Российская академия наук
Вычислительный центр РАН ФИЦ ИУ РАН
Центр хранения и анализа больших данных МГУ
Центр компетенций НТИ по направлению
«Искусственный интеллект» МФТИ

Математические методы распознавания образов

19-я Всероссийская конференция с международным участием

Москва, 2019

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

Математические методы распознавания образов: Тезисы докладов 19-й Всероссийской конференции с международным участием, г. Москва 2019 г. — М.: Российская академия наук, 2019. — 420 с.

ISBN 978-5-907036-76-5

В сборнике представлены тезисы докладов 19-й Всероссийской конференции «Математические методы распознавания образов», проводимой Российской академией наук, Вычислительным центром Федерального исследовательского центра «Информатика и управление» РАН, Центром компетенций НТИ по направлению «Искусственный интеллект» на базе Московского физико-технического института, Центром хранения и анализа больших данных на базе Московского государственного университета имени М. В. Ломоносова.

Конференция проводится регулярно, начиная с 1983 г., и является представительным научным форумом в области интеллектуального анализа данных, машинного обучения, распознавания образов, анализа изображений, обработки сигналов, дискретного анализа.

Сайт конференции `http://mmro.ru`.

ISBN 978-5-907036-76-5

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

Mathematical Methods for Pattern Recognition: Book of abstract of the 19th Russian National Conference with International Participation, Moscow, 2019. — Moscow: Russian Academy of Sciences, 2019. — 420 p.

ISBN 978-5-907036-76-5

The volume contains the abstracts of the 19th Russian National Conference "Mathematical Methods for Pattern Recognition". The conference is organized by the Russian Academy of Sciences, Federal Research Center "Computer Science and Control" of RAS, Center of big data storage and analysis technologie at the Moscow State University, and the competence Center of the National Technological Initiative "Artificial intelligence" at the Moscow Institute of Physics and Technology.
The conference has being held biennially since 1983. 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://mmro.ru/en/`.

ISBN 978-5-907036-76-5

Оргкомитет

Председатель: Журавлев Юрий Иванович, *акад. РАН*, *ФИЦ ИУ РАН*

Заместитель: Чехович Юрий Викторович, *к.ф.-м.н.*

Борисова Татьяна Игоревна
Горнов Александр Юрьевич, *д.т.н.*
Грабовой Андрей Валериевич
Громов Андрей Николаевич
Инякин Андрей Сергеевич, *к.ф.-м.н.*
Кокошкин Андрей Афанасьевич, *акад. РАН*
Мотренко Анастасия Петровна, *к.ф.-м.н.*
Помазкова Евгения Владимировна
Рейер Иван Александрович, *к.т.н.*
Соколов Игорь Анатольевич, *акад. РАН*
Татарчук Александр Игоревич, *к.ф.-м.н.*
Чехович Юлия Викторовна
Шананин Александр Алексеевич, *чл.-корр. РАН*

Программный комитет

Председатель: Рудаков Константин Владимирович, *акад. РАН*, *ФИЦ ИУ РАН*
Зорин Денис Николаевич, *проф.*, *CIMS NYU USA*

Ученый секретарь: Стрижов Вадим Викторович, *д.ф.-м.н.*

Воронцов Константин Вячеславович, *д.ф.-м.н.*
Гимади Эдуард Хайрутдинович, *д.ф.-м.н.*
Громова Ольга Алексеевна, *д.м.н.*
Двоенко Сергей Данилович, *д.ф.-м.н.*
Кельманов Александр Васильевич, *д.ф.-м.н.*
Краснопрошин Виктор Владимирович, *д.т.н.*
Матвеев Иван Алексеевич, *д.т.н.*
Местецкий Леонид Моисеевич, *д.т.н.*
Моттль Вадим Вячеславович, *д.т.н.*
Осипов Геннадий Семенович, *д.ф.-м.н.*
Пытьев Юрий Петрович, *д.ф.-м.н.*
Рязанов Владимир Васильевич, *д.ф.-м.н.*
Сойфер Виктор Александрович, *акад. РАН*
Чуличков Алексей Иванович, *д.ф.-м.н.*
Хачай Михаил Юрьевич, *д.ф.-м.н.*

Organizing Committee

Chair: Yury Zhuravlev, *acad. of RAS*, *FRCCSC*

Secretary: Yury Chekhovich, *C.Sc.*

Tatiana Borisova
Alexander Gornov, *D.Sc.*
Andrey Grabovoy
Andrey Gromov
Andrey Inyakin, *C.Sc.*
Andrey Kokoshkin, *acad. of RAS*
Anastasiya Motrenko, *C.Sc.*
Evgenia Pomazko
Ivan Reyer, *C.Sc.*
Igor Sokolov, *acad. of RAS*
Alexander Tatrshuk, *D.Sc.*
Yulia Chekhovich
Alexander Shananin, *corr. member of RAS*

Program Committee

Chair: Konstantin Rudakov, *acad. of RAS*, *FRCCSC*
Denis Zorin, *professor of computer CIMS NYU USA*

Secretary: Vadim Strijov, *D.Sc.*

Konstantin Vorontsov, *D.Sc.*
Edward Gimadi, *D.Sc.*
Olga Gromova, *D.Sc.*
Sergey Dvoenko, *D.Sc.*
Alexander Kel'manov, *D.Sc.*
Viktor Krasnoproshin *D.Sc.*
Ivan Matveev *D.Sc.*
Leonid Mestetskiy, *D.Sc.*
Vadim Mottl, *D.Sc.*
Genady Osipov, *D.Sc.*
Yury Pytiev, *D.Sc.*
Vladimir Ryazanov, *D.Sc.*
Viktor Soyfer, *acad. of RAS*
Alexey Chulichkov, *D.Sc.*
Michael Khachay, *D.Sc.*

Рецензенты

Адуенко А. А.
Анциперов В. Е.
Бахтеев О. Ю.
Бунакова В. Р.
Вальков А. С.
Ветров Д. П.
Визильтер Ю. В.
Владимирова М. Р.
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.
Дударенко М. А.
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.
Ишкина Ш. Х.
Карасиков М. Е.
Каркищенко А. Н.
Катруца А. М.
Копылов А. В.
Кочедыков Д. А.
Кочетов Ю. А.
Красоткина О. В.
Крымова Е. А.
Кудинов М. С.
Кузнецов М. П.
Кузнецова М. В.
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.
Матросов М. П.
Местецкий Л. М.
Миркин Б. Г.
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.
Хританков А. С.
Царьков С. В.
Черепанов Е. В.
Чичева М. А.
Чуличков А. И.
Янина А. О.

Reviewers

Aduenko A.
Antsiperov V.
Bakhteev O.
Bunakova V.
Cherepanov E.
Chicheva M.
Chulichkov A.
Dudarenko M.
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.
Ishkina Sh.
Ivakhnenko A.
Karasikov M.
Karkishchenko A.
Katrutsa A.
Khachay M.
Khritankov A.
Kochedykov D.
Kochetov Yu.
Kopylov A.
Krasotkina O.
Krymova E.
Kudinov M.
Kulunchakov A.
Kushnir O.
Kuz'min A.
Kuznetsov M.
Kuznetsova M.
Lange M.
Lomov N.
Lukashevich N.
Maksimov Yu.
Matrosov M.
Matveev I.
Maysuradze A.
Mestetskiy L.
Mikheeva A.
Mirkin B.
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.
Tsar'kov S.
Turdakov D.
Val'kov A.
Vetrov D.
Vizil'ter Yu.
Vladimirova M.
Volodin S.
Vorontsov K.
Yanina A.
Zagorujko N.
Zajtsev A.
Zharikov I.
Zhivotovskiy N.

Краткое оглавление

Brief contents

Поиск минимальных нечастых и максимальных частых наборов в частично упорядоченных данных

Драгунов Никита Аркадьевич[1]⋆ nikitadragunovjob@gmail.com
Дюкова Елена Всеволодовна[2] edjukova@mail.ru
[1]Москва, МГУ им. М.В. Ломоносова
[2]Москва, ВЦ ФИЦ ИУ РАН

Задача поиска минимальных нечастых (максимальных частых) наборов в данных занимает важное место в области информационного поиска. В случае бинарных данных эта задача ставится следующим образом.

Дано множество элементов V. Подмножества $X \subseteq V$ называются наборами. Пусть D — база данных, содержащая некоторые, не обязательно различные наборы. Наборы, содержащиеся в D, называются транзакциями. Под частотой набора $\nu(X)$ понимается доля транзакций в D, содержащих X. Если $\nu(X) \geqslant s$, где $s \in [0, 1]$, то набор X называется s-частым, иначе он называется s-нечастым. Если набор нечастый и при этом он не содержит в себе никакого другого нечастого набора, то такой набор называется минимальным нечастым. Если набор частый и он не содержится ни в каком другом частом наборе, то он называется максимальным частым. Требуется найти все минимальные нечастые (максимальные частые) наборы при заданном s. В более общей постановке каждый элемент имеет некоторое множество числовых значений, и вместо наборов элементов рассматриваются наборы их значений.

Важным приложением поиска частых и нечастых наборов является нахождение ассоциативных правил в базах данных. Поиск ассоциативных правил осуществляется в два этапа. На первом этапе происходит поиск частых наборов, на втором этапе из найденных частых наборов формируются ассоциативные правила. Задача поиска нечастых наборов фактически решается на втором этапе.

