-анализа. // Управляющие системы и машины. 2009. № 5. С. 3–13.

Methodology for constructing diagnostic evaluation prognostic scales and models

Goncharova Anastasia[1]⋆ a.goncharova@spbu.ru
Arzhanik Alexandra[1] arzh_sasha@mail.ru
Vil' Maria[1] st054723@student.spbu.ru
[1]Saint Petersburg, Saint Petersburg State University

The most valuable tool used in medical practice is the prognostic scales. From the perspective of wide population applicability, prognostic models based on statistical analysis of population data are most valuable. In 1952, at the 27th Annual Congress of Anesthesiologists, W. Apgar presented the world's first and still relevant score scale for diagnosing asphyxia and assessing its severity for the newborn [**?**]. For a broad assessment of the severity of a condition in adults, the SAPS, APACHE, SOFA, MODS, and their modifications, as well as other scales, are used. For children, the PRISM, DORA, PIM, and other scales are used. The EUROSCORE, MPM for cancer patients, and Glasgow Coma Score was developed; scales of trauma severity, assessment of treatment and care effectiveness, specific nosology scales was developed. There is no standard methodology for constructing the scales, as the parameters taken into consideration in building the scales are diverse. It is important to note the dependence of scales on the populations for which they are developed and on the sample size used to construct the scale. The only important requirement for scales is their high sensitivity and universal applicability. It is recommended to evaluate the prognosis based on the clinical recommendations of the Ministry of Health of the Russian Federation for community-acquired pneumonia for all patients on the CURB-65 scale [2], which includes an analysis of 5 signs, the values of which are presented in Table 3 in the second column.

The presence of each feature is estimated at 1 point, the total amount can vary from 0 to 5 points, the risk of death increases as the amount of points increases. The sample size of the study in the case of the CURB-65 scale, developed in 2001 according to data from the UK, New Zealand and Netherlands, made up 1068 patients [3].

In the course of a study conducted by the Federal State Budgetary Educational Institution of the Russian Ministry of Health, a database was compiled for all patients with community-acquired pneumonia cared for at hospitals in Tomsk in 2017, it contains data on two hundred parameters measured on 1412 patients, as well as the outcome of the disease. It was necessary to analyze the parameters, determine which affect the outcome of the disease, and develop a system for predicting the outcome. It was necessary to reduce the dimensionality of the data, that is, to identify those data that affect mortality. To build a prediction scale, the initial base was divided in the ratio 2:1, verification was carried out on 411 patients of the test sample.

To determine the differences between the groups of surviving and deceased patients , the following methods were used:

1. Mann-Whitney criterion for quantitative variables [4];

2. Pearson's criterion was applied for nominal variables, small samples for variables taking only two values for two types of outcomes the exact Fisher criterion [4] was applied.

For the significance level $p < 0.05$, 23 signs with quantitative values and 16 signs with nominal values were selected from the training sample of the initial base. Among the significant variables, all with the highest level of significance ($p<0.0001$) were distinguished, then they were ranked by the value of the criterion. The results are presented in tables 1 and 2.

Table 1. Calculation results of the Mann-Whitney criterion.

Quantitative attribute	**Criterion value** U
Pulse rate	25305,5
Age	23786
SGPT	22312,5
Systolic Pressure	18605,5
Diastolic Pressure	18000
Creatinine	17289,5
Respiratory Rate	15115
Temperature	14319,5
Segmentonuclear	10976
Saturation	9144,5
Urea	9096
Total Protein	8371,5

Table 2. Results of the calculation of the criterion χ^2.

Nominal attribute	**Criterion value** χ^2
Heart Rate	340,32
Neurological Disorders	86,64

Using the separation points - the Yoden index [5], the CURB-65 scale for our population was modified for the first time (the third column of Table 3).

It should be noted that the parameters included in the CURB-65 scale are easily measured in primary care institutions. Table 4 shows the characteristics of models based on the parameters of the CURB-65 scale, on a test sample. Models of the Bayesian classifier and logistic regression were built on all significant parameters, however, the best result is shown by models built on the basis of parameters included in the CURB-65 scale. This result is explained by the fact that the selection of model parameters is conditioned by the inadmissibility of the error of the first kind.

Table 3. Comparison of features of the classical and modified CURB-65 scale.

Attribute	**CURB-65**	**Modified CURB-65 scale**
Confusion	yes	yes
Urea	>7 mmol/l	> 9.5 mmol/l
Respiratory rate	⩾ 30/min	> 21/min
Blood pressure	SBP <90 mm Hg or DBP ⩽60 mm Hg	SBP ⩽105 mm Hg or DBP ⩽ 65 mm Hg
Age	⩾65 years	> 72 years

Table 4. Performance indicators of the constructed models.

Model	**Sensitivity**	**Specificity**	**AUC**
CURB-65	0.875	0.763	0.868
Modification CURB-65	0.875	0.863	0.906
Bayesian classifier	0.875	0.891	0.961
Logistic regression	1.000	0.741	0.945

[1] *Aleksandrovich Yu. S., Gordeev V. I.* Evaluation and prognostic scales in critical condition medicine. Sotis Publishing House, 2007. 140 P.

[2] *Ministry of Health of the Russian Federation.* Clinical recommendations. Community-acquired pneumonia in adults. Publishing House of the Ministry of Health of the Russian Federation, 2021. 117 P.

[3] *Lim W., van der Eerden M., Laing R. et al.* Defining community acquired pneumonia severity on presentation to hospital: an international derivation and validation study. // Thorax. 2003. 58 (5). p. 377–382.

[4] *Arzhanik A. A., Goncharova A. B. et al.* Detection of the Community-Acquired Pneumonia Factors Leading to Death. // SCP 2020. Lecture Notes in Control and Information Sciences. Springer Nature, Stability and Control Processes. 2022. p. 545–550.

[5] *Feinsilberg L. S., Zhuk T. N.* Guaranteed evaluation of the effectiveness of diagnostic tests based on enhanced ROC analysis. // Control systems and machines. 2009. No. 5. p. 3–13.

Обнаружение опасных аритмий при помощи нейронной сети долгой краткосрочной памяти по описанию коротких сигналов ЭКГ в частотной области

Немирко Анатолий Павлович apn-bs@yandex.ru
Попадьина Алина Олеговна* alinaP-1998@mail.ru
Манило Людмила Алексеевна lmanilo@yandex.ru
Санкт-Петербург, Санкт-Петербургский государственный электротехнический университетет "ЛЭТИ"

Наиболее важной функцией систем диагностики и наблюдения за состоянием кардиологических больных, является оперативное распознавание аритмий, непосредственно угрожающих жизни пациента. К таким катастрофическим аритмиям относятся: асистолия (остановка сердца), выраженная брадикардия, фибрилляция желудочков сердца (ФЖ), желудочковая тахикардия (ЖТ) и трепетание желудочков (ТЖ). Существуют также ограничения на время анализа ЭКГ, которое в разных исследованиях варьируется от 2 до 8 с. Фактор сокращения времени анализа представляется чрезвычайно важным, поскольку мгновенная индикация опасного нарушения, особенно в имплантированных кардиостимуляторах, помогает пациенту сохранить жизнь.

Учитывая отмеченные выше моменты, в данном исследовании была поставлена следующие задача построить алгоритмы классификации на основе анализа спектрального описания 2 с фрагментов ЭКГ.

При решении схожих задач авторы прибегают к использованию машинного обучения или глубокого обучения. В данной работе рассмотрели наиболее популярные методы решения – метод ближайшего соседа (k-nearest neighbors; kNN), машину опорных векторов (support vector machine; SVM) и рекуррентную нейронную сеть долгой краткосрочной памяти (Long short-term memory; LSTM).

В данной работе использованы фрагменты ЭКГ из базы ЭКГ-данных MIT-BIH Malignant Ventricular Ectopy Database v1.0.0 PhysioNet [1]. Данные представляют собой одномерное спектральное представление сигнала электрокардиограммы длительностью 2 секунды. Для анализа использовали сглаженные спектры 0–15 Гц с шагом 1 и 1,5 Гц. Полное описание процесса обработки данных при подготовке к обучению описано в [1].

Разметка данных производилась специалистами ФГБУ «НМИЦ им. В. А. Алмазова». В работе рассмотрено два случая разбиения данных. В первом случае данные были разделены на классы: «опасные» для жизни аритмии и «неопасные». К «опасным» аритмиям были отнесены следующие патологии: трепетание желудочков (VFL); желудочковая фибрилляция (VF); пируэтная тахикардия (VTTdP); желудочковая тахикардия с высокой частотой сердечных сокращений (VTHR) – всего 578 записей.

Как «неопасные» рассматривали следующие категории: желудочковая тахикардия с низкой частотой сердечных сокращений (VTLR); бигеминия (B); высокая степень вентрикулярной эктопической активности (HGEA); желудочковый ритм (VER); фибрилляция предсердий (AFIB); наджелудочковая тахикардия (SVTA); синусовая брадикардия (SBR); блокада сердца первой степени (BI); узловой (a-v) ритм (NOD); синусовый ритм с блокадой ножек пучка Гиса (BBB); нормальный ритм (N); норма с экстрасистолией (Ne) – 438 записей. Итого объем данных составил 1016 записей. Во втором случае набор данных содержал только опасные для жизни аритмии с метками VFL и VF (опасные аритмии), и объекты BBB, N, Ne (неопасные аритмии).

Для первичной оценки классификации использовали приложение Classification Learner в программной среде MATLAB. Classification Learner позволяет с высокой скоростью оценить точность классификации классическими методами машинного обучения. Установили кросс-валидацию k-fold, где k = 5. Обучение провели для данных сглаженного спектра с шагом 1 Гц и с шагом 1,5 Гц. Также оценили эффективность использования полного спектра. Полученный результаты приведены в таблице 1.

Таблица 1. Показатели точности, чувствительности и специфичности на тестовой выборке

Метод	Все классы			Синусовый ритм и опасные желудочковые аритмии		
	OA,%	Se,%	Sp,%	OA,%	Se,%	Sp,%
	Сглаженный спектр (0-15 Гц) с шагом 1,5 Гц					
SVN	94,09	95,33	92,47	98,88	98,22	100
kNN	92,91	92,21	93,84	98,14	97,03	100
LSTM	92,76	92,80	92,68	97,97	99,09	94,60
	Сглаженный спектр (0-15 Гц) с шагом 1 Гц					
SVM	94,69	94,46	94,98	98,51	98,21	99,00
kNN	93,50	97,70	95,89	97,77	97,03	99,00
LSTM	94,08	96,36	95,50	98,65	99,09	97,30

Для оценки эффективности алгоритма использовали следующие показатели: общую точность (Overall Accuracy), чувствительность (Sensitivity) и специфичность (Specificity). Данные были разбиты на тестовую и обучающую. Объем обучающей выборки составил 70%, объем тестовой – 30%. Подробные результаты исследования приведены в [2]. В соответствии с таблицей 1, точность моделей практически не отличается для сглаженного спектра с шагом 1 Гц и 1,5 Гц.

Классические методы SVM и k-NN при работе с объектами малого размера способны обеспечить точность классификации, сравнимую с показателями для

сети LSTM. Классификаторы обеспечили точность 98% для задачи разделения опасных для жизни аритмий и нормы, и точность 94% для разделения между «опасными» и «неопасными» аритмиями.

Работа поддержана грантом РФФИ No. 19-29-01009.

[1] *Nemirko A., Manilo L., Tatarinova A., Alekseev B., Evdakova E.* Fragment Database for the Exploration of Dangerous Arrhythmia (version 1.0.0) // PhysioNet. 2022.

[2] *Немирко А. П., Попадьина А. О.* Распознавание опасных для жизни аритмий по спектральным характеристикам короткой записи электрокардиограммы // Биотехносфера. 2022, №1, — С. 3–8.

Detection of dangerous arrhythmias using a long-short-term memory neural network with a description of short ECG signals in the frequency domain

Nemirko Anatoly apn-bs@yandex.ru
***Popadina Alina**⋆* alinap-1998@mail.ru
Manilo Liudmila lmanilo@yandex.ru
Saint Petersburg, Saint Petersburg Electrotechnical Universuty "LETI"

The most important function of systems for diagnosing and monitoring the condition of cardiac patients is the high-speed recognition of arrhythmias that directly threaten the patient's life. Dangerous arrhythmias mentioned above include asystole (cardiac arrest), severe bradycardia, ventricular fibrillation (VF), ventricular tachycardia (VT), and ventricular flutter (VFL). There are also limitations on the ECG analysis time, which varies from 2 to 8 seconds in different studies. The factor of reducing the analysis time is extremely important, since the instant indication of a dangerous violation, especially in implanted pacemakers, helps the patient to save life.

Considering the points noted above, in this study the following task was set to build classification algorithms based on the analysis of the spectral description of 2 seconds ECG fragments.

When solving similar problems, the authors resort to using machine learning or deep learning. In this paper, we considered the most popular solution methods - the k-nearest neighbors' method (kNN), the support vector machine (SVM) and the recurrent neural network of long short-term memory (LSTM).

In this work, we used ECG fragments from the MIT-BIH Malignant Ventricular Ectopy Database v1.0.0 PhysioNet ECG database [1]. The data is a one-dimensional spectral representation of the 2 seconds duration electrocardiogram signal. For the analysis, smoothed spectra of 0–15 Hz with steps of 1 and 1.5 Hz were used. A complete description of the data processing in preparation for the training is described in [2].

Data labeling was carried out by specialists of the Almazov National Medical Research Centre. In the work, we considered two cases of data partitioning. In the first case, the data were divided into classes: "life-threatening" arrhythmias and "non-dangerous". The following pathologies were attributed to "life-threatening" arrhythmias: ventricular flutter (VFL); ventricular fibrillation (VF); ventricular tachycardia torsade de pointes (VTTdP); high-rate ventricular tachycardia (VTHR) - 578 records in total.

The following categories were considered "non-dangerous": low-rate ventricular tachycardia (VTLR); ventricular bigeminy (B); high degree of ventricular ectopic activity (HGEA); ventricular escape rhythm (VER); atrial fibrillation (AFIB); supraventricular tachycardia (SVTA); sinus bradycardia (SBR); first degree heart block (BI); nodal (av) rhythm (NOD); sinus rhythm with bundle branch block

(BBB); normal sinus rhythm (N); normal rhythm with single extrasitols (Ne) - 438 records. The total amount of data was 1016 records.