С ростом размерности современных баз данных искать все частые (нечастые) наборы становится неэффективно как по времени, так и по памяти в силу экспоненциального роста числа таких наборов. Одним из решений данной проблемы является поиск максимальных частых наборов $X_{\max}$ и минимальных нечастых наборов $Y_{\min}$, что позволяет компактно хранить информацию о всех частых и нечастых наборах соответственно. Наиболее изучены алгоритмы поиска $X_{\max}$ и $Y_{\min}$ в бинарных базах данных. Большинство из них работает по принципу, основанному на последовательном удалении отдельных элементов нечастых наборов и последовательном добавлении элементов частых наборов. Поэтому время работы таких алгоритмов существенно зависит от числа всех нечастых (частых) наборов [Dao-I Lin, Zvi M. Kedem, 1999].

В [1] рассмотрена задача поиска $X_{\max}$ и $Y_{\min}$ в частично упорядоченных данных.

Пусть $\mathcal{P} = \mathcal{P}_1 \times \ldots \times \mathcal{P}_n$ – декартово произведение частично упорядоченных множеств с порядком $\preceq$ и пусть $R \subseteq \mathcal{P}$. Тогда $R^- = \{x \in \mathcal{P} | \exists a \in R, x \preceq a\}$. Мно-

жество $I(R^-)$, состоящее из всех минимальных элементов множества $\mathcal{P} \setminus R^-$, называется минимальным независимым от R. Аналогично определяется максимальное независимое от R множество $I(R^+)$. Задачи поиска $I(R^-)$ и $I(R^+)$ являются одними из центральных труднорешаемых перечислительных задач дискретной математики. Каждая из этих задач называется задачей дуализации над произведением частично упорядоченных множеств. Произвольная совокупность наборов из $\mathcal{P}$ называется частично упорядоченной базой данных и обозначается $\mathcal{D}(\mathcal{P})$.

Поиск $X_{\max}$ и $Y_{\min}$ при заданной $\mathcal{D}(\mathcal{P})$ путем последовательного построения множеств $X_{\max}$ и $Y_{\min}$ является достаточно очевидным. Первое множество строится алгоритмом Apriori, а второе — путём дуализации первого. При последовательном построении $X_{\max}$ и $Y_{\min}$ используется свойство двойственности: $I(X^-_{\max}) = Y_{\min}$, $I(Y^+_{\min}) = X_{\max}$.

Предложенный в [1] подход к поиску $X_{\max}$ и $Y_{\min}$ основан на «совместном» перечислении этих множеств и автором экспериментально не изучен. Метод работает итеративно. В результате строятся две последовательности: $X_1 \subset X_2 \subset \cdots \subset X_{\max}$, $Y_1 \subset Y_2 \subset \cdots \subset Y_{\min}$. На первом шаге $X_1 = \{x\}$, $Y_1 = \{y\}$, где x и y ищутся алгоритмом Apriori. На $i+1$ ($i \geqslant 1$) шаге строится либо $I(X_i^-)$, либо $I(Y_i^+)$ и формируются X_{i+1}, Y_{i+1}. Таким образом, происходит дуализация всё больших по мощности множеств, вследствие чего, как показано в настоящей работе, метод не применим на практике для задач большой размерности.

Основным результатом представляемой работы является разработка и обоснование нового подхода к поиску $X_{\max}$ и $Y_{\min}$ в случае частично упорядоченных данных. Предложенный метод является синтезом последовательного и «совместного» подходов, описанных выше, и работает итеративно. Положим $X_0 = \varnothing$. Строится одна последовательность $X_1 \subset X_2 \subset \cdots \subset X_{\max}$. На первом шаге $X_1 = \{x\}$, где x ищется алгоритмом Apriori. На $i+1$ ($i \geqslant 1$) шаге решается задача дуализации множества $X_i \setminus X_{i-1}$ и формируется X_{i+1}. Множество $Y_{\min} = I(X^-_{\max})$ получается путём дуализации $X_{\max}$. Корректность метода базируется на приведённых ниже утверждениях 1 и 2.

Утверждение 1. *Пусть $X \subset X_{max}$. Тогда $I(X^-)$ содержит частые и нечастые наборы. При этом, если $y \in I(X^-)$ — нечастый набор, то y — минимальный нечастый набор.*

Утверждение 2. *Если $X \subseteq X_{max}$, $Y \subseteq Y_{min}$ и $I(X^-) = Y$, то $X = X_{max}$ и $Y = Y_{min}$.*

Проведено экспериментальное сравнение трёх описанных выше методов в случае, когда $\mathcal{P}$ — произведение цепей, и выявлены условия их эффективности. Экспериментально показано, что эффективность рассматриваемых подходов зависит как от числа всех максимальных частых и минимальных нечастых наборов, так и от соотношения между числом частых наборов и числом нечастых наборов. Если число частых наборов примерно равно числу нечастых наборов,

то наиболее эффективным является последовательно-совместный подход, предложенный в настоящей работе. Иначе наиболее эффективным является метод последовательного поиска $X_{\max}$ и $Y_{\min}$.

Работа частично финансирована РФФИ (проект № 19-01-00430-а).

[1] *Elbassioni K. M.* On Finding Minimal Infrequent Elements in Multi-dimensional Data Defined Over Partially Ordered Sets // arXiv,

Finding Minimal Infrequent and Maximal Frequent Sets in Partially Ordered Data

Nikita Dragunov[1]⋆ nikitadragunovjob@gmail.com
Elena Djukova[2] edjukova@mail.ru
[1]Moscow, Lomonosov MSU
[2]Moscow, CC FRC CSC RAS

Finding minimal infrequent (maximal frequent) sets in data is an important task of information retrieval. In binary case this problem is formulated as follows.

Given an arbitrary attributes set V. Subsets $X \subseteq V$ are called sets of attributes. Given database D containing some sets of attributes, which are not necessarily different. Sets of attributes contained in D are called transactions. Set frequency $\nu(X)$ is the ratio number of transactions containing X to the number of all transactions. If $\nu(X) \geqslant s$, $s \in [0,1]$, then set X is called s-frequent, else it is called s-infrequent. If set is frequent and it does not contain another frequent subset then it is called minimal infrequent. Similarly, you can define maximal frequent set of attributes. It is required to find all minimal infrequent (maximal frequent) sets of attributes with given s. In a more general statement, each attribute has a set of numerical values, and sets of values are considered instead of the sets of attributes.

An important application of finding frequent and infrequent sets of attributes is an association rules mining. Association rules mining consists of two steps. First, frequent sets are searched, then association rules are constructed from frequent sets. During the association rules constructing the problem of infrequent sets searching is solved.

With the growth of the dimension of modern databases, it becomes inefficient to search for all frequent (infrequent) sets both in time and in memory due to the exponential growth of the number of such sets. One solution to this problem is to find set of all maximal frequent subsets $X_{\max}$ and set of all minimal infrequent subsets $Y_{\min}$, that allow to store information about all frequent and infrequent subsets compactly. Typically, the search of $X_{\max}$ and $Y_{\min}$ is reduced to the binary case. Most of the algorithms works sequentially decreasing infrequent sets and sequentially decreasing frequent sets. Thus the running time of such algorithms sufficiently depends on number of all infrequent (frequent) sets [Dao-I Lin, Zvi M. Kedem, 1999].

In [1] the concept of non-binary partially ordered databases was introduced.

Assume $\mathcal{P} = \mathcal{P}_1 \times \ldots \times \mathcal{P}_n$ – Cartesian product of partially ordered sets with binary relation $\preceq$, $R \subseteq \mathcal{P}$. $R^- = \{x \in \mathcal{P} | \exists a \in R, x \preceq a\}$. Set $I(R^-)$ containing all minimal elements of set $\mathcal{P} \backslash R^-$ is called minimal independent of R. Similarly, you can define maximal independent of R set $I(R^+)$. Problems of mining $I(R^-)$ and $I(R^+)$ are computationally hard discrete problems called dualization. Arbitrary collection of $\mathcal{P}$ subsets is called partially ordered database and denotes $\mathcal{D}(\mathcal{P})$.

Sequential finding of $X_{\max}$ and $Y_{\min}$ with given $\mathcal{D}(\mathcal{P})$ is quite obvious. The first set is enumerated by Apriori algorithm, then the second one is enumerated

by dualization. Duality property is used during sequential finding: $I(X^-_{\max}) = Y_{\min}$, $I(Y^+_{\min}) = X_{\max}$

In [1] a method of finding $X_{\max}$ and $Y_{\min}$ based on joint sets enumeration was proposed, but was not studied experimentally. The method works iterative. Two sequences are constructed: $X_1 \subset X_2 \subset \cdots \subset X_{\max}$, $Y_1 \subset Y_2 \subset \cdots \subset Y_{\min}$. In the first step $X_1 = \{x\}$, $Y_1 = \{y\}$, where x and y are found by Apriori algorithm. In $i+1$ ($i \geqslant 1$) step $I(X_i^-)$ or $I(Y_i^+)$ is constructed and sets X_{i+1}, Y_{i+1} are found. Thus, on every step dualization of sets with increasing cardinality is solved, thus the method can not be used on big tasks in practise.

The main result of this work is the development and rationale for a new approach to finding $X_{\max}$ and $Y_{\min}$ in case of partially ordered data. Proposed method is a synthesis of sequential and joint approaches and works iterative. Assume $X_0 = \varnothing$. A sequence $X_1 \subset X_2 \subset \cdots \subset X_{\max}$ constructing. In the first step $X_1 = \{x\}$, where x is found by Apriori. In the $i+1$ ($i \geqslant 1$) step the dualization of $X_i \setminus X_{i-1}$ is solved and X_{i+1} is enumerated. Set $Y_{\min} = I(X^-_{\max})$ is constructed by dualization of set $X_{\max}$. The method is based on statements 1 and 2.

Statement 1. *Assume $X \subset X_{max}$. Then $I(X^-)$ consists of frequent and infrequent sets. If $y \in I(X^-)$ — infrequent set, then y — minimal infrequent set.*

Statement 2. *If $X \subseteq X_{max}$, $Y \subseteq Y_{min}$ and $I(X^-) = Y$, then $X = X_{max}$ and $Y = Y_{min}$.*

The three methods described above are experimentally compared in the case where $\mathcal{P}$ is the product of chains, and the conditions for their efficiency are revealed. It is shown experimentally that the efficiency of the considered approaches depends both on the number of all maximal frequent and minimal infrequent sets and on the ratio between the number of frequent sets and the number of infrequent sets. If the number of frequent sets is approximately equal to the number of infrequent sets, then the sequential-joint approach is the most efficient. If the number of frequent sets is substantially greater (less) than the number of infrequent sets, then sequential search of $X_{\max}$ and $Y_{\min}$ is the most efficient.

This research is partially financial supported by RFBR, grant 19-01-00430-a.

[1] *Elbassioni K. M.* On Finding Minimal Infrequent Elements in Multi-dimensional Data Defined Over Partially Ordered Sets // arXiv,

О поиске ассоциативных правил в небинарных данных

Генрихов Игорь Евгеньевич[1]⋆ ingvar1485@rambler.ru
Дюкова Елена Всеволодовна[2] edjukova@mail.ru
[1]Химки, ООО «Мобайл парк ИТ»
[2]Москва, ВЦ ФИЦ ИУ РАН

Задача поиска ассоциативных правил в данных является одной из центральных задач интеллектуального анализа информации и актуальна для многих прикладных областей. Эта задача впервые поставлена в 1993 г. Р. Агравалом, Т. Имелинским и А. Свами. Приведём её стандартную постановку в случае бинарных данных.

Дано некоторое множество P, элементы которого называются атрибутами. Дана база данных D, содержащая некоторые подмножества множества P, не обязательно различные. Подмножества множества P называются наборами атрибутов, а те из них, которые содержатся в D, называются транзакциями. *Ассоциативное правило* это пара непересекающихся наборов атрибутов X и Y, которые одновременно содержатся минимум в одной транзакции. Ассоциативное правило, порождаемое X и Y, обычно обозначается через $X \Rightarrow Y$. *Поддержкой* (*support*) правила $X \Rightarrow Y$ называется отношение числа транзакций, содержащих $X \cup Y$, к числу всех транзакций. *Достоверностью* (*confidence*) правила $X \Rightarrow Y$ называется отношение числа транзакций, содержащих $X \cup Y$, к числу транзакций, содержащих X. Требуется найти ассоциативные правила с поддержкой не менее s, $s \in [0, 1]$, и с достоверностью не менее c, $c \in [0, 1]$.

Поиск ассоциативных правил обычно осуществляется в два этапа. Сначала находятся все так называемые s-частые наборы атрибутов. Набор атрибутов Z называется s-частым, если отношение числа транзакций, содержащих Z, к числу всех транзакций не менее s (в противном случае Z называется s-нечастым). Затем для каждого найденного s-частого набора Z путем разбиения Z на два непересекающихся подмножества X и Y строятся ассоциативные правила вида $X \Rightarrow Y$ с достоверностью не менее c.

В более общей постановке каждый атрибут имеет некоторое множество числовых значений и вместо наборов атрибутов рассматриваются наборы их значений. Как правило, поиск ассоциативных правил сводится к бинарному случаю путем задания для каждого атрибута некоторого числа (порога), позволяющего перекодировать исходные небинарные данные в бинарные. Результат существенно зависит от выбора варианта бинаризации. Однако перебор по всем возможным вариантам бинаризации данных требует больших временных затрат.

На практике часто возникают задачи поиска зависимостей в частично упорядоченных данных. В [1] введены понятия s-частого и s-нечастого элемента для множества $P = P_1 \times \ldots \times P_n$, где $P_1, \ldots, P_n$ — конечные частично упорядоченные множества и элемент $y = (y_1, \ldots, y_n) \in P$ следует за элементом $x = (x_1, \ldots, x_n) \in P$ ($x \preceq y$), если y_i следует за x_i при $i = 1, 2, \ldots, n$. С целью

компактного хранения ассоциативных правил дано понятие неприводимого ассоциативного правила, опирающееся на понятия «минимального» s-нечастого элемента и «максимального» s-частого элемента множества P. Показано, что поиск таких правил может быть осуществлён путём решения вычислительно сложной дискретной задачи, называемой дуализацией над произведением частичных порядков. В случае, когда $P_i = \{0, 1\}$ $(0 \preceq 1,\ 0 \neq 1)$ для всех $i = 1, 2, \ldots, n$, при поиске неприводимых ассоциативных правил решается задача дуализации монотонной конъюнктивной нормальной формы (в матричной формулировке это задача построения неприводимых покрытий булевой матрицы).

В настоящей работе рассмотрен общий случай, а именно, когда P — произведение конечных частичных порядков и для этого случая дано понятие ассоциативного правила, приведённое ниже.

Элементы $x, y \in P$ называются *сравнимыми*, если либо $x \preceq y$, либо $y \preceq x$. В противном случае x и y называются *несравнимыми*.

Предполагается, что каждое множество P_i имеет *наименьший элемент*, т.е. такой элемент l_i, для которого выполнено $l_i \preceq x_i$ для любого $x_i \in P_i$. Элемент $x_i \in P_i$ называется существенным значением элемента $x = (x_1, \ldots, x_i, \ldots, x_n) \in P$, если $x_i \neq l_i$. Предполагается также, что база данных D, $D \subseteq P$, не содержит транзакцию $l = (l_1, \ldots, l_n)$. Через $S_D(x)$, $x \in P$, обозначается число транзакций z в D таких, что $x \preceq z$.

Два несравнимых элемента $x = (x_1, \ldots, x_n)$ и $y = (y_1, \ldots, y_n)$ множества P называются *непересекающимися*, если для любого $i \in 1, 2, \ldots, n$ хотя бы один из элементов x_i и y_i равен l_i. Из определения следует, что если x и y — непересекающиеся элементы множества P, то $x \neq l$, $y \neq l$.

Пусть $x, y \in P$, $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$ — непересекающиеся элементы. Через $x \odot y = (u_1, \ldots, u_n)$ обозначается элемент множества P, в котором $u_i = l_i$, если $x_i = y_i = l_i$, иначе $u_i = x_i$, если x_i — существенное значение для x, и $u_i = y_i$, если y_i — существенное значение для y.

Ассоциативным (s, c)-правилом, $s \in [0, 1]$, $c \in [0, 1]$, называется пара непересекающихся элементов x и y множества P таких, что $\frac{S_D(x \odot y)}{|D|} \geqslant s$, $S_D(x \odot y)/S_D(x) \geqslant c$. Ассоциативное (s, c)-правило, порождаемое элементами $x, y \in P$, называется *неприводимым*, если $\frac{S_D(x)}{|D|} < s$, $\frac{S_D(z)}{|D|} \geqslant s$ для любого z, такого что $z \preceq x$, $x \neq z$, и $\frac{S_D(z)}{|D|} < s$ для любого z, такого что $x \odot y \preceq z$, $x \odot y \neq z$ (т. е. ассоциативное (s, c)-правило $x \Rightarrow y$ является неприводимым, если x — минимальный s-нечастый элемент в P, а $x \odot y$ — максимальный s-частый элемент в P).

Ассоциативное (s, c)-правило $x \Rightarrow y$ указывает на определённую зависимость между набором существенных значений элемента x и набором существенных значений элемента y и является обобщением понятия ассоциативного правила, введённого выше для бинарных данных. Приведены иллюстративные модельные примеры. Кроме того, для поиска s-частых элементов в небинарных

данных, в том числе в частично упорядоченных, предложено модифицировать конструкцию классического бинарного FP-дерева (Frequent Pattern Tree) путём введения для каждого не бинарного атрибута дополнительной вершины, содержащей информацию о возможных вариантах бинаризации значений этого атрибута.

Работа частично финансирована РФФИ (проект № 19-01-00430-а).

[1] *Elbassioni K. M.* On Finding Minimal Infrequent Elements in Multi-dimensional Data Defined over Partially Ordered Sets // arXiv:1411.2275. — 2014. — 30 p.

On the search of association rules in nonbinary data

Genrikhov Igor Evgenyevich[1]⋆ ingvar1485@rambler.ru
Djukova Elena Vsevolodovna[2] edjukova@mail.ru
[1]LLC ≪Mobile Park IT≫, Khimki, Russia
[2]CC FRC CSC RAS, Moscow, Russia

Finding association rules in data is a central problem in pattern recognition data analysis, and it is important for many applications. This problem was first stated in 1993 by R. Agrawal, T. Imielinski, and A. Swami. Its standard statement for binary data is as follows.

A set P, the elements of which are called attributes, is given. There is also a database D containing subsets of P, which are not necessarily different. The subsets of P are called sets of attributes, and the subsets contained in D are called transactions. An association rule is a pair of nonintersecting sets of attributes X and Y that simultaneously belong to one and the same transaction. The association rule generated by X and Y are usually denoted by $X \Rightarrow Y$. The support of the rule $X \Rightarrow Y$ is the ratio of the number of transactions containing $X \cup Y$ to the number of all transactions. The confidence of the rule $X \Rightarrow Y$ is the ratio of the number of transactions containing $X \cup Y$ to the number of transactions containing X. It is required to find the association rules with the support not less than s, $s \in [0, 1]$, and with the confidence not less than c, $c \in [0, 1]$.