In the second case, the dataset contained only life-threatening arrhythmias labeled VFL and VF (dangerous arrhythmias), and objects BBB, N, Ne (non-dangerous arrhythmias).

For the primary assessment of the classification, the Classification Learner application in the MATLAB software environment was used. Classification Learner allows you to quickly evaluate the accuracy of classification using classical machine learning methods. We set up k-fold cross-validation, where k = 5. Training was performed for the data of the smoothed spectrum with a step of 1 Hz and with a step of 1.5 Hz. The results obtained are shown in Table 1.

Table 1. Indicators of accuracy, sensitivity and specificity on the test set

Method	All classes			Sinus Rhythm and Dangerous Ventricular Arrhythmias		
	OA,%	Se,%	Sp,%	OA,%	Se,%	Sp,%
	Smoothed spectrum (0-15 Hz) in 1.5 Hz steps					
SVN	94,09	95,33	92,47	98,88	98,22	100
kNN	92,91	92,21	93,84	98,14	97,03	100
LSTM	92,76	92,80	92,68	97,97	99,09	94,60
	Smoothed spectrum (0–15 Hz) in 1 Hz steps					
SVM	94,69	94,46	94,98	98,51	98,21	99,00
kNN	93,50	97,70	95,89	97,77	97,03	99,00
LSTM	94,08	96,36	95,50	98,65	99,09	97,30

To evaluate the effectiveness of the algorithm, the following indicators were used: Overall Accuracy (OA), Sensitivity (Se) and Specificity (Sp). The data was divided into test and training sets. The volume of the training sample was 70%, the volume of the test sample was 30%. Detailed results of the study are given in [2].

According to Table 1, the accuracy of the models is practically the same for the smoothed spectrum with a step of 1 Hz and 1.5 Hz.

The classical SVM and k-NN methods, when working with small objects, are able to provide classification accuracy comparable to that for the LSTM network. The classifiers provided 98% accuracy for the task of separating life-threatening arrhythmias from normal, and 94% accuracy for separating between "life-threatening" and "non-dangerous" arrhythmias.

This research is funded by RFBR, grant 19-29-01009.

[1] *Nemirko A., Manilo L., Tatarinova A., Alekseev B., Evdakova E.* Fragment Database for the Exploration of Dangerous Arrhythmia (version 1.0.0) // PhysioNet. 2022.

[2] *Nemirko A. P., Popadina A. O.* Recognition of life-threatening arrhythmias by the spectral characteristics of a short electrocardiogram recording // Biotechnosfera. 2022, no 1, — pp. 3–8.

Распознавание застойной сердечной недостаточности по критерию хаотичности ритмограммы

Манило Людмила Алексеевна[1] lmanilo@yandex.ru
Холматов Достон Умиджонович[1]⋆ xolmatov.2000@mail.ru
Немирко Анатолий Павлович[1] apn-bs@yandex.ru
[1]Санкт-Петербург, Санкт-Петербургский государственный электротехнический университет "ЛЭТИ" им. В.И. Ульянова (Ленина)

В работе исследуется возможность распознавания застойной сердечной недостаточности (ЗСН) на фоне нормального синусового ритма (НСР) с применением графиков Пуанкаре, представляющих собой отображение сердечного ритма в пространстве фазовых координат. В качестве основных параметров, характеризующих особенности графических данных, используются специальная функция комплексной корреляционной меры, а также статистические показатели. Выбор такого описания связан с тем, что при ЗСН благодаря активации компенсаторных механизмов системы регуляции сердечного ритма наблюдаются сложные хаотические изменения в структуре ритма.

График Пуанкаре был представлен в виде двумерной диаграммы рассеяния последовательных точек, координатами которых являются величины смежных RR-интервалов сердечного ритма. По геометрическим формам рассеяния точек на графике Пуанкаре можно судить о наличии сердечных патологий. Для этого обычно применяют стандартные статистические показатели SD1 и SD2. Они характеризуют дисперсию облака рассеяния вдоль и перпендикулярно линии тождества, что характерно для некоторых нарушений ритма. Однако применимы они в основном к эллиптическим формам рассеяния на графике Пуанкаре, поскольку имеют ряд ограничений. Такой подход не распознаёт временные изменения сердечного ритма, появление нелинейных компонент и кластерных образований, например, в виде лепестков, что характерно для некоторых видов аритмий. Поэтому в данной работе исследовался новый показатель комплексной корреляционной меры (ССМ), отражающий множественную корреляцию запаздывания, что важно для анализа структуры временного ряда при ЗСН [1]. Кроме этого оценивались значения часто используемых стандартных статистических показателей.

Для классического отображения Пуанкаре с задержкой $m = 1$ показатель ССМ вычисляется оконным образом с помощью движущегося треугольного окна, состоящего из последовательных точек $\mathrm{P}_i(\mathrm{RR}_i; \mathrm{RR}_{i+1})$, $\mathrm{P}_{i+1}(\mathrm{RR}_{i+1}; \mathrm{RR}_{i+2})$, $\mathrm{P}_{i+2}(\mathrm{RR}_{i+2}; \mathrm{RR}_{i+3})$. На каждом i-ом шаге находится оконная площадь $\mathrm{S}(i)$ через определитель матрицы:

$$\mathrm{S}(i) = \frac{1}{2} \begin{vmatrix} \mathrm{RR}_i & \mathrm{RR}_{i+1} & 1 \\ \mathrm{RR}_{i+1} & \mathrm{RR}_{i+2} & 1 \\ \mathrm{RR}_{i+2} & \mathrm{RR}_{i+3} & 1 \end{vmatrix}.$$

Тогда показатель CCM вычисляется как:

$$\mathrm{CCM}(m) = \frac{1}{\mathrm{C}_n(N-2)} \sum_{i=1}^{N-2} |\mathrm{S}(i)|.$$

где N – количество точек на графике Пуанкаре, m – величина задержки (запаздывания), C_n – площадь эллипса, в который вписывается облако рассеивания, вычисляется как $\mathrm{C}_n = \pi \cdot \mathrm{SD2} \cdot \mathrm{SD1}$.

Допуская наличие влияния вегетативной нервной системы не только на каждый последующий RR-интервал (величина запаздывания $m = 1$), но и через задержки большей величины, в работе предложено исследовать зависимость функции $\mathrm{CCM}(m)$ от ряда значений $m = 1, ..., 10$.

В таком случае, каждую исследуемую запись ритма (каждый объект) можно представить 10-мерным вектором $\mathbf{X}_i = (\mathrm{CCM}(1), ..., \mathrm{CCM}(10))$. Для распознавания объектов класса ЗСН и НСР в 10-мерном пространстве признаков был применен линейный дискриминантный анализ. При исследовании двух классов находился опорный вектор $\mathbf{W} = (w_1, \ldots, w_{10})$, на котором проекции объектов $\mathrm{Y}_i = \mathbf{W}^{\mathrm{T}} \cdot \mathbf{X}_i$ показывали наибольшее удаление между классами. Для нахождения $\mathbf{W}$ и построения решающей функции использовался линейный дискриминант Фишера.

Помимо получения решающего правила, проведен сравнительный анализ статистической значимости различий показателей SD1, SD2, SD1/SD2 и CCM(1), а также проекций Y_i для 10-мерных векторов $\mathbf{X}_i$. С использованием статистического Т-теста Стьюдента было установлено, что наибольшей статистической мощностью для распознавания классов ЗСН и НСР обладают CCM(1) и Y_i. Так при критическом значении t-критерия, равном 2.002, его расчётное значение для показателя CCM(1) составило 17.85. В то же время показатели SD1, SD2 оказались слабо эффективными, поскольку расчётные значения t-критерия составили 0.59 и 4.15, соответственно.

Оценка качества классификации проведена с использованием специально подготовленной базы ритмограмм, включающей по 30 фрагментов для классов ЗСН и НСР длительностью 300 кардиоинтервалов [2]. Записи отобраны из двух сертифицированных баз данных портала PhysioNet [3]. При отборе фрагментов учитывалось отсутствие выбросов, тренда, а также экстрасистол. Показано, что использование полного описания функции CCM позволяет получить безошибочное распознавание ЗСН.

Как показали результаты исследования, для повышения эффективности распознавания класса ЗСН по параметрам сердечного ритма целесообразно использовать функцию CCM для разных временных задержек. При этом полученную совокупность значений этого показателя следует рассматривать как многомерный вектор, проекции которого на ось весового вектора дают наилучшее разделение объектов исследуемых классов.

Работа поддержана грантом РФФИ №19-29-01009.

[1] *Karmakar C. K.*, *Gubbi J.*, *Palaniswami M.* Complex Correlation Measure: a novel descriptor for Poincaré plot // BioMedical Engineering OnLine, 8(17), 2009. — С. 17–35.

[2] *Manilo L. A.*, *Kholmatov D. U.* Recognition of Congestive Heart Failure Based on a Complex Correlation Measure of the Heart Rate Signal // Pattern Recognition and Image Analysis, 32(3), 2022. — С. 586–590.

[3] *Goldberger A. A.*, *Amaral L.*, *Glass L.*, *Hausdorff J*, *Ivanov P. C.*, *Mark R.*, *Stanley H. E.* PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals // PCirculation, 101(23), 2000. — С. 215–220.

Recognition of congestive heart failure by the criterion of chaoticity of rhythmogram

Manilo Lyudmila[1] lmanilo@yandex.ru
Kholmatov Doston[1]⋆ xolmatov.2000@mail.ru
Nemirko Anatoly [1] apn-bs@yandex.ru
[1]Saint Petersburg, Saint Petersburg Electrotechnical University "LETI"

The work investigates the possibility of recognition of congestive heart failure (CHF) on the background of normal sinus rhythm (NSR) using Poincaré plots, which are a representation of heart rhythm in the space of phase coordinates. A special function of the complex correlation measure as well as statistical indicators are used as the main parameters characterizing the features of graphical data. The choice of such a description is due to the fact that in CHF due to the activation of compensatory mechanisms of the heart rhythm regulation system, complex chaotic changes in the rhythm structure are observed.

The Poincaré plot was represented as a two-dimensional scatter plot of consecutive points whose coordinates were the values of adjacent RR intervals of the heart rate. The geometric shapes of the scatter points in the Poincaré plot can be used to judge the presence of cardiac abnormalities. Standard statistical indicators SD1 and SD2 are usually used for this purpose. They characterize the dispersion of the scattering cloud along and perpendicular to the identity line, which is characteristic of some rhythm disturbances. However, they apply mainly to elliptic scattering shapes on the Poincaré plot, since they have a number of limitations. This approach does not recognize temporal changes in heart rhythm, the appearance of non-linear components and cluster formations, for example, in the form of "petals" which is typical for some types of arrhythmias. Therefore, in this paper, a new measure of complex correlation measure (CCM), reflecting the multiple correlation of delay, which is important for analyzing the structure of time series in CHF was investigated [1]. In addition, the values of frequently used standard statistical indicators were assessed.

For the classical Poincaré plot with delay $m = 1$, the indicator of the CCM is calculated windowly using a moving triangular window consisting of sequential points $\mathrm{P}_i(\mathrm{RR}_i;\mathrm{RR}_{i+1})$, $\mathrm{P}_{i+1}(\mathrm{RR}_{i+1};\mathrm{RR}_{i+2})$, $\mathrm{P}_{i+2}(\mathrm{RR}_{i+2};\mathrm{RR}_{i+3})$. On each i-th step there is a window area $\mathrm{S}(i)$ through the determinant of the matrix:

$$\mathrm{S}(i) = \frac{1}{2}\begin{vmatrix} \mathrm{RR}_i & \mathrm{RR}_{i+1} & 1 \\ \mathrm{RR}_{i+1} & \mathrm{RR}_{i+2} & 1 \\ \mathrm{RR}_{i+2} & \mathrm{RR}_{i+3} & 1 \end{vmatrix}.$$

Then the indicator of the SSM is calculated as:

$$\mathrm{CCM}(m) = \frac{1}{\mathrm{C}_n(N-2)} \sum_{i=1}^{N-2} |\mathrm{S}(i)|.$$

where N is the number of points on the Poincaré plot, m is the value of delay (lag), C_n is the geometric area of the ellipse, into which the scattering cloud fits and is calculated as $C_n = \pi \cdot SD2 \cdot SD1$.

Allowing the influence of the vegetative nervous system not only on each subsequent RR interval (the value of the delay $m = 1$), but also through delays of a larger size, it is proposed to investigate the dependence of the CCM(m) function on a number of values $m = 1, ..., 10$.

In this case, each studied rhythm record (each object) can be represented by a 10-dimensional vector $\mathbf{W} = (w_1, \ldots, w_{10})$. For recognition of objects of the CHF and NSR class in the 10-dimensional space of signs, linear discriminant analysis was used. In the study of two classes, there was a support vector $\mathbf{W} = (w_1, \ldots, w_{10})$, on which the projections of objects $Y_i = \mathbf{W}^T \cdot \mathbf{X}_i$ showed the greatest removal between the classes. To find $\mathbf{W}$ and building a decisive function, Fisher's linear discriminant was used.

Assessment of the quality of the classification was carried out using a specially prepared base of rhythmograms, which includes 30 fragments for classes of the CHF and NSR duration of 300 cardio intervals [2]. The records are selected from two certified databases of the Physionet portal [3]. When selecting fragments, the absence of outliers, a trend, and also extrasystoles was taken into account. It is shown that the use of a full description of the CCM function allows you to obtain unmistakable recognition of the CHF.

In addition to obtaining a decisive rule, a comparative analysis of the statistical significance of the differences of the indicators SD1, SD2, SD1/SD2 and CCM(1), as well as the projections of Y_i for 10-dimensional vectors $\mathbf{X}_i$, was carried out. Using the statistical t-test, it was found that the greatest statistical power for recognizing the classes of the CHF and the NSR have CCM(1) and Y_i. So with a critical value of t statistic, equal to 2.002, its calculated value for the indicator of the CCM(1) was 17.85. At the same time, the indicators of the SD1 and SD2 were weakly effective, since the calculated values of t statistic were 0.59 and 4.15, respectively.