Association rules are usually sought in two steps. First, all so-called s-frequent sets of attributes are found. A set of attributes Z is said to be s-frequent if the ratio of the number of transactions containing Z to the number of all transactions is not less than s (otherwise, Z is called s-infrequent). Next, for each found s-frequent set Z, association rules $X \Rightarrow Y$ with confidence not lower than c are found by decomposing Z into two nonintersecting subsets X and Y.

In a more general statement, each attribute has a set of numerical values, and sets of values are considered instead of the sets of attributes. Typically, the search for association rules is reduced to the binary case by specifying a number (threshold) for each attribute that makes it possible to represent nonbinary data as binary data. The result significantly depends on the choice of binarization method. However, the search through all data binarization variants is computationally costly.

In practice, one often encounters problem of finding dependencies in partially ordered data. In [1] introduced the concepts of s-frequent and s-infrequent element for the set $P = P_1 \times \ldots \times P_n$, where $P_1, \ldots, P_n$ are finite partially ordered sets and the element $y = (y_1, \ldots, y_n) \in P$ succeeds the element $x = (x_1, \ldots, x_n) \in P$ if y_i succeeds x_i for $i = 1, 2, \ldots, n$. For the purpose of compact storage of association rules, the concept of irreducible association rule was introduced, based on the concepts of the "minimum" s-infrequent element and the "maximum" s-frequent element of the set P. It was shown that the search for such rules can be accomplished by solving a computationally complex discrete problem called dualization over the product of

partial orders. If $P_i = 0, 1$ ($0 \preceq 1$, $0 \neq 1$) for all $i = 1, 2, \ldots, n$, then the problem of dualization of a monotone conjunctive normal form is solved in the searching for irreducible association rules (in the matrix formulation, this is the problem of constructing irreducible covers of a Boolean matrix).

In this paper the general case in which P is the product of finite partial orders was considered, and the concept of association rule was defined for this case. This concept is a generalization of the concept of the association rule introduced above for binary data. Illustrative model examples are discussed. In addition, for finding s-frequent elements in nonbinary data, including partially ordered data, it is proposed to modify the construction of the classical binary FP-tree (Frequent Pattern Tree) by introducing for each not binary attribute additional vertex containing information about possible variants of values binarization for this attribute.

This research is partial financial supported of RFBR, grant 19-01-00430-a.

[1] *Elbassioni K. M.* On Finding Minimal Infrequent Elements in Multi-dimensional Data Defined over Partially Ordered Sets // arXiv:1411.2275. "— 2014. "— 30 p.

Субъективное восстановление пропусков данных измерений объекта исследования и его математической модели

Пытьев Юрий Петрович[1] yuri.pytyev@gmail.com
Фаломкина Олеся Владимировна[1] olesya.falomkina@gmail.com
Чуличков Алексей Иванович[1]⋆ achulichkov@gmail.com
[1]Москва, МГУ им. М. В. Ломоносова

В докладе даны решения задач эмпирического восстановления субъективной математической модели объекта исследования (ОБИ) и субъективной интерпретации данных его измерений, искажённых шумом, математическая модель которого неизвестна, и «пропусками» данных измерений ОБИ, причём математическая модель измерений также неизвестна. Такая неопределённая и нечёткая априорная информация о рассмотренных задачах трудно формализуема и требует специального математического аппарата как для математической формулировки их постановки, так и для математических методов их решения.

Для постановки и для решения названных задач использован математический формализм субъективного моделирования (МФСМ) [1], [2], позволяющий математически сформулировать как субъективную модель ОБИ, так и субъективную модель его измерений и их субъективной интерпретации с учётом восстановленных данных измерений в «пропусках». Для этого использованы субъективные представления модельера-исследователя (м-и) о физических свойствах ОБИ и о средствах его измерений, о математических свойствах шума и т. п.; вся подобная субъективная информация основана на научном опыте м-и и на его интуиции учёного.

МФСМ создан для моделирования неопределённости, отражающей неполноту и недостоверность субъективной информации об ОБИ, нечёткости и случайности, характерной для её содержания, характеризующего математическую модель ОБИ. Математическую модель субъективных суждений о значениях неизвестного параметра $x \in X$ модели $M(x)$ ОБИ модельер-исследователь (м-и) задает как пространство $(X, \mathcal{P}(X), \mathrm{Pl}^{\widetilde{x}}, \mathrm{Bel}^{\widetilde{x}})$ с мерами *правдоподобия* $\mathrm{Pl}^{\widetilde{x}}$ и *доверия* $\mathrm{Bel}^{\widetilde{x}}$, где $\widetilde{x}$ – неопределенный элемент (ноэ) со значениями в X, моделирующий субъективные суждения м-и о неизвестном $x \in X$, меры $\mathrm{Pl}^{\widetilde{x}}$ и $\mathrm{Bel}^{\widetilde{x}}$ моделируют модальности его субъективных суждений об истинности каждого $x \in X$: значение $\mathrm{Pl}^{\widetilde{x}}(\widetilde{x} = x)$ определяет, насколько, по его мнению, относительно правдоподобно равенство $\widetilde{x} = x$, а значение $\mathrm{Bel}^{\widetilde{x}}(\widetilde{x} \neq x)$ определяет, насколько следует относительно доверять неравенству $\widetilde{x} \neq x$, где «относительно» означает, что значения $\mathrm{Pl}(\cdot)$ и $\mathrm{Bel}(\cdot)$ не важны, имеет смысл лишь их упорядоченность.

МФСМ, в отличие от «стандартного» математического моделирования, позволяет м-и математически моделировать как точные формализованные знания модели ОБИ, так и неформализованные, неполные и недостоверные, начиная с «абсолютного незнания» вплоть до «точного знания» модели ОБИ, вычис-

лять относительные правдоподобия и доверия истинности любых характеристик ОБИ, обусловленных его субъективной моделью $M(\widetilde{x})$.

Субъективная модель измерений ОБИ и интерпретации данных его измерений определены как элементы параметрического семейства сглаживающих сплайнов [3], [4], представляющих данные измерений в виде суммы двух слагаемых: гладкого, оптимально оценивающего данные измерений, «очищенные» от шума, и слагаемого, оптимально оценивающего шум, причем «оптимальность» м-и определяет субъективно, изменяя параметры сплайна. В докладе исследованы зависимости качества решений перечисленных задач от количества «пропусков» данных измерений и относительного числа «пропущенных» данных измерений.

Работа поддержана грантами РФФИ № 17-07-00832 А, 18-07-00424 А.

[1] *Pyt'ev Y. P.* Modeling of subjective judgments made by a researcher-modeler about the model of the research object // Mathematical Models and Computer Simulations. — 2013. — Vol. 5, no. 6. — P. 538–557.

[2] *Ю. П. Пытьев.* Вероятность, возможность и субъективное моделирование в научных исследованиях. Математические и эмпирические основы, приложения. Москва. ФИЗМАТЛИТ, 2018, 268 стр.

[3] *Ю. П. Пытьев, О. В. Фаломкина, С. А. Шишкин, А. И. Чуличков.* Математический формализм субъективного моделирования //Машинное обучение и анализ данных, Москва, 2018. — Т. 4, № 2. — С. 108–121.

[4] *Yu. P. Pyt'ev, O. V. Falomkina, S. A. Shishkin.* Subjective Restoration of Mathematical Models for a Research Object, Its Measurements, and Measurement-Data Interpretation // Pattern Recognition and Image Analysis, № 4, 2019 (принята к печати).

Subjective restoration of missing measurement data of the research object and its mathematical model

Pyt'ev Yuri Petrovich[1] yuri.pytyev@gmail.com
Falomkina Olesya Vladimirovna[1] olesya.falomkina@gmail.com
Chulichkov Alexey Ivanovich[1]⋆ achulichkov@gmail.com
[1]Moscow, Lomonosov Moscow State University

The report provides solutions to the problems of empirical reconstruction of the subjective mathematical model of a research object (RO) and the subjective interpretation of its measurement data distorted by noise, the mathematical model of which is unknown, and by "gaps" in RO measurement data, and the mathematical model of measurements is also unknown. Such uncertain and fuzzy a priori information about the considered problems is difficult to formalize and requires a special mathematical apparatus both for the mathematical formulation of their statement and for the mathematical methods of their solution.

For the statement and for solving these tasks we used the mathematical formalism of subjective modeling (MFSM) [1] that allows to mathematically formulate both a subjective mathematical RO model and the subjective model of measurements and their subjective interpretation taking into account the recovered measurement data in "gaps". For this purpose the subjective judgments made by the researcher–modeler (r-m) about RO physical properties and means of its measurements, about mathematical properties of noise etc. are used. All such subjective information is based on r-m's scientific experience and his scientific intuition.

MFSM was created for modeling of uncertainty reflecting incompleteness and unreliability of subjective information about RO, fuzziness and randomness characteristic of its contents characterizing mathematical model of RO. A mathematical model of subjective judgments about the values of an unknown parameter $x \in X$ of the RO model $M(x)$ is defined by r-m as a space $(X, \mathcal{P}(X), \mathrm{Pl}^{\widetilde{x}}, \mathrm{Bel}^{\widetilde{x}})$ with measures of *plausibility* $\mathrm{Pl}^{\widetilde{x}}$ and *belief* $\mathrm{Bel}^{\widetilde{x}}$, where $\widetilde{x}$ is the indeterminate element (i. el.) taking values in X, modeling r-m's subjective judgments about the unknown $x \in X$. The measures $\mathrm{Pl}^{\widetilde{x}}$ and $\mathrm{Bel}^{\widetilde{x}}$ model the modalities of r-m's subjective judgments about the truth of each $x \in X$: namely, the value $\mathrm{Pl}^{\widetilde{x}}(\widetilde{x} = x)$ determines how, in the r-m's opinion, relatively plausible the equality $\widetilde{x} = x$, and the value $\mathrm{Bel}^{\widetilde{x}}(\widetilde{x} \neq x)$ determines to what extent the inequality $\widetilde{x} \neq x$ should be relatively believed. "Relatively" means that the numerical values of the measures $\mathrm{Pl}(\cdot)$ and $\mathrm{Bel}(\cdot)$ which differ from 0 and 1 cannot be meaningfully interpreted, and only their rank order is significant.

The MFSM, in contrast to standard mathematical modeling, enables the r-m to model both accurate, formalized and unformalized, unreliable knowledges, starting from "absolute ignorance" up to "complete knowledge" of the model of the RO and to calculate the plausibility and belief distributions of the truth of any characteristics of the RO which are of interest and caused by its subjective model $M(\widetilde{x})$.

Subjective model of the RO measurements and interpretation of its measurements defined as an element of the parametric family of smoothing splines [2], [3], representing measurement data in the form of a sum of two contributions: a smooth term, optimally evaluating the measurement data, "cleaned" out the noise and a term, optimally estimating the noise, and "optimality" r-m determines subjectively by changing the parameters of the spline. The report investigated the dependence of the quality of solutions to these problems on the number of "missing" measurement data and the relative number of "missed" measurement data.

This research is funded by RFBR, grants 17-07-00832 A, 18-07-00424 A.

[1] *Pyt'ev Y. P.* Modeling of subjective judgments made by a researcher-modeler about the model of the research object // Mathematical Models and Computer Simulations. — 2013. — Vol. 5, no. 6. — P. 538–557.

[2] *Pyt'ev Y. P., Falomkina O. V., Shishkin S. A., Chulichkov A. I.* Mathematical formalism for subjective modeling // Machine Learning and Data Analysis, 2018. Vol. 4, No 5. P. 108–121.

[3] *Yu. P. Pyt'ev, O. V. Falomkina, S. A. Shishkin.* Subjective Restoration of Mathematical Models for a Research Object, Its Measurements, and Measurement-Data Interpretation // Pattern Recognition and Image Analysis, No 4, 2019 (to be printed).

Использование качественной субъективной информации в виде «мягких» неравенств при оценке состава инвестиционного портфеля

Ашарин Влад Викторович[1]⋆ asharin.vlad@gmail.com
Шапошник Григорий Леонидович[1] shaposhnik_grigorii@mail.ru
Фадеев Егор Павлович[1] fadeevegor@yandex.ru
Зубюк Андрей Владимирович[1] zubuk@cmpd2.phys.msu.ru
[1]Москва, МГУ имени М.В.Ломоносова, физический факультет

Часто инвестиционные компании, являющиеся финансовыми посредниками между инвесторами и биржевым рынком, предлагают диверсифицировать (распределить) средства по биржевым активам, вместо того, чтобы вложить все средства в один единственный актив, даже если его ожидаемая доходность велика. Диверсификация позволяет снизить возможные риски при минимальном уменьшении доходности. Таким образом, формируется *инвестиционный портфель*, который можно описать совокупностью *бета-коэффициетов* $\beta_1, \ldots, \beta_n$, где n — общее количество биржевых активов, а β_i — доля средств, вложенный в i-ый актив, $\beta_i \geqslant 0$, $\sum_{i=1}^{n} \beta_i = 1$. Администраторы портфеля, однако, рассматривают его структуру $\beta_1, \ldots, \beta_n$ как свою комерческую тайну. Как правило, инвестиционные компании обязаны предоставлять информацию о *доходности инвестиционного портфеля* d в течнение отчетного периода, которых обычно несколько за все время функционирования портфеля. В свою очередь доходность всех ценных бумаг на мировом фондовом рынке x_i находится в открытом доступе в течение всего периода торговли. Используя эту информацию, можно оценить состав портфеля.

Таким образом для оценки значений β-коэффициентов необходимо решить следующую систему уравнений:

$$\sum_{i=1}^{n} \beta_i x_i = d, \quad \sum_{i=1}^{n} \beta_i = 1, \quad \beta_i \geqslant 0, \tag{1}$$

которая, как мы видим, недоопределена (количестиво уравнений меньше чем количество переменных).

Эта система имеет бесконечно много решений.

Для того чтобы конкретизировать решение, в настоящей работе предлагается принять во внимание субъективные суждения стороннего финансового эксперта о том, в какие именно активы в большей или меньшей степени могла вложить средства та или иная инвестиционная компания. Такие суждения предлагается выражать на языке «мягких» неравенств вида $\beta_{i_k} \gtrsim \beta_{j_k}$, $k = 1, \ldots, r$. «Мягкое» неравенство $\beta_{i_k} \gtrsim \beta_{j_k}$ можно читать как «β_{i_k} скорее всего превосходит β_{j_k}», оно означает, что эксперту портфели, для которых $\beta_{i_k} \geqslant \beta_{j_k}$, кажутся субъективно более правдоподобными, более предпочтительными, чем портфели, для которых $\beta_{i_k} \leqslant \beta_{j_k}$.

Математически «мягкое» неравенство в настоящей работе предлагается определить как нечёткое бинарное отношение действительных чисел, т. е. как нечёткое подмножество плоскости $\mathbb{R} \times \mathbb{R}$. Согласно теории возможностей в варианте Ю. П. Пытьева математической моделью такого нечёткого множества является распределение возможности $\pi_{\gtrsim} : 2^{\mathbb{R}\times\mathbb{R}} \to [0, 1]$, заданное на множестве $2^{\mathbb{R}\times\mathbb{R}}$ всех подмножеств плоскости $\mathbb{R} \times \mathbb{R}$.

В работе исследовано понятие нечёткого неравенства, отвечающее следующим условиям:

1. чем больше разность $a - b$ действительных чисел a и b, тем больше возможность $\pi_{\gtrsim}(R)$ того, что реализация $R \subset \mathbb{R} \times \mathbb{R}$ исследуемого нечёткого бинарного отношения накроет упорядоченную пару чисел (a, b),
2. если реализация R накрыла пару (a, b), то она накрыла и все пары (a', b'), для которых $a' \geqslant a$ и $b \geqslant b'$.

Показано, что задача поиска наиболее возмжных портфелей (являющихся наиболее правдоподобными по субъективному мнению эксперта) среди всех решений задачи (1) может быть сведена к задаче линейного программирования:

$$\begin{cases} z \to \max \\ z \leqslant \beta_{i_k} - \beta_{j_k}, \quad k = 1, \ldots, r \\ \sum_{i=1}^{n} \beta_i x_i = d, \quad \sum_{i=1}^{n} \beta_i = 1, \quad \beta_i \geqslant 0. \end{cases} \tag{2}$$

Проведено исследование предложенного метода.

Работа поддержана грантом РФФИ № 18-07-00424.

Return-based investment portfolio analysis using qualitative subjective information in the form of "soft" inequalities

Asharin Vlad[1]⋆ asharin.vlad@gmail.com
Shaposhnik Grigory[1] shaposhnik_grigorii@mail.ru
Fadeev Egor[1] fadeevegor@yandex.ru
Zubuk Andrew[1] zubuk@cmpd2.phy.msu.ru
[1]Moscow, Lomonosov Moscow State University

Often financial investment companies that are intermediaries between investors and the stock market suggest to diversify (distribute) funds among assets instead of investing all funds in one single asset even if its expected return is high. Diversification reduces potential risks with minimal decrease in profitability. Thus formed *investment portfolio* that can be described collectively with *beta coefficients* $\beta_1, \ldots, \beta_n$, where n is the total number of assets and β_i is the share of funds invested in the i active, $\beta_i \geqslant 0$, $\sum_{i=1}^{n} \beta_i = 1$. Portfolio administrators, however, consider the structure $\beta_1, \ldots, \beta_n$ as its commercial secret. As a rule, investment companies are required to put information about *profitability of the investment portfolio* d during the reporting period, there are usually several periods for all portfolio functioning time. Profitability in turn of all securities in the global stock market x_i is located in open access throughout the trading period. Using this information, you can evaluate the structure of the portfolio.

Thus, to estimate the values β coefficients, it is necessary to solve the following system of equations:

$$\sum_{i=1}^{n} \beta_i x_i = d, \quad \sum_{i=1}^{n} \beta_i = 1, \quad \beta_i \geqslant 0, \tag{1}$$

which, as can be seen, is underdetermined (the number of equations less than the number of variables).

This system has infinitely many solutions.

In order to specify the solution, in this work we propose to take into account subjective judgments of a third-party financial expert about what assets an investment company prefers. Such propositions suggest in the term of "soft" inequalities of the form $\beta_{i_k} \gtrsim \beta_{j_k}$, $k = 1, \ldots, r$. The soft inequality $\beta_{i_k} \gtrsim \beta_{j_k}$ can be read as "β_{i_k} is likely to exceed β_{j_k}", it means that the expert portfolios for which $\beta_{i_k} \geqslant \beta_{j_k}$, seem to be subjectively more believable, more preferred than portfolios, for which $\beta_{i_k} \leqslant \beta_{j_k}$.

Mathematically "soft" inequality in this paper is proposed to define as a fuzzy binary relation between real numbers, i.e., as a fuzzy subset of the plane $\mathbb{R} \times \mathbb{R}$. According to the theory of possibilities in the variant of Yu. P. Pytiev the mathematical model of such a fuzzy set is distribution of the possibility $\pi_{\gtrsim} : 2^{\mathbb{R}\times\mathbb{R}} \to [0, 1]$, given on the set $2^{\mathbb{R}\times\mathbb{R}}$ of all subsets of the plane $\mathbb{R} \times \mathbb{R}$.