As the results of the study have shown, it is advisable to use the function of CCM for different temporary delays to increase the efficiency of the CHF class by heart rhythm parameters. In this case, the resulting set of values of this indicator should be considered as a multidimensional vector, the projections of which on the axis of the weight vector give the best separation of objects of the studied classes.

This research is funded by RFBR, project 19-29-01009.

[1] *Karmakar C. K.*, *Khandoker A. H.*, *Gubbi J.*, *Palaniswami M.* Complex Correlation Measure: a novel descriptor for Poincaré plot // BioMedical Engineering OnLine, 8(17), 2009. — p. 17–35.

[2] *Manilo L. A.*, *Kholmatov D. U.* Recognition of Congestive Heart Failure Based on a Complex Correlation Measure of the Heart Rate Signal // Pattern Recognition and Image Analysis, 32(3), 2022. — p. 586–590.

[3] *Goldberger A. A.*, *Amaral L.*, *Glass L.*, *Hausdorff J*, *Ivanov P. C.*, *Mark R.*, *Stanley H. E.* PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals // PCirculation, 101(23), 2000. — p. 215–220.

Применение нейронных сетей для диагностирования очаговых образований в печени по изображениям и видеозаписям ультразвуковых исследований

Камгуя Феукви Херве[1]⋆ st093241@student.spbu.ru
Козубова Ксения Вячеславовна[2] dr.kozubova@yandex.ru
Бусько Екатерина Александровна[1,2] katrn@mail.ru
[1]Санкт-Петербург, Санкт-Петербургский государственный университет
[2]Санкт-Петербург, ФГБУ Научный медицинский исследовательский центр онкологии им. Н.Н. Петрова

Сверточные нейронные сети (CNN) являются мощным инструментом для извлечения особенностей из изображений и широко используются в обработке изображений. Видео - это набор изображений, записанных в разные моменты времени. Таким образом, видео имеет как пространственное, так и временное измерение. Для врача ультразвуковой диагностики, визуализирующего очаговое поражение печени и определяющего, является ли оно злокачественным или нет, важно изменение паренхимы данного органа в разных срезах и в различные временные промежутки. При выполнении ультразвукового исследования (УЗИ) печени важно визуализировать очаговое изменение эхогенности данного органа, которое трактуется как наличие поражения [1]. При УЗИ выделяют 4 вида эхогенности образования: анэхогенное- полностью поглощает ультразвуковые волны, чаще всего жидкостная структура, изоэхогенное-имеет такую же плотность как и неизменённая ткань, гипоэхогенное- структурное изменение ткани исследуемого органа, характеризующееся сниженной плотностью и, соответственно, гиперэхогенное- образование, имеющее высокую плотность для прохождения ультразвуковых волн. В работе рассматриваются прогностические модели, способные извлекать признаки и учитывать временное измерение.

Для исследования возможностей применения нейронных сетей для диагностирования метастатического поражения печени по данным ультразвуковых исследований специалистами проводится обследование пациентов на базе ФГБУ «НМИЦ онкологии им. Н.Н. Петрова» Минздрава России. В процессе исследования создается набор данных, содержащий видеозаписи длительностью от 7 до 30 секунд, к каждой видеозаписи врач ультразвуковой диагностики создает текстовый документ с описанием состояния печени на основании следующих ультразвуковых параметров: оценка структуры и эхогенности паренхимы, оценка сосудистой архитектоники органа, выявление признаков диффузных и очаговых изменений, а так же при наличии очагового поражения полное описание изменений согласно общеутверждённым нормам. В Таблице 1 приводится краткое описание собранных видео.

Тип	**Количество видео записей**
Нормальная печень	17
Диффузные изменения печени (жировой гепатоз) без очаговых образований	8
Добкачественные очаговые образования (кисты, гемангиомы)	5
Очаги метастазирования	8

Таблица 1. Описание набора данных.

В общей сложности было получено 38 видео, выполненных либо при продольном сканировании, либо визуализация из поперечного сканирования. Из каждого видео извлекается максимальное количество кадров. Полученный набор разделяется на обучающий, валидационный и тестовый (60%, 20%, 20% соответственно).

Первая модель представляет собой комбинацию слоев CNN [2] и слоя с распределенным временем (Time Distributed Layer). Слои CNN свертки применяются к кадрам для вычисления объектов только из пространственных измерений. Распределенный по времени слой позволяет применять слой в "N"заданных временных измерениях. Это означает, что наша модель после идентификации части печени (например, поражения) будет пытаться узнать любое изменение, которое произойдет в следующих "N"кадрах. В этой модели операции свертки применяются отдельно на разных кадрах видео, таким образом, веса обучаются отдельно, но во временной последовательности.

Вторая модель - это 3D CNN-модель [3]. Эта модель применяет свертку как пространственную, так и временную. Разница между этой моделью и предыдущей в том, что операция свертки выполняется одновременно на всех кадрах.

Целью данного исследования является изучение возможностей применения нейронных сетей для диагностирования очаговых образований печени и их дифференциальной диагностики по данным ультразвуковых исследований

Сравнение результатов обоих моделей говорит о наличии связи качества предсказания и динамики кровотока в различной фазе. Предполагается собрать объединенную модель на базе двух предложенных моделей. Далее предполагается использовать карту псевдоколоризации, чтобы очертить области, идентифицированные моделью как поражения.

На основе результатов, полученных при тестировании каждой модели по отдельности, каждой модели будет присвоен вес в объединенной модели. Кроме того, в процессе дальнейшей работы планируется увеличить базу видео изображений для проведения дальнейшего тестирования.

[1] *Бусько Е. А., Козубова К. В. и др.* Сравнительный анализ эффективности КТ и контрастно-усиленного УЗИ в диагностике метастазов колоректального рака печени // Анналы хирургической гепатологии. 2022. Т. 27. № 1., С. 22–32.

[2] *Siyuan Z., Yifan W. et al.* CNN-Based Medical Ultrasound Image Quality Assessment. // Hindawi Complexity. 2021. Article ID: 9938367. 9 p.

[3] *Ji S., Xu W., Yang M., Yu K.* 3D Convolutional Neural Networks for Human Action Recognition. // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2013. Vol. 35. no.1 p. 221-231.

Application of neural networks to photos and recordings of ultrasound exams to identify focal forms in the liver

Kamguia Feukwi[1]⋆ st093241@student.spbu.ru
Kozubova Ksenia[2] dr.kozubova@yandex.ru
Busko Ekaterina[1,2] katrn@mail.ru
[1]Saint Petersburg, Saint Petersburg State University
[2]Saint Petersburg, N.N. Petrov National Medicine Research Center of oncology

Convergent Neural Networks (CNNs) are powerful tools for extracting features from images and are widely used in image processing. A video is a set of images recorded at different time steps. Thus, video has both spatial and temporal dimensions. For an ultrasound diagnostic physician, imaging a focal liver lesion and determining whether it is malignant or not, the change in the parenchyma of that organ in different frames and at different time intervals is important. When performing an ultrasound examination of the liver, the visualisation of a focal change in the echogenicity of this organ, can be interpreted as the presence of a lesion [1]. Four different formation echogenicity categories are identified during ultrasound: Hyperechoic is a formation with a high density for the passage of ultrasonic waves, whereas anechoic completely absorbs ultrasonic waves and is typically a liquid structure. Other types of formations include isoechoic, which has the same density as the surrounding tissue, hypoechoic, which refers to structural changes in the tissue of the organ under study. The paper considers predictive models capable of extracting features while taking into account the temporal dimension.

To investigate the possibility of using neural networks to diagnose metastatic liver lesions based on ultrasound examinations, specialists examine patients at N.N. Petrov National Medicine Research Center of oncology in Saint Petersburg. During the study a set of data is created containing video recordings with duration of 7 to 30 seconds. For each video recording, the ultrasound diagnostic physician creates a text document describing the state of the liver based on the following ultrasound parameters: assessment of the structure and echogenicity of the parenchyma, assessment of the vascular architectonics of the organ, identification of signs of diffuse and focal changes, as well as the presence of a focal lesion, and a complete description of the changes according to generally approved standards. The brief description of the collected videos is presented in the Table 1.

A total of 38 videos were obtained from either the longitudinal scan or the imaging from the cross-sectional scan. The maximum number of frames is extracted from each video. The resulting set is divided into a training, validation, and test set (60%, 20%, and 20%, respectively).

The first model is a combination of CNN layers [2] and Time Distributed Layer. CNN convolution layers are applied to frames to compute objects from spatial dimensions only. The Time Distributed Layer allows to apply the layer in "N" given time dimensions. This means that our model, after identifying a part of the liver

Category	Number of video recordings
Normal liver	17
Diffuse liver changes (fatty hepatosis) without focal formations	8
Benign focal formations (cysts, hemangiomas)	5
Metastasis formations	8

Table 1. Description of the data set.

(e.g., a lesion), will try to learn any change that occurs in the next "N" frames. In this model, convolution operations are applied separately on different frames of the video, so the weights are trained separately, but in a temporal sequence.

The second model is the 3D CNN model [3]. This model makes use of both spatial and temporal convolution. The difference between this model and the previous one is that the convolution operation is performed on all frames simultaneously.

The goal of this research is to look into the possibilities of using neural networks to diagnose focal liver formations and their differential diagnosis based on ultrasound studies.

A comparison of the results of both models indicates that there is a connection between the quality of prediction and the dynamics of blood flow in different phases. A combined model based on the two proposed models will be built. Next, a saliency map will be used to outline the areas identified by the model as lesions.

Based on the results obtained by testing each model separately, each model will be assigned a weight in the combined model. In addition, in the process of further work it is planned to increase the base of video images for further testing.

[1] *Busko E. A.,Kozubova K. V. et al.* Comparative assessment of diagnostic value of computed tomography and contrast-enhanced ultrasound in colorectal cancer liver metastases diagnosis. //
Annaly khirurgicheskoy gepatologii = Annals of HPB Surgery. 2022. Vol. 27. No. 1. p. 377-382. (In Russ.)

[2] *Siyuan Z., Yifan W., Jiayao J., Jingxian D., Weiwei Y., and Wenguang H.* CNN-Based Medical Ultrasound Image Quality Assessment. // Hindawi Complexity. 2021.

[3] *Ji S., Xu W., Yang M., Yu K.* 3D Convolutional Neural Networks for Human Action Recognition. // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2013. Vol. 35. No.1 p. 221–231.

Программный комплекс для прямого моделирования данных электрофизиологической активности

Бойко Анна Ивановна[1]⋆ a.boyko@list.ru
Рыкунов Станислав Дмитриевич[1] rykunov@impb.ru
Устинин Михаил Николаевич[1] u_m_n@mail.ru

[1] Пущино, ИМПБ РАН – филиал ИПМ им. М.В. Келдыша РАН

Разработан комплекс программ для компьютерного моделирования многоканальных временных рядов, регистрируемых в различных экспериментах по изучению электромагнитных полей, порождаемых человеческим телом. В качестве моделей приборов могут быть использованы наборы координат и направлений магнитных энцефалографов нескольких типов, электроэнцефалографов и магнитных кардиографов. Программный комплекс обеспечивает: - интерактивное размещение источников поля в пространстве головы или тела; - редактирование амплитудно-временной зависимости; - пакетную загрузку большого количества источников; -моделирование шумов; -моделирование малоканальных планарных магнитометров различных порядков, с заданием формы прибора, количества датчиков и их параметров. Для изучения мозга человека в качестве моделей головы используются магнитно-резонансные томограммы, для изучения сердца используется модель тела в виде полупространства с плоской границей. Источники размещаются в модельном пространстве, для них решается прямая задача в физической модели, соответствующей используемому прибору. Для каждого источника задается временная зависимость и вычисляется многоканальный временной ряд. Затем производится суммирование временных рядов от всех источников и добавляется шумовая компонента. Программа состоит из трех модулей: модуля ввода-вывода, расчетного модуля и модуля визуализации. Модуль ввода-вывода отвечает за загрузку моделей приборов, моделей головного мозга и параметров источников поля. Расчетный модуль отвечает за непосредственный расчет поля и преобразование координат между индексной системой и системой головы. Модуль визуализации отвечает за изображение модели мозга, положения источников поля, графического представления амплитудно-временной зависимости источников поля и рассчитанных значений суммарного поля. Рассмотрены магнитные и электрические поля, производимые источниками в зонах мозга, ответственных за обработку речевых стимулов. Полученный многоканальный сигнал может использоваться для тестирования различных методов анализа данных и планирования экспериментов.

Работа поддержана грантами РФФИ 20-07-00733, 20-07-00842.

[1] *Бойко А. И., Рыкунов С. Д., Устинин М. Н.* Программный комплекс для моделирования данных электрофизиологической активности // Математическая биология и биоинформатика, 2022. —Т. 17(1) —С. 1–9.

A Software Package for the Direct Modeling of Electrophysiological Activity Data

Boyko Anna[1]⋆ a.boyko@list.ru
Rykunov Stanislav[1] rykunov@impb.ru
Ustinin Mikhail[1] u_m_n@mail.ru
[1]Pushchino, IMPB RAS - Branch of KIAM RAS

A complex of programs has been developed for computer modeling of multichannel time series recorded in various experiments on electromagnetic fields created by the human body. Sets of coordinates and directions of sensors for magnetic encephalographs of several types, electroencephalographs and magnetic cardiographs are used as models of devices. To study the human brain, magnetic resonance tomograms are used as head models; to study the heart, a body model in the form of a half-space with a flat boundary is used. The sources are placed in the model space, for them the direct problem is solved in the physical model corresponding to the device used. For a magnetic encephalograph and an electroencephalograph, an equivalent current dipole model in a spherical conductor is used, for a magnetic cardiograph, an equivalent current dipole model in a flat conductor or a magnetic dipole model is used. For each source, a time dependence is set and a multichannel time series is calculated. Then the time series from all sources are summed and the noise component is added. The program consists of three modules: an input-output module, a calculation module and a visualization module. The input-output module is responsible for loading device models, brain models, and field source parameters. The calculation module is responsible for directly calculating the field and transforming coordinates between the index system and the head system. The visualization module is responsible for the image of the brain model, the position of the field sources, a graphical representation of the amplitude-time dependence of the field sources and the calculated values of the total field. The user interface has been developed. The software package provides: interactive placement of field sources in the head or body space and editing of the amplitude-time dependence; batch loading of a large number of sources; noise modeling; simulation of low-channel planar magnetometers of various orders, specifying the shape of the device, the number of sensors and their parameters. Magnetic and electric fields produced by sources in the brain areas responsible for processing speech stimuli are considered. The resulting multichannel signal can be used to test various data analysis methods and for the experiment planning.