The paper explores the concept of fuzzy inequality, which answers following conditions:

1. the greater the difference $a - b$ of the real numbers a and b, the greater the possibility $\pi_{\gtrsim}(R)$ of the realization $R \subset \mathbb{R} \times \mathbb{R}$ of the investigated fuzzy binary relation will cover a pair of numbers (a, b),
2. if the implementation of R covers the pair (a, b), then it covers all pairs (a', b'), for which $a' \geqslant a$ and $b \geqslant b'$.

It is shown that the problem of searching the most possible portfolios (which are the most likely subjective expert opinion) among all solutions to problem (1) can be reduced to the linear programming problem:

$$\begin{cases} z \to \max \\ z \leqslant \beta_{i_k} - \beta_{j_k}, \quad k = 1, \ldots, r \\ \sum_{i=1}^{n} \beta_i x_i = d, \quad \sum_{i=1}^{n} \beta_i = 1, \quad \beta_i \geqslant 0. \end{cases} \tag{2}$$

A study of the proposed method was conducted.

This research is funded by RFBR grant No 18-07-00424.

Классификация над произведением частичных порядков

Дюкова Елена Всеволодовна[1] edjukova@mail.ru
Масляков Глеб Олегович[2]⋆ gleb-mas@mail.ru
Прокофьев Пётр Александрович[3] p_prok@mail.ru
[1]Москва, ВЦ ФИЦ ИУ РАН
[2]Москва, МГУ им. М.В. Ломоносова
[3]Москва, ИМАШ РАН

Актуальность исследования обусловлена существованием прикладных задач машинного обучения, качественное решение которых невозможно в рамках классической постановки логического анализа данных. На основе обобщения базовых понятий предложена схема синтеза корректных логических процедур классификации по прецедентам, ориентированная на задание отношений частичных порядков на множествах значений признаков.

Основное достоинство логического подхода к задаче классификации (распознавания) — возможность получения результата при отсутствии дополнительных предположений вероятностного характера и при небольшом числе прецедентов. Анализ прецедентной информации сводится к поиску определенных закономерностей или элементарных классификаторов, различающих объекты из разных классов. По их наличию или, наоборот, отсутствию в описании распознаваемого объекта, решается вопрос о его классификации. При этом большое внимание уделяется вопросам синтеза корректных алгоритмов, т.е. алгоритмов безошибочно классифицирующих материал обучения. Примерами являются тестовые алгоритмы, алгоритмы голосования по представительным наборам и по покрытиям классов [Дмитриев А.И., Журавлев Ю.И., Кренделев Ф.П., 1966], [Дюкова Е.В., Песков Н.В., 2002].

Существуют сложные задачи, когда не удается найти достаточное количество информативных корректных элементарных классификаторов. Подобная ситуация возникает, например, в случае целочисленных данных высокой значности. Проблема решается применением логических корректоров — корректных распознающих алгоритмов, основанных на построении корректных наборов из некорректных элементарных классификаторов [Дюкова Е.В., Журавлёв Ю.И., Рудаков К.В., 1996] [Дюкова Е.В., Журавлёв Ю.И., Прокофьев П.А., 2017].

При больших размерах признакового пространства возникает необходимость рассматривать сложные в вычислительном плане задачи, которые в теории алгоритмической сложности дискретных задач называют труднорешаемыми. Среди этих задач центральное место принадлежит монотонной дуализации – задаче построения максимальных конъюнкций монотонной булевой функции, заданной конъюнктивной нормальной формой. Задача допускает матричную формулировку с использованием понятия неприводимого покрытия булевой матрицы. Лидерами по скорости счёта являются асимптотически оптимальные алгоритмы [Дюкова, 1977], [Дюкова Е.В., Прокофьев П.А., 2015].

Прикладные задачи классификации не всегда могут быть описаны в рамках классической постановки логической классификации, когда отдельные значения признака сравниваются с использованием отношения равенства. В [1] предложена схема синтеза корректных логических алгоритмов классификации при условии, что на множествах значений целочисленных признаков заданы частичные порядки. Обобщены базовые понятия, используемые при логическом анализе целочисленных данных в задаче классификации по прецедентам, и приведены условия корректности основных логических процедур классификации. Установлено, что анализ прецедентной информации с частичными порядками приводит к необходимости решать задачу дуализации над произведением конечных частичных порядков, простейшим частным случаем которой является монотонная дуализация. Дана матричная формулировка задачи дуализации общего вида и показано, что эта задача сводится к перечислению специальных покрытий целочисленной матрицы. На модельных и реальных данных показана зависимость качества логической классификации от выбора частичных порядков на множествах значений признаков.

Работа частично финансирована РФФИ (проект № 19-01-00430-а)

[1] *Djukova E. V., Masliakov G. O., Prokofyev P. A.* Logical Classification of Partially Ordered Data // arXiv preprint arXiv:1907.08962, 2019.

Classification over partially ordered data

Elena Djukova[1] edjukova@mail.ru
Gleb Masliakov[2]⋆ gleb-mas@mail.ru
Petr Prokofyev[3] p_prok@mail.ru
[1]Moscow, CC FRC CSC RAS
[2]Moscow, MSU
[3]Moscow, IMASH RAS

The importance of this study is caused by the existence of applied machine learning problems that cannot be adequately solved in the classical statement of the logical data analysis. Based on a generalization of basic concepts, a scheme for synthesizing correct supervised classification procedures is proposed. These procedures are focused on specifying partial order relations on sets of feature values.

The main advantage of the logical approach to the classification (recognition) problem is the possibility to obtain a results without additional probabilistic assumptions and using a small number of training objects (using a small number of precedents). The analysis of training data is reduced to finding certain dependences or elementary classifiers (which are subsets of feasible values of some features) that differentiate objects belonging to different classes. An object is classified judging by the presence or absence of such elementary classifiers in the object's description. Special attention is given to synthesizing correct algorithms, i.e., algorithms that unmistakably classify the training objects. Examples are classification by the vote of tests or by the vote of representative sets or by the vote of class coverings [Dmitriev A.N., Zhuravlev Yu.I. and Krendelev F.P., 1966], [Djukova E.V. and Peskov N.V., 2002].

There are complicated problems in which no sufficient number of informative correct elementary classifiers can be found. For example, such a situation occurs when the features can take a large number of possible values. Features that can take real values are often treated as integer valued features with a large number of possible values. A way to solve such problems is to use logical correctors. Here we mean correct recognition algorithms based on constructing correct sets elementary classifiers of from incorrect elementary classifiers [Djukova E.V., Zhuravlev Yu.I. and Rudakov K.V., 1996] [Djukova E.V., Zhuravlev Yu.I., and Prokofyev P.A., 2017].

If the feature space is large, computationally complex (intractable) problems have to be solved. The central place among these problems is occupied by the monotone dualization problem, i.e., the problem of constructing a reduced disjunctive normal form of a monotone Boolean function specified by a conjunctive normal form. This problem can be formulated in terms of matrices using the concept of irreducible covering of a Boolean matrix. The intractability of the monotone dualization problem has two aspects—an exponential growth of the number of solutions as the problem size increases and the complexity of finding (enumerating) these solutions. The

fastest algorithms are the asymptotically optimal algorithms [Djukova, E.V., 1977], [Djukova E.V., Prokofyev P.A., 2015].

Applied classification problems cannot always be described within the classical statement of logical classification in which feature values are compared for equality. In [1], we propose a scheme for synthesizing correct logical algorithms under the condition that partial order relations are specified on the sets of values of integer-valued features. The basic concepts used in the logical analysis of integer data in the supervised classification problem are generalized, and conditions for the correctness of the basic logical classification procedures are obtained. It is found that the analysis of training samples with partial orders requires the dualization problem over the product of finite partial orders to be solved; a simple special case of this problem is monotone dualization. We give a matrix formulation of the general dualization problem and show that this problem is reduced to enumerating special coverings of an integer matrix. The concept of the ordered irredundant covering of an integer matrix is a generalization of the well-known concept of irreducible covering of a Boolean matrix used in the matrix formulation of the monotone dualization problem. Using model and real-life data, we establish the dependence of the quality of logical classification on the choice of partial orders on the sets of feature values.

This study was partially supported by the Russian Foundation for Basic Research, project no. 19-01-00430-a.

[1] *Djukova E. V., Masliakov G. O., Prokofyev P. A.* Logical Classification of Partially Ordered Data // arXiv preprint arXiv:1907.08962, 2019.

Реконструкция треков заряженных частиц с помощью машинного обучения

Шульгин Егор Владимирович[1]⋆ shulgin.ev@phystech.edu
Ратников Федор Дмитриевич[2] fedor.ratnikov@cern.ch
[1]Москва, Московский физико-технический институт
[2]Москва, Высшая школа экономики

В работе рассматривается задача восстановления траекторий ионизованных частиц, зарегистрированных в детекторе после столкновения встречных пучков в коллайдере. Предлагается использовать подход основанный на кластеризации: частицы из одного трека должны попадать в одну группу.

Исходные данные представляют из себя набор трехмерных неупорядоченных векторов, каждый из которых рассматривается как узел графа. Это позволяет применять методы из активно развивающейся области геометрического глубокого обучения. На практике применяется графовая сверточная нейронная сеть для построения нового признакового описания в пространстве более высокой размерности для каждой исходной точки. Далее подсчитывается матрица попарных расстояний между объектами в новом пространстве и применяется кластерный анализ.

Для предложенного метода был поставлен ряд численных экспериментов на специальным образом сгенерированных синтетических данных. Использовались искусственные спиралевидные траектории с добавлением шума из нормального распределения. Проведено сравнение различных архитектур нейронных сетей для решения данной задачи. Также показано, что среднее значение качества кластеризации незначительно уменьшается при увеличении числа треков в событии.

[1] *LHCb collaboration, Ratnikov F. et al.* Measurement of the electron reconstruction efficiency at LHCb // `https://cds.cern.ch/record/2688983`.

Machine Learning for particle tracks reconstruction

Egor Shulgin[1]⋆ shulgin.ev@phystech.edu
Fedor Ratnikov[2] fedor.ratnikov@cern.ch
[1]Moscow, Moscow Institute of Physics and Technology
[2]Moscow, Higher School of Economics

This work considers the problem of reconstructing the trajectories of ionized particles that left a signal ("hit") in radiation detector after beams collision. The authors propose to cluster particles tracks.

It is assumed that input data consists of three-dimensional sparsely measured points that are treated as nodes of a graph. This allows applying methods from Geometric Deep Learning that were developed quite well during the recent past years. Thus dynamic graph convolutional neural network is used for constructing a new higher dimension feature representation for each registered hit and then a clustering algorithm is applied. This leads to grouping the points from the same tracks.

The suggested method was tested on synthetic data: artificially generated spiral-shaped trajectories with added Gaussian noise. It was shown that the average value of the clusterisation quality slightly decreases with an increase in the number of tracks in the event.

[1] *LHCb collaboration, Ratnikov F. et al.* Measurement of the electron reconstruction efficiency at LHCb // https://cds.cern.ch/record/2688983.

Доменное состязательное обучение для понижения смещения прогноза при поиске бозона Хиггса в детекторе ATLAS

Фатхуллин Ильяс Фаизович[1]⋆ ilyas.fn979@gmail.com
Стрижов Вадим Викторович[1] strijov@phystech.edu
[1]Долгопрудный, Московский физико-технический институт

Детальное изучение бозона Хиггса, включая его редкие режимы рождения, является целью экспериментов на Большом адронном коллайдере [**?**]. Анализируется обнаружение бозона Хиггса, образующегося в ассоциации с топ анти-топ кварковой парой (известной как режим рождения ttH). Измерение этого процесса проверяет взаимодействие Юкавы между бозоном Хиггса и топ-кварком. Наиболее вероятным распадом бозона Хиггса является распад на два нижних кварка (измеренные как b-джеты). Данный сигнал требуется отличать от большого фона tt+b-джетов. Это задача бинарной классификации с 41-им признаком, которые имитируют отклик детектора ATLAS на рождение бозона Хиггса. Соответствующим типом события, меткой класса, является сигнал или фон. Симулированные размеченные наборы данных Монте-Карло используются для обучения модели классификации. Однако смещение при обучении в сторону конкретного Монте-Карло генератора снижает обобщающую способность модели. Текущие результаты [1] поиска ttH (H $\to$ bb) ограничиваются моделированием неопределенности симуляции фона, рассчитываемой как расхождение отклика классификатора (смещение при обучении) на различных генераторах Монте-Карло. Цель данной работы состоит в том, чтобы уменьшить это расхождение при обучении.

Два набора данных S_1 и S_2 состоят из 100000 ttH(bb) сигнальных событий в результате моделирования MadGraph/Herwig6. Каждый из них также содержит 100000 фоновых событий. Но S_1 заполнен фоновыми событиями из MadGraph/Pythia6, а S_2 из Powheg Pythia8. Каждое событие описывается 41-м признаком такими как *количество джетов*, *потери поперечной энергии*, *масса ближайших b-джетов* и т.д.

Распределение признаков фоновых событий немного отличается для S_1 и S_2. Целью данного исследования является построение модели классификации, которая показывает адекватное качество на S_2 при обучении на S_1. Качество классификации оценивается по трем критериям: точность, площадь под *ROC*-кривой и значимость. Последнее оценивается как

$$Z_A = \frac{s}{\sqrt{b + \sigma_b^2}}, \tag{1}$$

где s, b соответствуют числу сигнальных и фоновых событий, превышающих фиксированное значение на графике распределения ответов классификатора, а

σ_b^2 есть разница между уровнями двух фонов, превышающих фиксированное значения на графике распределения ответов классификатора.

Изучено применение нейросетей с градиентным реверсивным слоем к поиску ttH(bb) в ATLAS. Это полносвязная трехслойная нейронная сеть, в которую встроена дополнительная нейронной сетью под названием *классификатор доменов*. Классификатор доменов используется только во время обучения и предназначен для сближения распределений признаков в скрытом пространстве для двух заданных наборов данных.

Эксперименты показали, что состязательная доменная адаптация снижает смещение прогноза к заданной симуляции, сохраняя при этом адекватное качество классификации.

[1] *Collaboration ATLAS.* Search for the Standard Model Higgs boson produced in association with top quarks and decaying into a $b\bar{b}$ pair in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector, 2019. `arXiv:1712.08895`.

Domain Adversarial Learning to Reduce Training Bias in ttH(bb) Search at ATLAS

Fatkhullin Ilyas[1]⋆ ilyas.fn979@gmail.com
Strijov Vadim[1] strijov@phystech.edu
[1]Dolgoprudny, Moscow Institute of Physics and Technology

The detailed study of the Higgs boson including its rare production modes is a primary activity of the Large Hadron Collider (LHC) experiments [1]. The detection of the Higgs boson produced in the association with a top anti-top quark pair (known as ttH production mode) is analysied. A measurement of this process tests the Yukawa coupling between the Higgs boson and the top quark. The most probable decay of the Higgs boson is to two bottom quarks (measured as b-jets) and the signal needs to be distinguished from a large background of tt+b-jets. Formally, this is a binary classification problem with 41 input variables (features) that simulate ATLAS detector response on the processes. Its corresponding event type (class label) is either signal or background. The simulated labelled Monte Carlo datasets are utilised to train the classification model. However, the training bias towards a specific Monte Carlo generator reduces the generalisation capabilities of the model. The current results [1] for the ttH (H $\to$ bb) search was limited to model the uncertainty of the background simulation, calculated as the discrepancy of the classifier response (training bias) to different Monte Carlo generators. The aim is to reduce this training bias.

Two datasets S_1 and S_2 consist of 100000 ttH(bb) signal events from MadGraph/Herwig6 simulation. Each of them also have 100000 background events. But S_1 is filled with background events from MadGraph/Pythia6, while S_2 from Powheg Pythia8. Each event is described by 41 features such as *number of jets*, *missing transverse energy*, *mass of closest b-jets* etc.

The feature distributions of the background events appear to be slightly different for S_1 and S_2. The goal of this study is to build a classification model, which shows an adequate performance on S_2 while being trained on S_1. The classification performance is mainly assessed based on accuracy, area under ROC-curve, and significance. The latter is approximated as follows:

$$Z_A = \frac{s}{\sqrt{b + \sigma_b^2}}, \tag{1}$$

where s, b correspond to the number of signal and background events, above the fixed cut in the classifier response plot, and σ_b^2 is the difference between two backgrounds above the cut.

The application of Neural Networks with a gradient reversal layer to the ttH(bb) search at ATLAS is studied. This is a simple Feed Forward NN with an additional NN called *domain classifier*. The domain classifier is used during training only

and aimed to match the feature distributions in the latent space for the two given datasets.

Extensive experiments have demonstrated that adversarial domain adaptation reduces the training bias towards a given simulation while preserving adequate classification performance.

[1] *Collaboration ATLAS.* Search for the Standard Model Higgs boson produced in association with top quarks and decaying into a $b\bar{b}$ pair in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector, 2019. `arXiv:1712.08895`.

Введение отношения порядка на множестве параметров нейронной сети

Грабовой Андрей Валериевич[1]⋆ grabovoy.av@phystech.edu
Бахтеев Олег Юриевич[1] 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

Оптимизация глубоких нейронных сетей является задачей высокой вычислительной сложности и требует больших временных затрат и вычислительных мощностей. При этом оптимизация сходится по большинству параметров сети уже после небольшого числа итераций.

Данная работа предлагает метод введения отношения порядка на множестве параметров нейронной сети. Порядок задается при помощи ковариационной матрицы градиентов функции ошибки по параметрам модели. Порядок используется для фиксации параметров модели во время решения оптимизационной задачи. Предполагается, что после небольшого числа интеграций метода оптимизации некоторые параметры модели можно зафиксировать без значимой потери качества модели. Это позволит существенно снизить размерность задачи оптимизации.

Для анализа качества представленного метода проводились вычислительные эксперименты на синтетических и реальных данных. Сравнивались модели, в которых параметры фиксируются в соответсвии с заданым порядком с моделями, в которых параметры фиксируются произвольным образом. Показано, что порядок заданный при помощи ковариационной матрицы является адекватным, так как позволяет зафиксировать значимое число параметров без значимой потери качества. Также показано, что предложенный порядок является устойчивым и не меняется от запуска к запуску метода оптимизации.

Работа выполнена при поддержке РФФИ (проекты 19-07-1155, 19-07-0875) и НТИ (проект 13/1251/2018).

[1] *Грабовой А. В. Бахтеев О. Ю. Стрижов В. В.* Задания порядка на множестве параметров аппроксимирующих моделей // Информатика и ее применения, 2020.