This research is funded by RFBR, grant 20-07-00733, 20-07-00842.

[1] *Boyko A. I., Rykunov S. D., Ustinin M. N.* A Software Package for the Modeling of Electrophysiological Activity Data // Mathematical Biology and Bioinformatics, 2022. — V. 17(1)— p. 1–9.

Разделение магнитной энцефалограммы на "мозговые" и "внемозговые" физиологические сигналы на основе совместного анализа функциональных томограмм и магнитно-резонансных томограмм

Рыкунов Станислав Дмитриевич[1]⋆ rykunov@impb.ru
Бойко Анна Ивановна[1] a.boyko@list.ru
Устинин Михаил Николаевич[1] u_m_n@mail.ru
[1]Пущино, ИМПБ РАН – филиал ИПМ им. М.В. Келдыша РАН

В работе рассматривается задача разделения данных энцефалографии на два временных ряда, генерируемых мозгом и генерируемых другими электрическими источниками, расположенными в голове человека. Магнитные энцефалограммы и магнитно-резонансные томограммы головы были записаны в Центре нейромагнетизма Медицинской школы имени Гроссмана при Нью-Йоркском университете. Также использовались данные, полученные в Университете Макгилла и Монреальском университете. Запись производилась в помещении с магнитным экранированием, а конструкция градиентометров предусматривала подавление внешних шумов, что позволяло исключить их из анализа данных. Магнитные энцефалограммы были проанализированы методом функциональной томографии, основанным на преобразовании Фурье и решении обратной задачи для всех частот. В этом методе каждому частотному компоненту присваивается одно пространственное положение. По магнитно-резонансные томограммам головы было посторено аннотированое пространство для анализа. Это пространство было разделено на две части: «мозговую» и «немозговую». Частотные составляющие классифицировались по признаку их включения в ту или иную часть. Набор частот, обозначенный как «мозг», представлен частичным спектром сигнала мозга, тогда как набор частот, обозначенный как «не мозг», представлен частичным спектром физиологического шума, производимого головой. Оба парциальных спектра имеют одну и ту же полосу частот. Из парциальных спектров были реконструированы временные ряды сигнала области «мозга» и шума головы «не мозговой» области. Суммарная спектральная мощность сигнала оказалась в десять раз больше шума. Предлагаемый метод позволяет детально анализировать как сигнальную, так и шумовую составляющие энцефалограммы, фильтровать магнитную энцефалограмму.

Работа поддержана грантами РФФИ 20-07-00733, 20-07-00842.

[1] *Llinás R. R., Rykunov S., Walton K. D., Boyko A., Ustinin M.* Splitting of the magnetic encephalogram into «brain» and «non-brain» physiological signals based on the joint analysis of frequency-pattern functional tomograms and magnetic resonance images // Frontiers in Neural Circuits, 2022. — V. 16.

Splitting of the magnetic encephalogram into “brain” and “non-brain” physiological signals based on the joint analysis of frequency-pattern functional tomograms and magnetic resonance images

Rykunov Stanislav[1]⋆ rykunov@impb.ru
Boyko Anna[1] a.boyko@list.ru
Ustinin Mikhail[1] u_m_n@mail.ru

[1]Pushchino, IMPB RAS - Branch of KIAM RAS

The work considers the problem of dividing the encephalography data into two time series, that generated by the brain and that generated by other electrical sources located in the human head. The magnetic encephalograms and magnetic resonance images of the head were recorded in the Center for Neuromagnetism at NYU Grossman School of Medicine. Data obtained at McGill University and Montreal University were also used. Recordings were made in a magnetically shielded room and the gradiometers were designed to suppress external noise, making it possible to eliminate them from the data analysis. Magnetic encephalograms were analyzed by the method of functional tomography, based on the Fourier transform and on the solution of inverse problem for all frequencies. In this method, one spatial position is assigned to each frequency component. Magnetic resonance images of the head were evaluated to annotate the space to be included in the analysis. The included space was divided into two parts: ≪brain≫ and ≪non-brain≫. The frequency components were classified by the feature of their inclusion in one or the other part. The set of frequencies, designated as ≪brain≫, represented the partial spectrum of the brain signal, while the set of frequencies designated as ≪non-brain≫, represented the partial spectrum of the physiological noise produced by the head. Both partial spectra shared the same frequency band. From the partial spectra, a time series of the ≪brain≫ area signal and ≪non-brain≫ area head noise were reconstructed. Summary spectral power of the signal was found to be ten times greater than the noise. The proposed method makes it possible to analyze in detail both the signal and the noise components of the encephalogram and to filter the magnetic encephalogram.

This research is funded by RFBR, grant 20-07-00733, 20-07-00842.

[1] *Llinás R. R., Rykunov S., Walton K. D., Boyko A., Ustinin M.* Splitting of the magnetic encephalogram into ≪brain≫ and ≪non-brain≫ physiological signals based on the joint analysis of frequency-pattern functional tomograms and magnetic resonance images // Frontiers in Neural Circuits, 2022. — V. 16.

Структурный мотив 3β-уголок

Руднев Владимир Ремович[1]$\star$ v.r.rudnev@gmail.com
Никольский Кирилл Сергеевич[1] glucksistemi@gmail.com
Петровский Денис Витальевич[1] petro2017@gmail.com
Куликова Людмила Ивановна1 likulikova@mail.ru
Мальсагова Кристина Ахмедовна1 kristina.malsagova86@gmail.com
Кайшева Анна Леонидовна1 kaysheva3@gmail.com

[1]Группа Биобанкинга, Обособленное подразделение «Научно-практический образовательный центр» Федерального государственноого бюджетного научного учреждения «Научно-исследовательский институт биомедицинской химии имени В.Н. Ореховича» (ИБМХ)

Структурные мотивы белковых молекул имеют уникальные укладки составляющих их последовательно идущих по полипептидной цепи элементов вторичных структур и образуют компактные пространственные конструкции. Уникальность структур и их возможность быть зародышами в процессе сворачивания белков или использованными в качестве стартовых структур для поиска возможных укладок полипептидной цепи при моделировании структуры белков определяют актуальность и важность всестороннего изучения структурных мотивов. Полученные в ходе исследований знания могут быть весьма полезны для решения как фундаментальных, так и прикладных задач.

Данная работа посвящена исследованию поведения структурного мотива 3β-уголок в эксперименте молекулярной динамики. Структурный мотив 3β-уголок имеет уникальную укладку в пространстве и часто встречается как в гомологичных, так и негомологичных белках.

На первом этапе нами была решена задача создания набора данных – представительной выборки объектов исследования, состоящей из 330 3β-уголков. Данные структуры были распознаны и отобраны из белковых структур, аннотированных в банке белковых структур PDB (https://www.rcsb.org/). Основным инструментом поиска, распознавания и сегментации 3β-уголков выступили нейронные сети. Нами предложена модель с новой сетевой архитектурой глубокого обучения, использующей интегративную синергию графовых нейронных сетей, сверточных нейронных сетей и рекуррентных нейронных сетей. Для сегментации структурных мотивов этот метод машинного обучения использует основные характеристики геометрии мотивов — длины элементов вторичных структурных и расстояния между ними, торсионные углы, координаты Cα-атомов и др.

На следующем этапе с помощью эксперимента молекулярной динамики показана автономная устойчивость исследуемого структурного мотива в водной среде (вне белковой глобулы). В эксперименте участвовали 330 структурных мотивов, отобранных из негомологичных белков различного происхождения. В ходе эксперимента молекулярной динамики 3β-уголки сохранили основные свои геометрические характеристики: радиус гирации, доступная для растворителя

площадь, торсионные углы, количество водородных связей. На данном этапе исследования также были решены следующие задачи:

— сравнительный анализ структурных параметров мотива 3β-уголок в белке и автономно (вне белка) в ходе эксперимента молекулярной динамики;
— оптимизация продолжительности молекулярно-динамического эксперимента;
— анализ взаимодействий, участвующих в стабилизации структурного мотива 3β-уголок.

Таким образом, основные результаты, полученные в ходе данной работы:

✓ структурный мотив 3β-уголок является автономно стабильной структурой;
✓ 3β-уголок может выступать как самостоятельный объект для изучения в области структурной биологии.

Работа выполнена в рамках Программы фундаментальных научных исследований в Российской Федерации на долгосрочный период (2021–2030 годы) ($No.$122092200056-9).

[1] *Vladimir R. Rudnev, Kirill S. Nikolsky, Denis V. Petrovsky, Liudmila I. Kulikova, Anton M. Kargatov, Kristina A. Malsagova, Alexander A. Stepanov, Arthur T. Kopylov, Anna L. Kaysheva and Alexander V. Efimov* 3β-Corner Stability by Comparative Molecular Dynamics Simulations // Int. J. Mol. Sci., 2022.

Structural motif 3β-corner

Rudnev Vladimir[1] v.r.rudnev@gmail.com
Nikolsky Kirill[1] glucksistemi@gmail.com
Petrovsky Denis[1] petro2017@gmail.com
Kulikova Lyudmila1 likulikova@mail.ru
Malsagova Kristina1 kristina.malsagova86@gmail.com
Kaysheva Anna1 kaysheva3@gmail.com

[1]Biobanking Group, Branch of Institute of Biomedical Chemistry "Scientific and Education Center", 109028 Moscow, Russia

Structural motifs of protein molecules have unique arrangements of their secondary structures constituent elements sequentially going along the polypeptide chain and form compact spatial structures. The uniqueness of the structures and their ability to be starting points of the protein folding process or used as starting structures to search for possible folds of the polypeptide chain when modeling protein structures determine the relevance and importance of structural motifs comprehensive study. The knowledge gained in the course of research can be very useful solve both fundamental and applied problems.

This work is devoted to study the behavior of the 3β-corner structural motif in a molecular dynamics experiment. The structural motif of the 3β-corner has a unique fold in space and is often found in both homologous and non-homologous proteins.

At the first stage, we solved the dataset creation problem - a representative sample of research objects, consisting of 330 3β-corners. These structures were recognized and selected from protein structures annotated in the PDB protein structure bank (https://www.rcsb.org/). The main tool for searching, recognizing and segmenting 3β-corners was neural networks. We have proposed a model with a new deep learning network architecture using the integrative synergy of graph neural networks, convolutional neural networks, and recurrent neural networks. For segmentation of structural motifs, this machine learning method uses the main characteristics of the geometry of motifs — the lengths of secondary structural elements and the distance between them, torsion angles, coordinates of Cα-atoms, etc.

At the next stage, using a molecular dynamics experiment, the autonomous stability of the studied structural motif in an aqueous medium (outside the protein globule) was shown. The experiment involved 330 structural motifs selected from nonhomologous proteins of various origins. In the course of the molecular dynamics experiment, the 3β-corners retained their main geometric characteristics: the gyration radius, the solvent-accessible area, torsion angles, and the number of hydrogen bonds. At this stage of the study, the following tasks were also solved:

— comparative analysis of the structural parameters of the 3β-corner motif in the protein and autonomously (outside the protein) during a molecular dynamics experiment;

— molecular dynamics experiments duration optimization;
— analysis of the interactions involved in 3β-corner structural motif stabilization process.

The main results obtained in this work's course are:

✓ the structural motif 3β-corner is an autonomously stable structure;
✓ 3β-corner can act as an independent object for study in the field of structural biology.

This research is funded within the framework of the Long-Term Program of Fundamental Scientific Research in the Russian Federation (2021–2030) (No.122092200056-9).

[1] *Vladimir R. Rudnev, Kirill S. Nikolsky, Denis V. Petrovsky, Liudmila I. Kulikova, Anton M. Kargatov, Kristina A. Malsagova, Alexander A. Stepanov, Arthur T. Kopylov, Anna L. Kaysheva and Alexander V. Efimov* 3β-Corner Stability by Comparative Molecular Dynamics Simulations // Int. J. Mol. Sci., 2022.

Исследование нейрофизиологических закономерностей болезни Паркинсона на первой стадии с помощью метода анализа всплескообразной электрической активности мышц

Сушкова Ольга Сергеевна[1]⋆ o.sushkova@mail.ru
Морозов Алексей Александрович[1] morozov@cplire.ru
Габова Александра Васильевна[2] agabova@yandex.ru
Чигалейчик Лариса Анатольевна[3] chigalei4ick.lar@yandex.ru
Карабанов Алексей Вячеславович[3] doctor.karabanov@mail.ru

[1]Москва, ИРЭ им. В.А. Котельникова РАН
[2]Москва, ИВНД и НФ РАН
[3]Москва, ФГБНУ «Научный центр неврологии»

Для исследования нейрофизиологических закономерностей (НЗ) в электромиографических (ЭМГ) сигналах у пациентов с болезнью Паркинсона (БП) был применён инструментарий разработанного ранее авторами метода анализа всплескообразной электрической активности (ВЭА). Изначально параметры ВЭА в сигналах изучались авторами в качестве метрик, позволяющих эффективно распознавать некоторые нейродегенеративные заболевания, такие как БП и эссенциальный тремор (ЭТ). В настоящем исследовании показано, что данные метрики отражают НЗ протекания БП на 1 стадии. Были исследованы параметры разных видов тремора, а именно, паркинсонического (3–7 Гц) (ПТ) и физиологического (8–20 Гц) (ФТ) тремора. С помощью анализа ВЭА была обнаружена отрицательная корреляция между количеством всплесков ПТ и ФТ у пациентов на 1 стадии БП. Кроме того, была обнаружена корреляция между количеством всплесков ФТ и возрастом пациентов. Выявленные НЗ проливают свет на нейрофизиологические процессы, протекающие в коре головного мозга, и позволяют выдвинуть гипотезы, уточняющие механизмы взаимодействия различных структур мозга у пациентов с БП.

Исследование выполнено за счёт гранта Российского научного фонда No. 22-75-10079, https://rscf.ru/project/22-75-10079/.

[1] *Sushkova O. S., Morozov A. A., Gabova A. V., Karabanov A. V., Illarioshkon S. N.* The method of wave train electrical activity analysis for investigation of neurophysiological regularities of Parkinson's disease at the first stage // Proceedings of PhREME–2022. — Vladimir, 2022. — p. 46–50.

[2] *Sushkova O. S., Morozov A. A., Petrova N. G., Khokhlova M. N., Gabova A. V., Karabanov A. V., Chigaleichik L. A., Sarkisova K. Y.* Method of wave train electrical activity analysis – the theoretical basis and application // RENSIT, 2022. — V. 14. — Issue. 3. — p. 317–330. `http://en.rensit.ru/vypuski/article/457/14(3)317-330e.pdf`.

[3] *Sushkova O. S., Morozov A. A., Gabova A. V., Karabanov A. V., Illarioshkin S. N.* A statistical method for exploratory data analysis based on 2D and 3D area under curve diagrams: Parkinson's disease investigation // Sensors, MDPI, 2021. — V. 21. — Issue. 14. — p. 4700.

The investigation of neurophysiological regularities of Parkinson's disease at the first stage using the method of wave train electrical activity analysis of muscles

Sushkova Olga[1]⋆ o.sushkova@mail.ru
Morozov Alexei[1] morozov@cplire.ru
Gabova Alexandra[2] agabova@yandex.ru
Chigaleichick Larisa[3] chigalei4ick.lar@yandex.ru
Karabanov Alexei[3] doctor.karabanov@mail.ru

[1]Moscow, Kotel'nikov IRE RAS
[2]Moscow, IHNA&NPh RAS
[3]Moscow, FSBI "Research Center of Neurology"

We used the tools of the method developed earlier by the authors for analyzing wave train electrical activity to investigate the neurophysiological regularities in electromyographic (EMG) signals in patients with Parkinson's disease (PD). Initially, the parameters of wave train electrical activity in signals were studied by the authors as metrics to effectively recognize some neurodegenerative diseases, such as PD and essential tremor (ET). In the present study, we have shown that these metrics reflect the neurophysiological regularities of the course of PD at the 1 stage. The parameters of different types of tremor were investigated, namely, PD (3–7 Hz) and physiological (8–20 Hz) tremor. A negative correlation was found between the number of wave trains of PD and physiological tremor in patients at the 1 stage of PD using the analysis of wave train electrical activity. In addition, a correlation was found between the number of wave trains of physiological tremor and the age of patients. The revealed regularities shed light on the neurophysiological processes occurring in the cerebral cortex and make it possible to put forward hypotheses that explain the mechanisms of interaction between various brain structures in patients with PD.

This research is funded by Russian Science Foundation No. 22-75-10079, https://rscf.ru/en/project/22-75-10079/.

[1] *Sushkova O. S., Morozov A. A., Gabova A. V., Karabanov A. V., Illarioshkon S. N.* The method of wave train electrical activity analysis for investigation of neurophysiological regularities of Parkinson's disease at the first stage // Proceedings of PhREME–2022. — Vladimirl, 2022. — p. 46–50. http://freme.vlsu.ru/doc/works/FREME_2022_ISBN.pdf.

[2] *Sushkova O. S., Morozov A. A., Petrova N. G., Khokhlova M. N., Gabova A. V., Karabanov A. V., Chigaleichik L. A., Sarkisova K. Y.* Method of wave train electrical activity analysis – the theoretical basis and application // RENSIT, 2022. — V. 14. — Issue. 3. — p. 317–330.

[3] *Sushkova O. S., Morozov A. A., Gabova A. V., Karabanov A. V., Illarioshkin S. N.* A statistical method for exploratory data analysis based on 2D and 3D area under curve diagrams: Parkinson's disease investigation // Sensors, MDPI, 2021. — V. 21. — Issue. 14. — p. 4700.

Методы комплексирования данных при нейросетевом решении обратной задачи разведочной геофизики

Исаев Игорь Викторович[1,2]⋆ isaev_igorl@mail.ru
Оборнев Иван Евгеньевич[1] o_ivano@mail.ru
Оборнев Евгений Александрович[3] eugenyo@mail.ru
Родионов Евгений Александрович[3] evgeny_980@list.ru
Шимелевич Михаил Ильич[3] Shimelevich-M@yandex.ru
Доленко Сергей Анатольевич[1] dolenko@srd.sinp.msu.ru

[1]Москва, НИИ ядерной физики имени Д.В. Скобельцына МГУ имени М.В. Ломоносова
[2]Москва, Институт радиотехники и электроники им. В.А.Котельникова РАН
[3]Москва, Российский государственный геологоразведочный университет имени Серго Орджоникидзе

Обратные задачи разведочной геофизики заключаются в восстановлении пространственного распределения свойств среды в толще Земли по геофизическим полям, измеряемым на ее поверхности. В данной работе рассматриваются обратные задачи гравиметрии, магнитометрии и магнитотеллурического зондирования, а также их комплексирование, подразумевающее одновременное использование различных геофизических полей для восстановления искомого распределения.

Для обеспечения возможности комплексирования различных геофизических методов необходимо, чтобы определяемые параметры для каждого из методов были одинаковыми. Это может быть достигнуто пространственной постановкой задачи, в которой задачей является определение границ геофизических объектов. В предыдущих исследованиях мы рассматривали схему параметризации, где обратная задача заключалась в определении нижней границы геологических слоев, а каждый слой характеризовался переменными значениями глубины нижней границы по разрезу и фиксированными значениями плотности, намагниченности, и удельного сопротивления, как для слоя, так и для всего набора данных. Было показано, что комплексирование геофизических методов дает значительно лучшие результаты, чем использование каждого из методов по отдельности.

Настоящая работа является продолжением работ в данном направлении и здесь рассматривалась схема параметризации с фиксированными свойствами среды внутри слоя и переменными свойствами слоев по набору данных.

Исследование выполнено за счёт гранта Российского Научного фонда, проект № 19-11-00333, https://rscf.ru/project/19-11-00333/.

Data integration methods for neural network solution of the exploration geophysics inverse problem

Isaev Igor[1,2]⋆ isaev_igorl@mail.ru
Obornev Ivan[1] o_ivano@mail.ru
Obornev Eugeny[3] eugenyo@mail.ru
Rodionov Eugeny[3] evgeny_980@list.ruu
Shimelevich Mikhail[3] Shimelevich-M@yandex.ru
Dolenko Sergey[1] dolenko@srd.sinp.msu.ru

[1]Moscow, D.V. Skobeltsyn Institute of Nuclear Physics,
M.V. Lomonosov Moscow State University
[2]Moscow, Kotelnikov Institute of Radioengineering and Electronics,
Russian Academy of Sciences
[3]Moscow, Sergo Ordjonikidze Russian State University for Geological Prospecting

The inverse problems of exploration geophysics consist in reconstructing the spatial distribution of the medium properties in the thickness of the earth from the geophysical fields that are measured on its surface. This study considers the inverse problems of gravimetry, magnetometry and magnetotelluric sounding, and also their integration, which means simultaneous use of various geophysical fields to reconstruct the desired distribution.

To provide the possibility of the integration of various geophysical methods, it is necessary that the determined parameters for each of the methods are the same. This may be achieved by the spatial statement of the problem, in which the task is to determine the boundaries of geophysical objects. In the previous studies, we consider the parameterization scheme where the inverse problem was to determine the lower boundary of the geological layers, and each layer was characterized by variable values of the depth of the lower boundary along the section and fixed values of density, magnetization, and resistivity, both for the layer and for the entire data set. It is demonstrated that integration of geophysical methods provides significantly better results that use of each of the methods separately.

This study is a continuation of work in this direction and here we considered a parameterization scheme with fixed properties of medium within a layer and variable properties over a data set.

This study has been performed at the expense of the grant of the Russian Science Foundation (project no. 19-11-00333), https://rscf.ru/en/project/19-11-00333/.

Многоагентная иерархическая маршрутизация с временными окнами

Козлова Маргарита Геннадьевна[1] kozlovamg@cfuv.ru
Лукьяненко Владимир Андреевич[1] art-inf@yandex.ru
Макаров Олег Олегович[1]⋆ fantom2.00@mail.ru

[1]Симферополь, ФГАОУ ВО «Крымский федеральный университет имени В. И. Вернадского»

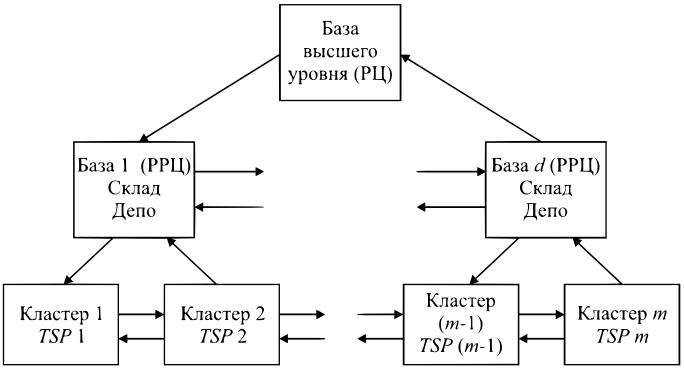
Рассматривается задача моделирования реальных логистических систем, устроенных иерархическим образом. Один из вариантов такой задачи предложен в [3]. Формируются кластеры потребителей нижнего уровня, отвечающие ограничениям временных окон для каждого потребителя и кластера в целом. На каждом таком кластере строится маршрут агента-коммивояжера и выделяется вершина, наиболее близкая к центральному узлу, которая является вершиной перегрузки товара с большегрузных транспортных средств (ТС) на малогрузные ТС, обслуживающие кластер потребителей. Вершины перевалки, в свою очередь, объединяются в маршруты коммивояжера более высокого уровня с учетом временных окон для маршрутов этого уровня. Программная реализация тестируется на известных сетях. Методика применима для синтеза центрального распределительного центра и системных распределительных центров нижнего уровня, а также для расчета необходимого числа транспортных средств (агентов).

Обзор источников показывает перспективность применения методов кластеризации для снижения размерности задачи и использования метаэвристик для TSP (Traveling Salesman Problem) [4] с временными окнами. В данной работе исследуется иерархическая модель маршрутизации с окнами.

Рассматривается двухуровневая логистическая сеть по доставке грузов. Предполагается наличие центрального склада или распределительного центра (РЦ), с которого груз доставляется на региональные распределительные центры (РРЦ) – склады, перевалочные пункты, с которых в свою очередь грузы распределяются по потребителям. Вариант такой структуры сети приведен на рисунке. Региональный центр (база высшего уровня) перераспределяет грузы по базам нижнего уровня (РРЦ), которые обслуживают один или несколько кластеров потребителей. Такую задачу будем называть иерархической многоагентной задачей коммивояжера HmTSP (Hierarchical multiple Traveling Salesman Problem).

Задачу HmTSP с временными окнами – временными ограничениями на доставку грузов, будем обозначать HmTSPTW. Временные окна (Time Windows, TW) могут быть связаны с режимом погрузочно-разгрузочных работ, перевозкой скоропортящихся грузов и др. Во многих случаях включение временных окон в математическую модель является обязательным. Рассматриваются варианты обхода кластера с учетом требуемого количества агентов (ТС) и потребностей клиентов (потребителей). Для численной реализации реше-

ния задачи маршрутизации ТС с временными окнами и несколькими маршрутами MVRPTW (Multi-Vehicle Routing Problem with Time Windows) выбран алгоритм, основанный на системе кооперации муравьиных колоний MACS-VRPTW (A Multiple Ant Colony System For Vehicle Routing Problems With Time Windows). Близкий алгоритм основан на многороевой кооперативной оптимизации роя частиц (A multi-swarm cooperative particle swarm optimizer, MCPSO), т.е. MCPSO-VRPTW (A multi-swarm cooperative particle swarm optimizer for Vehicle Routing Problem with Time Windows). MACS-VRPTW основана на системе муравьиных колоний (an Ant Colony System, ACS) [1, 2], и, в более общем смысле, на оптимизации муравьиной колонии (an Ant Colony Optimization, ACO), метаэвристическом подходе, основанном на поведении реальных колоний муравьев.



Иерархическая $mTSP$ с m кластерами, d базами и базой высшего уровня

Проблема маршрутизации транспортных средств с временными окнами VRPTW определяется как проблема минимизации времени и затрат в случае, когда ТС должны распределить товары со склада по множеству клиентов. VRPTW, минимизирует двухкритериальную, иерархическую целевую функцию: первая цель – минимизировать количество маршрутов (туров, агентов или ТС), а вторая – минимизировать общее время в пути. Решение с меньшим числом туров всегда предпочтительнее, даже если время в пути больше. Идея адаптации ACS к этим целям заключается в определении двух колоний ACS [1, 2], каждая из которых предназначена для оптимизации различных целевых функций. В MACS-VRPTW реализуется агентное управление, при котором колонии сотрудничают, обмениваясь информацией через обновление феромонов.

В литературе предложено множество алгоритмов ACO для решения различных типов задач комбинаторной оптимизации. В частности, было показано, что алгоритмы ACO очень эффективны в сочетании со специализированными процедурами локального поиска для решения симметричных и асимметричных задач коммивояжера [1, 2, 6]. Заметим, что MACS-VRPTW можно считать луч-

шим из существующих методов как по качеству решения, так и по времени вычислений. MACS-VRPTW улучшает решения, известные в литературе для некоторых экземпляров задач.

Для модели построения многоагентных маршрутов в сети потребителей, региональных баз и центральной базы использованы алгоритмы, обеспечивающие временные ограничения и оптимальность поставок в такой иерархической сети. На уровне региональных центров приводится более общий алгоритм, направленный на минимизацию количества агентов (ТС), так и на оптимальность маршрутов на кластерах с учетом временных окон.

Для решения многоагентной иерархической задачи HmTSPTW реализуется алгоритм MACS-VRPTW на основе оптимизации кооперации муравьиной колонии. Разработан алгоритм для решения задач маршрутизации транспортных средств с двумя целевыми функциями: минимизация количества агентов (или транспортных средств) и минимизация общего времени в пути, где минимизация количества агентов имеет приоритет над минимизацией времени в пути.

Решение многоцелевой задачи с помощью алгоритма оптимизации с несколькими муравьиными колониями является перспективным. Тестирование проводилось на наборах SOLOMON [5]. Полученные результаты подтверждают работоспособность и эффективность предложенных алгоритмов.

[1] *Dorigo M., Gambardella L. M.* (1997) Ant Colony System: A cooperative learning approach to the Traveling Salesman Problem // IEEE Tr. Evol. Comp. 1, 53-66.

[2] *Dorigo M., Gambardella L. M.* (1997) Ant Colonies for the Traveling Salesman Problem // BioSystems. 43(2): 73-81.

[3] *Германчук М. С., Козлова М. Г., Лукъяненко В. А.* (2013) Модели обобщенных задач коммивояжера в интеллектуализации поддержки принятия решений для геоинформационных систем // Географические и геоэкологические исследования в Украине и сопредельных территориях: сборник научных статей / под общ. ред. Б.А. Вахрушева. – Симферополь: ДИАЙПИ, 2013. – Т.1. – С. 413-415.

[4] *Germanchuk M. S., Lemtyuzhnikova D. V., Lukianenko V. A.* (2021) Metaheuristic Algorithms for Multiagent Routing Problems // Automation and Remote Control, 2021, Vol. 82, No. 10, pp. 1787–1801.

[5] Solomon benchmark. – URL: https://www.sintef.no/projectweb/top/vrptw/solomon-benchmark/

[6] *Stützle T.* (1998) Local Search Algorithms for Combinatorial Problems // Analysis, Improvements, and New Applications, PhD Thesis, Intellectics Group, Department of Computer Science, Darmstadt University of Technology, Germany.

Multiple hierarchical routing with time windows

Kozlova Margarita[1] kozlovamg@cfuv.ru
Lukianenko Vladimir[1] art-inf@yandex.ru
Makarov Oleg[1]⋆ fantom2.00@mail.ru
[1]Simferopol, V. I. Vernadsky Crimean Federal University

The problem of modeling real logistics systems arranged in a hierarchical manner is considered. One of the variants of such a task is proposed in [3]. Clusters of lower-level consumers are formed that meet the time window limits for each consumer and the cluster as a whole. On each such cluster, the route of the traveling salesman agent is built and the vertex closest to the central node is allocated, which is the vertex of the transshipment of goods from heavy-duty vehicles to small-load vehicles serving the consumer cluster. The transshipment peaks, in turn, are combined into higher-level traveling salesman routes, taking into account time windows for routes of this level. The software implementation is tested on known networks. The technique is applicable for the synthesis of the central distribution center and the system distribution centers of the lower level, as well as for calculating the required number of vehicles (agents).

A review of the sources shows the prospects of using clustering methods to reduce the dimension of the problem and the use of metaheuristics for TSP (Traveling Salesman Problem) [4] with time windows. In this paper, a hierarchical routing model with windows is investigated.

A two-level logistics network for cargo delivery is considered. It is assumed that there is a central warehouse or distribution center (DC), from which the cargo is delivered to regional distribution centers (RDC) – warehouses, transshipment points, from which, in turn, the goods are distributed to consumers. A variant of such a network structure is shown in the figure. The regional center (top-level base) redistributes cargo to lower-level bases (RDC) that serve one or more clusters of consumers. We will call such a Hierarchical multiple Traveling Salesman Problem (HmTSP).

The HmTSP task with time windows – time restrictions on the delivery of goods, we will denote HmTSPTW. Time windows (TW) may be associated with loading and unloading operations, transportation of perishable goods, etc. In many cases, the inclusion of time windows in the mathematical model is mandatory. Options for traversing the cluster are considered, taking into account the required number of agents (vehicles) and the needs of customers (consumers). An algorithm based on the Multiple Ant Colony System For Vehicle Routing Problems With Time Windows (MACS-VRPTW) ant colony cooperation system was chosen for numerical implementation of the solution of the Multi-Vehicle Routing Problem with Time Windows (MVRPTW). A close algorithm is based a Multi-swarm Cooperative Particle Swarm Optimizer (MCPSO), i.e. a Multi-swarm Cooperative Particle Swarm Optimizer for Vehicle Routing Problem with Time Windows (MCPSO-VRPTW). MACS-VRPTW

is based on the an Ant Colony System (ACS), and, more generally, on Ant Colony Optimization (ACO), a meta-heuristic approach based on the behavior of real ant colonies.

The problem of routing vehicles with VRPTW time windows is defined as the problem of minimizing time and costs in the case when the vehicle must distribute goods from the warehouse to a variety of customers. VRPTW minimizes a two–criteria, hierarchical objective function: the first goal is to minimize the number of routes (tours, agents or vehicles), and the second is to minimize the total travel time. A solution with fewer tours is always preferable, even if the travel time is longer. The idea of adapting ACS to these goals is to define two ACS colonies [1, 2], each of which is designed to optimize different objective functions. In MACS-VRPTW, agent-based management is implemented, in which colonies cooperate by exchanging information through pheromone updates.

Many ACO algorithms have been proposed in the literature to solve various types of combinatorial optimization problems. In particular, it was shown that ACO algorithms are very effective in combination with specialized local search procedures for solving symmetric and asymmetric traveling salesman problems [1, 2, 6]. Note that MACS-VRPTW can be considered the best of the existing methods both in terms of solution quality and calculation time. MACS-VRPTW improves solutions known in the literature for some instances of problems.

For the model of constructing multi-agent routes in a network of consumers, regional bases and a central base, algorithms were used to ensure time constraints and optimal supply in such a hierarchical network. At the level of regional centers, a more general algorithm is given, aimed at minimizing the number of agents (vehicles), and at optimality of routes on clusters, taking into account time windows.

To solve the multi-agent hierarchical HmTSPTW problem, the MACS-VRPTW algorithm is implemented based on the optimization of ant colony cooperation. An algorithm has been developed to solve vehicle routing problems with two objective functions: minimizing the number of agents (or vehicles) and minimizing the total travel time, where minimizing the number of agents takes precedence over minimizing travel time.

Solving a multi-purpose problem using an optimization algorithm with several ant colonies is promising. Testing was carried out on the sets of SOLOMON [5]. The results obtained confirm the efficiency and effectiveness of the proposed algorithms.

[1] *Dorigo M., Gambardella L. M.* (1997) Ant Colony System: A cooperative learning approach to the Traveling Salesman Problem // IEEE Tr. Evol. Comp. 1, 53-66.

[2] *Dorigo M., Gambardella L. M.* (1997) Ant Colonies for the Traveling Salesman Problem // BioSystems. 43(2): 73-81.

[3] *Germanchuk M. S., Kozlova M. G., Lukianenko V. A.* (2013) Models of generalized traveling salesman tasks in the intellectualization of decision support for Geoinformation systems // Geographical and geoecological research in Ukraine and adjacent

territories: collection of scientific articles / edited by B.A. Vakhrushev. – Simferopol: DIAYPI, 2013. – 1. – Pp. 413-415.

[4] *Germanchuk M. S., Lemtyuzhnikova D. V., Lukianenko V. A.* (2021) Metaheuristic Algorithms for Multiagent Routing Problems // Automation and Remote Control, 2021, Vol. 82, No. 10, pp. 1787–1801.

[5] Solomon benchmark. – URL: https://www.sintef.no/projectweb/top/vrptw/solomon-benchmark/

[6] *Stützle T.* (1998) Local Search Algorithms for Combinatorial Problems // Analysis, Improvements, and New Applications, PhD Thesis, Intellectics Group, Department of Computer Science, Darmstadt University of Technology, Germany.

Мажоритарное доминирование на графах с ограниченной степенью вершин

Лемтюжникова Дарья Владимировна[1] darabbt@gmail.com
Чеботарев Павел Юрьевич[2] pavel4e@gmail.com
Губко Михаил Владимирович[1] mgoubko@mail.ru
Кудинов Илья Дмитриевич[1]⋆ ilja@kdsli.ru
Шушко Никита Игоревич[1] shushko.ni@phystech.edu

[1]Москва, Институт проблем управления им. В.А. Трапезникова РАН
[2]Долгопрудный, Московский физико-технический институт

При голосовании о принятии важных для всего общества решений, когда решение принимается большинством голосов, социальные агенты (отдельные люди или небольшие сообщества) склонны голосовать, учитывая не только собственное мнение, но и мнение окружающих. Так, даже при собственном намерении голосовать «против», видя вокруг себя большинство намеренных проголосовать «за», агент порой голосует «за». Представляет интерес задача о минимальном числе намеренных голосовать «за», которого может быть достаточно, чтобы предложение было принято большинством голосов [1].

Пусть социальные связи в обществе описываются графом $G = (V, E)$, $|V| = n$, каждая вершина которого имеет петлю. Вершина $v \in V$ имеет множество соседей $N_v = \{w \in V \mid (v, w) \in E\}$, в которое входит и v. Степень вершины v – число $\deg(v) = |N_v|$. Каждая вершина $v \in V$ обладает *мнением* $f(v) \in \{-1, 1\}$.

Функция мнений обобщается для подмножеств $W \subseteq V$ как сумма мнений вершин: $f(W) = \sum_{v \in W} f(v)$. В ходе голосования вершина $v \in V$ ориентируется на мнение вершин из N_v, и голосует «за», если $f(N_v) \geqslant 1$, либо «против» в ином случае. Если число $|V^+| = |\{v \in V \mid f(N_v) \geqslant 1\}|$, голосующих «за», строго больше $|V|/2$, предложение, поставленное на голосование, принимается. *Числом доминирования* называется величина $\gamma_{\mathrm{maj}}(G) = \min\{f(V) \mid |V^+| > |V|/2\}$.

Рассматривается задача поиска такой функции мнений $f(v)$ графа G, которая реализует число доминирования $\gamma_{\mathrm{maj}}(G)$, то есть обеспечивает успех голосования при минимальном числе вершин v таких, что $f(v) = 1$. Эта задача называется *задачей доминирования большинством* [2]. Далее рассматривается задача нахождения графов G с определенными ограничениями на порядок и степени вершин, которые минимизируют $\gamma_{\mathrm{maj}}(G)$.

В работе исследованы классы графов заданного порядка n с петлями и ограничениями на степень вершин v вида $\deg(v) \leqslant k$ либо $k - 1 \leqslant \deg(v) \leqslant k$, где k фиксировано. Построен набор свойств, характеризующих каждый класс.

Получен аналитический вид зависимости минимального значения γ_{maj} от параметров n и k. В случае $\deg(v) \leqslant k$ это значение достигается на графе, изображённом на рис. 1, а в случае $k - 1 \leqslant \deg(v) \leqslant k$ оно достигается на k-регулярном графе (см. рис. 2).

Работа поддержана грантом РНФ No. 22-71-10131.

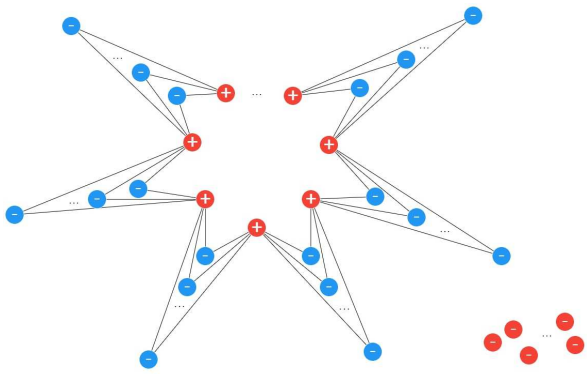


Рис. 1. Граф G порядка n с ограничением $\deg(v) \leqslant k$ для всех вершин v, на котором достигается минимальное значение γ_{maj} для заданных n и k. Для вершин v синего цвета $f(N_v) \geqslant 1$, для вершин красного цвета $f(N_v) < 1$. Вершины v, для которых $f(v) = -1$, обозначаются знаком «$-$», вершины w, для которых $f(w) = 1$, – знаком «$+$». Вне основной части графа присутствует набор вершин, соединённых произвольным образом.

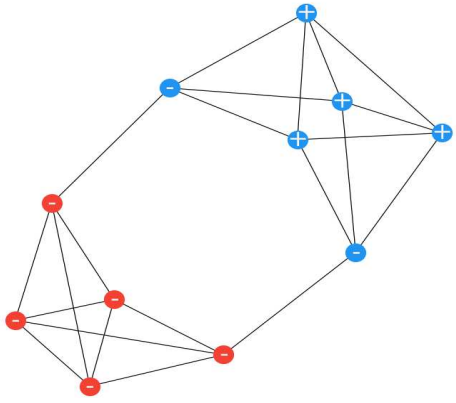


Рис. 2. k-ругялырный граф G порядка n с ограничением $k - 1 \leqslant \deg(v) \leqslant k$ для всех вершин v, на котором достигается минимальное значение γ_{maj} для заданных $n = 11$ и $k = 5$.

[1] *Chebotarev P., Peleg D.* The power of small coalitions under two-tier majority on regular graphs // CoRR, 2022.

[2] *Broere I., Hattingh J.H., Henning M.A., McRae A.A.* Majority domination in graphs // Discrete Mathematics, 1995. V. 138. No. 1–3. P. 125–135.

Majority domination problem for graphs with given maximum vertex degree

Lemtuzhnikova Darya[1] darabbt@gmail.com
Chebotarev Pavel[2] pavel4e@gmail.com
Goubko Mikhail[1] mgoubko@mail.ru
Kudinov Ilja[1]⋆ ilja@kdsli.ru
Shushko Nikita[1] shushko.ni@phystech.edu
[1]Moscow, Institute of Control Sciences of RAS
[2]Dolgoprudny, Moscow Institute of Physics and Technology

In voting on decisions important for the whole society, when the decision is made by a majority vote, social agents (individuals or small communities) tend to vote based not only on their own opinion but also on the opinion of those around them. Even with its own opinion to vote "against", seeing around it the majority those want to vote "for", the agent can as well vote "for". Of interest is the problem of finding the minimum number of intentional votes "for", which may be sufficient for the proposal to be accepted by a majority [1].

Let $G = (V, E)$, $|V| = n$ be a graph of the social relations in society. Each vertex has a loop. For any $v \in V$, $N_v = \{w \in V \mid (v, w) \in E\}$ is the set of neighbors of v, $v \in N_v$. The degree of v is the number $\deg(v) = |N_v|$. Each vertex $v \in V$ has a *private opinion* $f(v) \in \{-1, 1\}$.

For any $W \subseteq V$, let $f(W) = \sum_{v \in W} f(v)$. Vertex v makes its choice during the voting process not only based on its own opinion $f(v)$, but on the collective opinion of the vertices belonging to N_v. It is said that a vertex $v \in V$ is a *proponent* if $f(N_v) \geqslant 1$, and it is an *opponent* otherwise. If a number of proponents $|V^+| = |\{v \in V \mid f(N_v) \geqslant 1\}|$ exceeds $|V|/2$, then the proposal put to the vote is adopted. The *majority domination number* for a graph G is $\gamma_{\text{maj}}(G) = \min\{f(V) \mid |V^+| > |V|/2\}$.

We consider the problem of finding the opinion function $f(v)$, $v \in V$ for a graph $G = (V, E)$ that provides the minimum value of $\gamma_{\text{maj}}(G)$. The desired function f ensues the adopting of the proposal with the minimum number of vertices v such that $f(v) = 1$. This problem is known as *majority domination* [2]. Moreover, we solve the problem of finding the graphs G that obey certain restrictions on the order and vertex degree and minimize $\gamma_{\text{maj}}(G)$.

We study the classes of graphs of fixed order n with loops, for which $\deg(v) \leqslant k$ or $k - 1 \leqslant \deg(v) \leqslant k$, where k is a fixed parameter. The sets of properties characterizing these classes are provided.

We analytically obtain the function expressing the dependence of the minimum $\gamma_{\text{maj}}(G)$ on the parameters n and k. In the case of $\deg(v) \leqslant k$, this minimum is reached on the graph shown in Fig. 1. In the case of $k-1 \leqslant \deg(v) \leqslant k$, it is reached on the k-regular graph shown in Fig. 2.

This research is funded by RSF, grant 22-71-10131.

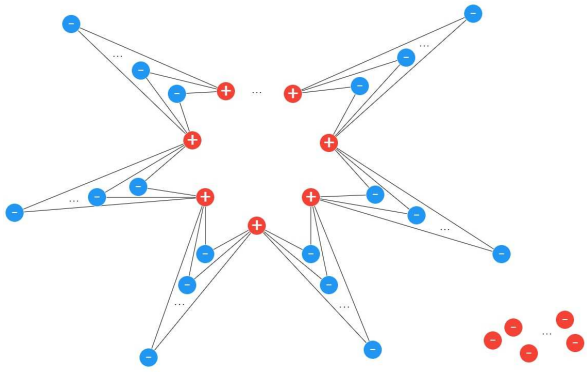


Fig. 1. An example of graph G of order n with restriction $\deg(v) \leqslant k$ for all vertices v, for which the minimum value of $\gamma_{\mathrm{maj}}(G)$ is reached for given parameters n and k. The blue vertices v are those with $f(N_v) \geqslant 1$; for the red vertices, $f(N_v) < 1$ holds. Vertices v with label "$-$" satisfy $f(v) = -1$; vertices w with label "$+$" satisfy $f(w) = 1$. There is a number of vertices outside the main part of the graph the connections between which are irrelevant.

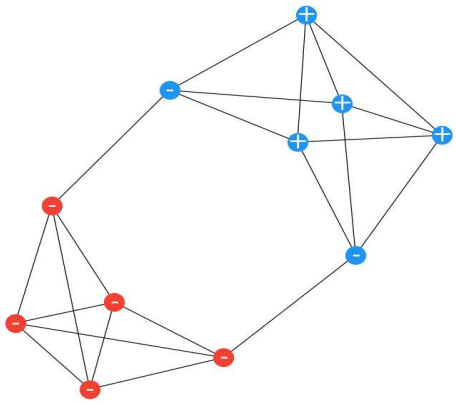


Fig. 2. An example of k-regular graph G of order n with restriction $k - 1 \leqslant \deg(v) \leqslant k$ for all vertices v, for which the minimum value of $\gamma_{\mathrm{maj}}(G)$ is reached for $n = 11$ and $k = 5$.

[1] *Chebotarev P., Peleg D.* The power of small coalitions under two-tier majority on regular graphs // CoRR, 2022.

[2] *Broere I., Hattingh J.H., Henning M.A., McRae A.A.* Majority domination in graphs // Discrete Mathematics, 1995. V. 138. No. 1–3. P. 125–135.

Проблематика исследования труднорешаемых задач

Лемтюжникова Дарья Владимировна[1] darabbt@gmail.com
Лукьяненко Владимир Андреевич[2]⋆ art-inf@yandex.ru
[1]Москва, Институт проблем управления им. В.А. Трапезникова РАН РФ
[2]Симферополь, ФГАОУ ВО «Крымский федеральный университет имени В. И. Вернадского»

Рассматривается проблематика исследования дискретных экстремальных задач, к которым относятся задачи дискретной оптимизации (ДО), комбинаторной оптимизации, псевдобулевой оптимизации или задачи выбора (некорректные задачи) и пр. Уточнение терминологии происходит на уровнях математической модели, задачи, метода, алгоритма в зависимости от специфики исходной прикладной задачи.

Основной является проблема алгоритмической разрешимости NP-трудных задач; выделение полиномиально разрешимых классов задач, содержащихся в класс NP-трудных (например, полиномиально разрешимых задач коммивояжера); разработка алгоритмов приближенного решения NP-трудных задач на основе их полиномиально разрешимых подклассов.

С каждым годом происходит все более тонкая классификация NP-сложных задач и пополняется список новых постановок NP-трудных задач, актуальных для приложений. Кроме выделения подкласса полиномиально разрешимых задач, важным является описание свойств и прецедентов заведомо трудно решаемых (неразрешимых). Распознавание NP-полных задач основывается на результатах С. Кука. Заметим, что экземпляры полиномиально разрешимых классов задач ДО требуют невообразимого времени для построения даже приближенного решения.

Можно поставить задачу описания NP-трудных задач, близких к полиномиально разрешимым и практически неразрешимым. Здесь возникает проблема выбора и описания близости. Введение метрик для близких задач и для соответствующих пространств решений существенно зависит от специфики самих задач, свойств множеств ограничений, целевых функций, геометрии соответствующих множеств [1, 4]. Применяемый комбинаторно-геометрический подход кроме наглядности связан с изучением комбинаторно-геометрических свойств NP-трудных задач и соответствующей интерпретацией алгоритмов решения [5]. В этом подходе рассматривается система «задача – алгоритм», которая исследовалась, начиная с работ Ю. И. Журавлева [6] в направлении получения оценок сложности задач и алгоритмов. Представляет интерес исследования классов близких задач в самых разных вариантах.

Подход, основанный на решении близких задач, опирается на систему «задача – близкая задача – алгоритм». В частности, для геометрического конструктивизма в [1] приводится ссылка на работу Й. Моравека, в которой «формализуется класс алгоритмов, основанных на линейных сравнениях, и предпринимается

попытка получения нижних оценок числа сравнений, необходимых для решения задачи». Здесь термины «сравнение», «близость» являются базовыми.

Для исследования труднорешаемой задачи, получения какой-либо информации о решении, может использоваться близкая простая (эталонная) задача, для которой имеется решение. Для этого исследуемую задачу включают в некоторое специальным образом построенное однопараметрическое семейство задач (гомотопирующее изучаемую задачу к эталонной), а затем это решение приближают по параметру к отыскиваемому решению исходной задачи. В работе [7] таким способом метод продолжения по параметру применен к исследованию различных классов экстремальных задач, в частности, к задачам математического программирования. Предполагается, что входящие в задачу отображения (функционалы, функции) определены в гильбертовых или банаховых пространствах и гладкие (дифференцируемые по Фреше, Гато или Липшицевы).

Параметризация экстремальных задач может быть основана на необходимых условиях экстремума, представленного в виде операторного уравнения $A(z) = 0$. Пусть уравнение удалось включить в однопараметрическое семейство уравнений $A(z, \lambda) = 0,\ 0 \leqslant \lambda \leqslant 1$, гладко зависящее от параметра λ. Причем $A(z, 0) = 0$ имеет решение z_0 и $A(z, 1) = A(z)$. Если $z(\lambda)$ продолжимо на промежуток $[0, 1]$, то $z(1)$ будет решением уравнения $A(z) = 0$. Гомотопический метод сводит решение уравнения $A(z) = 0$ к решению близкого (эталонного) уравнения $\tilde{A}(z) = A(z, 0)$. Наиболее простая связь между $\tilde{A}(z)$ и $A(z)$ определяется однопараметрическим семейством вида $\lambda A(z) + (1 - \lambda)\tilde{A}(z) = 0,\ \ 0 \leqslant \lambda \leqslant 1$, где эталонное уравнение $\tilde{A}(z)$ строится на основе имеющейся информации об исследуемом уравнении $A(z)$ (более простое по структуре; обеспечивающее близость решений z и $\tilde{z}$, $\tilde{A}\tilde{z} = 0$; оценку погрешности). С однопараметрическим семейством задач связан подход, применяемый в работах по теории расписаний [8].

Подход, основанный на близости задач и решений, применяется для разнообразных задач. Так для построения приближенных решений уравнений типа свертки используются две теоремы Ю. И. Черского о приближенном решении линейных уравнений [3].

Теоремы [3] и их обобщения с успехом применялись для приближенного решения краевых задач теории аналитических функций, систем линейных алгебраических уравнений, интегральных уравнений типа свертки, с том числе первого рода, уравнений типа Урысона и экстремальных задач, соответствующих некорректным труднорешаемым задачам уравнений первого рода [9].

Данная схема обеспечивает близость решений задач, моделей, структур. Здесь метрический подход базируется на метриках соответствующих нормированных (банаховые) пространств. Позволяет выделять классы задач, близких по решениям, входным данным в операторам преобразований. Возможны оценки погрешности решения задачи через решение близкой задачи на базе выбранной метрики.

Экстремальные задачи на графовых структурах большой размерности предполагают сравнение с прецедентами; построение вспомогательных структур, удовлетворяющих заданным свойствам и обеспечивающих полиномиальную разрешимость. В реальности возникают задачи на сложных структурах (качественных и количественных) для объектов и взаимодействующих агентов. Исследуемым реальным объектам ставятся в соответствие модели, задачи, наборы ограничений, предписаний, оценки экспертов и пр. Объекты, как правило, являются многопризнаковыми. Соответствующие признаки при формализации переносятся на модели и задачи. Тем самым, выделение требуемого класса задач связывается со сравнениями на близость по многим признакам (параметрам, векторам). Необходимая близость обеспечивается выбором соответствующих метрик. В этом случае подходящим математическим инструментом являются разработанные А. Б. Петровским [2] теоретические и практические положения мультимножеств. Существует выбор различных способов выделения метрик (псевдометрик) на σ-алгебрах измеримых множеств и мультимножеств. Для рассматриваемой проблематики применимы различные разновидности иерархического и неиерархического кластерного анализа, в частности, для задач классификации и упорядочения многопризнаковых объектов, которые могут существовать в нескольких вариантах с отличающимися значениями количественных и качественных признаков. Рассмотренные подходы могут быть перспективными для выделения классов полиномиально разрешимых задач, близких в NP-трудным; позволяют строить цепочки задач, алгоритмические процедуры приближенного решения и сложностные карты NP-трудных задач.

Работа поддержана грантом РНФ No. 22-71-10131.

[1] *Бондаренко В. А., Максименко А. Н.* Геометрические конструкции и сложность в комбинаторной оптимизации. — М.: Издательство ЛКИ, 2008. — 184 с.

[2] *Петровский А. Б.* Пространства множеств и мультимножеств. — М.: Едиториал УРСС, 2003. — 248 с.

[3] *Гахов Ф. Д., Черский Ю. И.* Уравнения типа свертки. — М.: Наука, 1978. — 296 с.

[4] *Деза М. М., Лоран М.* Геометрия разрезов и метрик. — М.: МЦНМО, 2001. — 736 с.

[5] *Гейл Д.* Соседние вершины на выпуклом многограннике // Линейные неравенства и смежные вопросы. — М.: ИЛ, 1959. — С. 355–362.

[6] *Журавлев Ю. И.* Избранные научные труды. — М.: Магистр, 1998. — 420 с.

[7] *Емельянов С. В., Коровин С. К., Бобылев Н. А., Булатов А. В.* Гомотопии экстремальных задач. — М.: Наука, 2001. — 350 с.

[8] *Lazarev A. A., Lemtyuzhnikova D. V., Werner F.* A metric approach for scheduling problems with minimizing the maximum penalty // Applied Mathematical Modelling. — 2021, V. 1902. — С. 1163–1176.

[9] *Belozub V., Kozlova M., Lukianenko V.* Approximated solution algorithms for Urysohn-type equations // Journal of Physics: Conference Series. — 2021, V. 89.

Issue of studying unsolvable problems

Lemtuzhnikova Dariana[1] darabbt@gmail.com
Lukianenko Vladimir[2] art-inf@yandex.ru
[1]Moscow, V.A. Trapeznikov Institute of Control Science of RAS
[2]Simferopol, V. I. Vernadsky Crimean Federal University

The problems of the study of discrete extremal problems, which include discrete optimization (DO) problems, combinatorial optimization, pseudo-Boolean optimization or selection problems (incorrect problems), etc., are considered. Terminology refinement takes place at the levels of the mathematical model, problem, method, algorithm, depending on the specifics of the original applied problem.

The main one is the problem of algorithmic solvability of NP-difficult problems; allocation of polynomial solvable classes of problems contained in the class of NP-difficult (for example, polynomial solvable traveling salesman problems); development of algorithms for approximate solutions of NP-difficult problems based on their polynomial solvable subclasses.

Every year there is an increasingly subtle classification of NP-complex problems and the list of new statements of NP-difficult problems relevant to applications is being updated. In addition to highlighting a subclass of polynomial solvable problems, it is important to describe the properties and precedents of notoriously difficult to solve (unsolvable). Recognition of NP-complete problems is based on the results of S. Cook. Note that instances of polynomial solvable classes of DO problems to require unimaginable time to construct even an approximate solution.

It is possible to set the task of describing NP-difficult problems that are close to polynomial solvable and practically unsolvable. Here there is a problem of choosing and describing proximity. The introduction of metrics for similar problems and for the corresponding solution spaces significantly depends on the specifics of the problems themselves, the properties of the sets of constraints, objective functions, and the geometry of the corresponding sets [1, 4]. The applied combinatorial-geometric approach, in addition to clarity, is associated with the study of the combinatorial-geometric properties of NP-difficult problems and the corresponding interpretation of the solution algorithms [5]. In this approach, the system "task – algorithm" is considered, which has been studied since the works of Yu. I. Zhuravlev [6] in the direction of obtaining estimates of the complexity of problems and algorithms. It is of interest to study classes of similar problems in a variety of variants.

The approach based on solving close problems is based on the system "task – close task – algorithm". In particular, for geometric constructivism in [1], a reference is given to the work of J. Moravek, in which "a class of algorithms based on linear comparisons is formalized and an attempt is made to obtain lower estimates of the number of comparisons necessary to solve the problem". Here the terms "comparison", "proximity" are basic.

To study a difficult-to-solve problem, to obtain any information about the solution, a close simple (reference) problem for which there is a solution can be used. To do this, the problem under study is included in a specially constructed one-parameter family of problems (homotoping the studied problem to the reference one), and then this solution is approximated by parameter to the solution of the original problem being sought. At work [7] in this way, the continuation method by parameter is applied to the study of various classes of extreme problems, in particular, to mathematical programming problems. It is assumed that the maps included in the problem (functionals, functions) are defined in Hilbert or Banach spaces and are smooth (differentiable by Frechet, Gato or Lipschitz).

Parametrization of extreme problems can be based on the necessary conditions of the extremum, represented as an operator equation $A(z) = 0$. Let the equation be included in a one-parameter family of equations $A(z, \lambda) = 0, 0 \leqslant \lambda \leqslant 1$, smoothly dependent on the parameter λ. Moreover, $A(z, 0) = 0$ has a solution of z_0 and $A(z, 1) = A(z)$. If $z(\lambda)$ is continued for the interval $[0, 1]$, then $z(1)$ will be the solution of equation $A(z) = 0$. The homotopy method reduces the solution of the equation $A(z)$ to solve the close (reference) equation $\tilde{A}(z) = A(z, 0)$. The simplest relationship between $\tilde{A}(z)$ and $A(z)$ is determined by a one-parameter family of the form $\lambda A(z) + (1 - \lambda)\tilde{A}(z) = 0, 0 \leqslant \lambda \leqslant 1$, where the reference equation $\tilde{A}(z)$ is constructed based on the available information about the equation under study $A(z)$ (simpler in structure; providing proximity of solutions z and $\tilde{z}$, $\tilde{A}\tilde{z} = 0$; error estimation). The approach used in the works on the theory of schedules is connected with the one-parameter family of problems [8].

The approach based on the proximity of tasks and solutions is applied to a variety of tasks. Thus, to construct approximate solutions of convolution-type equations, two theorems of Yu. I. Chersky on the approximate solution of linear equations are used [3].

Theorems [3] and their generalizations have been successfully applied to approximate solutions of boundary value problems of the theory of analytic functions, systems of linear algebraic equations, integral convolution-type equations, including the first kind, Urysohn-type equations and extreme problems corresponding to incorrect intractable problems of equations of the first kind [9].

This scheme ensures the proximity of solutions to problems, models, structures. Here the metric approach is based on the metrics of the corresponding normalized (Banach) spaces. Allows you to allocate classes of problems that are close in solutions, input data to transformation operators. It is possible to estimate the error of solving the problem by solving a close problem based on the selected metric.

Extreme problems on graph structures of large dimension involve comparison with precedents; construction of auxiliary structures satisfying the specified properties and providing polynomial solvability. In reality, problems arise on complex structures (qualitative and quantitative) for objects and interacting agents. Models, tasks, sets of restrictions, prescriptions, expert assessments, etc. are put in

accordance with the studied real objects. Objects, as a rule, are multisigned. The corresponding features are transferred to models and tasks during formalization. Thus, the allocation of the required class of tasks is associated with comparisons for proximity on many grounds (parameters, vectors). The necessary proximity is provided by the choice of appropriate metrics. In this case, a suitable mathematical tool is developed by A. B. Petrovsky [2] theoretical and practical provisions of multisets. There is a choice of different ways to allocate metrics (pseudometrics) on σ-algebras of measurable sets and multisets. Various types of hierarchical and non-hierarchical cluster analysis are applicable for the problems under consideration, in particular, for the tasks of classification and ordering of multi-sign objects, which can exist in several variants with different values of quantitative and qualitative features. The considered approaches can be promising for distinguishing classes of polynomial solvable problems that are close to NP-difficult; it allows you to build chains of problems, algorithmic procedures for approximate solutions and complexity maps of NP-difficult problems.

This research is funded by RSCF, grant 22-71-10131.

[1] *Bondarenko V. A., Maksimenko A. N.* Geometric constructions and complexity in combinatorial optimization. — Moscow: LKI Publishing House, 2008. — 184 p.

[2] *Petrovsky A. B.* Spaces of sets and multisets. — Moscow: Editorial URSS, 2003. — 248 p.

[3] *Gakhov F. D., Chersky Yu. I.* Convolution type equations. — Moscow: Nauka, 1978. — 296 p.

[4] *Deza M. M., Loran M.* Geometry of sections and metrics. — Moscow: MCNMO, 2001. — 736 p.

[5] *Gale D.* Adjacent vertices on a convex polyhedron // Linear inequalities and related issues. — Moscow: IL, 1959. — Pp. 355–362.

[6] *Zhuravlev Yu. I.* Selected scientific works. — Moscow: Magistr, 1998. — 420 p.

[7] *Emelianov S. V., Korovin S. K., Bobylev N. A., Bulatov A. V.* Homotopies of extreme problems. — Moscow: Nauka, 2001. — 350 p.

[8] *Lazarev A. A., Lemtyuzhnikova D. V., Werner F.* A metric approach for scheduling problems with minimizing the maximum penalty // Applied Mathematical Modelling. — 2021, V. 1902. — Pp. 1163–1176.

[9] *Belozub V., Kozlova M., Lukianenko V.* Approximated solution algorithms for Urysohn-type equations // Journal of Physics: Conference Series. — 2021, V. 89.

Многоэтапная стохастическая задача ориентирования: эвристические алгоритмы

Барашов Егор Борисович[1]⋆ barashov.eb@gmail.com
Лемтюжникова Дарья Владимировна[1,2] darabbt@gmail.com
Баттайа Ольга[3] olga.battaia@kedgebs.com
Садыков Руслан[4] ruslan.sadykov@inria.fr
[1]Москва, Институт Проблем Управления им. В.А. Трапезникова РАН
[2]Москва, Московский Авиационный Институт
[3]France, Kedge Business School
[4]France, INRIA Bordeaux

Рассматриваемая проблема ориентирования с временными окнами относится к классу проблем маршрутизации и планирования, возникающих при физическом распределении. Она рассматривает набор узлов (клиентов), каждый из которых имеет ассоциированную прибыль и продолжительность обслуживания (временное окно), и набор ребер, каждое из которых характеризуется определенным временем. Цель задачи - построить ациклический путь, начинающийся в заданном начале и заканчивающийся в заданном пункте назначения, который максимизирует общую прибыль при соблюдении ограничений на временное окно на всех узлах и не превышая заданного лимита времени. Задача классифицируется как NP-трудная, поэтому точный алгоритм, выполняющийся за разумное время, вряд ли существует.

Алгоритм должен быть реализован и протестирован на наборе случайно сгенерированных примеров. Рассматриваются следующие варианты генерации: примеры детерминированной задачи ориентирования с временными окнами и без них, с добавлением определенных сценариев, а также можно сгенерировать "туристические"примеры, основанные на реальных достопримечательностях некоторых городов.

Из-за значительной сложности задачи точный алгоритм будет работать только на относительно небольших примерах. Поэтому необходимо добавить некоторые эвристики для сравнения с точным решением на этих примерах, которые впоследствии можно использовать для решения больших задач. Можно использовать сложные релаксации, например, известную релаксацию NG-path. Такая релаксация может быть динамической: мы можем итеративно накладывать ограничения элементарности на основе расслабленного решения текущей релаксации. Этот подход, вероятно, будет более эффективным, чем базовый алгоритм маркировки.

Для сформулированной задачи необходимо ввести нетривиальную систему оценку пути. Вероятности перехода между вершинами являются дискретными величинами, поэтому предлагается использовать алгоритм муравьиной колонии. Он имеет достаточное количество преимуществ в виде параметризуемости, сходимости к субоптимальному решению за конечное время, а также

возможность внедрения локального поиска в этап алгоритма и на выходе не единственный маршрут, а целый набор, который соответствует ограничениям. Помимо этого подхода также можно воспользоваться комбинаторной оптимизацией, а корнкретно проблемой выбора эвристики перехода из i-го объекта в j-ый и проблемой оценки суммарной стоимости пройденного пути. Для их решения существуют различные рекомендательные системы оценки пути.

Задача состоит в определении многоэтапного решения, которое максимизирует ожидаемую прибыль. Решение принимает форму дерева решений, в котором на каждом этапе мы определяем следующую точку, в которую следует перейти в зависимости от реализации неопределенности в текущей точке.

Исследование выполнено при частичной финансовой поддержке РНФ в рамках научного проекта No.22-71-10131.

[1] *Bakker, H., Dunke, F., and Nickel, S.* A structuring review on multi- stage optimization under uncertainty: Aligning concepts from theory and practice. // Omega, 2020 – 96:102080.

[2] *Токарева М. М., Волкова Л. Л., Абдуллаев А. П. О.* О рекомендательной маршрутной системе, основанной на оценке предпочтений пользователя //Новые информационные технологии в автоматизированных системах. – 2016. – No.. 19. – С. 75-80.

[3] *Коцюба И. Ю., Назаренко А. Е.* Разработка рекомендательной системы для планирования туристических маршрутов в оптимизационной постановке // Моделирование, оптимизация и информационные технологии. – 2020. – Т. 8. – No.. 2. – С. 21-22.

The Multi-Stage Stochastic Orienteering Problem: heuristics

Barashov Egor[1]⋆ barashov.eb@gmail.com
Lemtuzhnikova Daria[1], 2 darabbt@gmail.com
Battaia Olga[3] olga.battaia@kedgebs.com
Sadykov Ruslan[4] ruslan.sadykov@inria.fr
[1]Moscow, V.A. Trapeznikov Institute of Control Science RAS
[2]Moscow, Moscow Aviation Institute
[2]France, Kedge Business School
[2]France, INRIA Bordeaux

The time-window orientation problem under consideration belongs to the class of routing and scheduling problems arising from physical allocation. It considers a set of nodes (clients), each with associated profit and processing time (time window), and a set of edges, each characterized by a particular time. The goal of the problem is to construct an acyclic path starting at a given origin and ending at a given destination that maximizes total profit while meeting the time window constraints on all nodes and not exceeding a given time limit. The problem is classified as NP-hard, so an exact algorithm that executes in a reasonable time is unlikely to exist.

The algorithm must be implemented and tested on a set of randomly generated examples. The following generation options are considered: examples of a deterministic orienteering problem with and without time windows, with the addition of certain scenarios, and it is also possible to generate "tourist" examples based on real landmarks of some cities.

Due to the considerable complexity of the problem, the exact algorithm will only work on relatively small examples. Therefore, it is necessary to add some heuristics to compare with the exact solution on these examples, which can then be used to solve larger problems. Complex relaxations, such as the well-known NG-path relaxation, can be used. Such a relaxation can be dynamic: we can iteratively impose elementarity constraints based on the relaxed solution of the current relaxation. This approach is likely to be more efficient than the basic labeling algorithm.

For the formulated problem, we need to introduce a non-trivial path estimation system. Transition probabilities between vertices are discrete quantities, so we propose to use the ant colony algorithm. It has sufficient advantages in the form of parameterizability, convergence to suboptimal solution in finite time, as well as the possibility of introducing local search in the algorithm step and the output is not a single route, but a whole set that meets the constraints. In addition to this approach, one can also take advantage of combinatorial optimization, and more specifically, the problem of choosing the heuristic of going from object i to object j and the problem of estimating the total cost of the path traveled. To solve them, there are various recommender path estimation systems.

The problem is to determine a multi-step solution that maximizes expected profits. The solution takes the form of a decision tree, in which at each stage we determine the next point to move to, depending on the realization of uncertainty at the current point.

The research was partially supported by RSF (project No.22-71-10131).

[1] *Bakker, H., Dunke, F., and Nickel, S.* A structuring review on multi- stage optimization under uncertainty: Aligning concepts from theory and practice. // Omega, 2020 – 96:102080.

[2] *Tokareva M. M., Volkova L. L., Abdullaev A. P. O.* Recommendation route system based on the evaluation of user preferences //New information technologies in automated systems. – 2016. – No.. 19. – P. 75-80.

[3] *Kotsyuba I. Yu., Nazarenko A. E.* Development of a recommendation system for planning tourist routes in the optimization statement // Modeling, optimization and information technology. – 2020. – Vol. 8. – No.. 2. – P. 21-22.

Содержание

Contents

Авторский указатель

Author index

MachineLearning.ru

http://www.machinelearning.ru/

Профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных. Цели ресурса — сконцентрировать информацию о достижения ведущих научных школ; способствовать обмену опытом, накоплению и распространению научных знаний; предоставить площадку для виртуальных научных семинаров и обсуждений.

Научное издание

Интеллектуализация обработки информации

Тезисы докладов
14-й Международной конференции

Подписано в печать 14.12.2022
Формат 60×84 1/8
Усл.-печ. л. 22,1. Уч.-изд. л. 23,2
Тираж 100 экз

Издатель — Российская Академия Наук

Печать — УНИД РАН

Отпечатано в экспериментальной цифровой типографии РАН

Издается по решению Научно-издательского совета
Российской академии наук (НИСО РАН) от 01.02.2022 г.
и распространяется бесплатно