Order on the set of neural network parameters

Andrey Grabovoy[1]⋆ grabovoy.av@phystech.edu
Oleg Bakhteev[1] bakhteev@phystech.edu
Vadim Strijov[1] strijov@phystech.edu
[1]Moscow, Moscow Institute of Physics and Technology
[2]Moscow, FRCCSC of the Russian Academy of Sciences

Optimization of deep neural networks is a high complexity computational task. It requires a lot of time and computational power. But the optimization converges in most network parameters after a small number of iterations.

This paper investigates a method for setting order on a set of the neural network parameters. It is proposed to set this order using the covariance matrix of the gradients. The order is used to freeze the model parameters during the optimization. It is assumed that after a few iterations of the optimization method, many model parameters can be freezed without a significant loss in model quality.

The proposed method was tested on the real dataset and synthetic data. The experiment shown that models which parameters were ordered using proposed method, are more resistant to parameters freezing. It is shown that proposed order allows to reduce a number of optimizing parameter without significant loss of quality. It is also shown that the proposed order is stable and does not change in different optimization running.

This research was supported by RFBR (projects 19-07-1155, 19-07-0875) and NTI (project 13/1251/2018).

[1] *Grabovoy A. Bakhteev O. Strijov V.* Automatic search for the relevance of neural network parameters // Informatics and Applications, 2020.

Численные методы оценки оптимального объёма выборки для логистической и линейной регрессии

Гадаев Тамаз Тезикоевич[1]⋆ gadaev.tt@phystech.edu
Грабовой Андрей Валерьевич[1] grabovoy.av@phystech.edu
Мотренко Анастасия Петровна[1] anastasiya.motrenko@phystech.edu
Стрижов Вадим Викторович[1,2] strijov@phystech.edu
[1]Москва, Московский физико-технический институт
[2]Москва, Вычислительный центр им. А.А. Дороницына ФИЦ ИУ РАН

Сбор данных для дальнейшего решения задачи классификации или регрессии может быть затратным. В связи с этим исследуется проблема нахождения оптимального объёма выборки для построения линейной или логистической регрессии.

Для построения модели необходимо, чтобы в выборке было достаточное число объектов, а также чтобы выборка не противоречила гипотезе порождения данных. В эту гипотезу входят предположения о составе выборки и о свойствах модели, оптимально описывающей выборку согласно принятому критерию. Модель, не противоречащая гипотезе порождения данных называется адекватной. В качестве базовых предположений принимаются предположения о простоте и однородности выборки. Выборка, необходимый объем которой требуется оценить, адекватно аппроксимируется одной обобщенно-линейной моделью. Предпочтительны методы определения объема выборки, позволяющие строить адекватные модели по меньшим выборкам.

Данная работа проводит анализ численных свойств методов, которые используются для оценки выборки и предлагает возможные варианты их улучшения. В анализ входят методы, оценивающие объем выборки исходя из гипотезы порождения данных, использующие эвристические предположения, а также методы, учитывающие структуру модели, которая будет построена. Вычислительный эксперимент включает часто используемые открытые выборки.

Работа выполнена при поддержке РФФИ (проекты 17-20-01212, 19-07-0885) и НТИ (проект 13/1251/2018).

Numerical methods of sample size estimation for linear and logistic regression

Tamaz Gadaev[1]⋆ gadaev.tt@phystech.edu
Andriy Grabovoy [1] grabovoy.av@phystech.edu
Anastasia Motrenko[1] anastasiya.motrenko@phystech.edu
Vadim Strijov[1,2] strijov@phystech.edu
[1]Moscow, Moscow Institute of Physics and Technology
[2]Moscow, FRCCSC of the Russian Academy of Sciences

Data collection for building the prognostic model can be very expensive. This work investigates a problem of estimation of sufficient sample size for building linear or logistic regression.

To build the model it is necessary to have a sample set that contains sufficient number of objects and that is not in controversy with data generation hypothesis. This hypothesis includes assumptions of sample set structure and properties of model which optimally describes the sample set according to the chosen criterion. Model which is not in a controversy with data generation hypothesis is called adequate. Basic assumptions is that objects in sample are independent and identically distributed. Sample set, which size is estimated, can be adequately approximated by one generalized linear model. Preferable methods of sample size estimation are which allow fitting models on the lesser samples.

This work investigates numerical properties of sample size estimation methods. Methods using data generation hypothesis are included in the analysis as well as methods based on the structure of the linear model that is built.

This research was supported by RFBR (projects 17-20-01212, 19-07-0885) and NTI (project 13/1251/2018).

Метрическая кластеризация ранжирований

Двоенко Сергей Данилович[1]⋆ sergdv@yandex.ru
Пшеничный Денис Олегович[1] denispshenichny@yandex.ru
[1]Тула, Тульский государственный университет

При обработке данных, представленных матрицами парных сравнений, применяются специально разработанные нами варианты алгоритмов машинного обучения (кластеризации, группировки, оптимального разделения). Для представления отсутствующих элементов множества (например, средних и др.) существенно эксплуатируется т.н. формула Торгерсона. С ее помощью новые элементы представлены своими сравнениями с остальными элементами множества (расстояниями или соответствующими скалярными произведениями) [1].

При решении задачи согласования индивидуальных мнений, представленных ранжированиями, часто применяется известный алгоритм построения медианы Кемени. Алгоритм Кемени основан на введении расстояний между ранжированиями. Но, как только это сделано, то появляется матрица парных расстояний между ранжированиями, позволяющая вычислить центральный элемент (это искомое ранжирование) множества, представленный своими расстояниями до остальных элементов множества. По смыслу – это тоже медиана. В общем случае такому ранжированию соответствуют разные элементы множества, представленные своими расстояниями до остальных, и наоборот (среди них и медиана Кемени). При согласовании большого числа ранжирований появление неразличимых альтернатив говорит о наличии групп сильно различающихся мнений экспертов. Таким образом, возникает задача кластеризации, где применение метрических медиан позволяет решать ее известными алгоритмами кластер-анализа.

Данный подход рассматривается нами как новый при решении проблемы метризации бинарных отношений.

Работа поддержана грантами РФФИ № 17-07-00319, 18-07-01087, 18-07-00942.

[1] *Dvoenko S.D., Pshenichny D.O.* On metric correction and conditionality of raw featureless data in machine learning // Pattern Recognition and Image Analysis, Pleiades Publishing, 2018. Vol. 28, No.4. p. 595–604.

A metric clustering of rankings

Dvoenko Sergey[1]⋆ sergdv@yandex.ru
Pshenichny Denis[1] denispshenichny@yandex.ru
[1]Tula, Tula State University

For data presented by matrices of paired comparisons, we use specially developed versions of machine learning algorithms (for clustering, grouping, optimal separation). To represent the missing elements of the set (for example, means, etc.), the so-called Torgerson formula is essentially used. Based on it, new elements are represented by their comparisons with the rest of elements of the set (by distances or corresponding scalar products) [1].

When solving the concordance problem of individual opinions represented by rankings, the well-known Kemeny's median algorithm is often used. Kemeny's algorithm is based on distances between rankings. But, as soon as the matrix of paired distances between the rankings appears, it allows to calculate the central element (this is the desired ranking) of the set, represented by its distances to the other elements of the set. Therefore, this is the median too, by sence. In general case, this ranking can correspond to different elements of the set, represented by their distances to the rest, and vice versa (the Kemeny's median is among them). The appearance of indistinguishable alternatives when to coordinate a large number of rankings indicates the presence of groups of quite different opinions of experts. Therefore, there is a problem of clustering, where using of metric medians allows to solve it by known algorithms of cluster analysis.

We consider this approach as a new one in solving the problem of metrization of binary relations.

This research is funded by RFBR, grants 17-07-00319, 18-07-01087, 18-07-00942.

[1] *Dvoenko S.D., Pshenichny D.O.* On metric correction and conditionality of raw featureless data in machine learning // Pattern Recognition and Image Analysis, Pleiades Publishing, 2018. Vol. 28, No.4. p. 595–604.

Матричная коррекция ограничений несобственных задач линейного программирования в задаче распознавания образов с пересекающимися классами

Ерохин Владимир Иванович[1] erohin_v_i@mail.ru
Красников Александр Сергеевич[2] askrasnikov@gmail.com
Волков Владимир Викторович[3]⋆ volkov@fizmat.net

[1]Санкт-Петербург, Военно-космическая академия имени А. Ф. Можайского
[2]Москва, Московский политехнический университет
[3]Борисоглебск, Борисоглебский филиал Воронежского государственного университета

Настоящая работа посвящена применению оптимальной матричной коррекции данных к задачам распознавания образов.

Рассмотрим простейшую формулировку задачи распознавания образов. Исходными данными являются описания объектов S в виде векторов значений признаков $S = (v_1(S), v_2(S), \ldots, v_n(S))$ и значения класса $\Omega(S)$, которому принадлежит объект.

Предполагается, что существует функциональная связь между признаками и классом. Задача распознавания состоит в определении значения класса $\Omega(S)$ некоторого объекта S по информации (обучающей выборке) о классах объектов $S_1, S_2, \ldots, S_m$.

Рассмотрим алгоритм распознавания, основанный на построении линейных разделяющих поверхностей (гиперплоскостей) [1]. Для простоты будем считать, что имеются два класса объектов.

Задача состоит в вычислении некоторой линейной относительно признаков функции $f(v) = \lambda_1 v_1 + \lambda_2 v_2 + \ldots + \lambda_n v_n + 1 = \lambda^T v + 1$ и использовании решающего правила $\Omega(S) = \begin{cases} 1, \text{ if } f(S)>0, \\ \Delta, \text{ if } f(S)=0, \\ 0, \text{ if } f(S)<0. \end{cases}$ [1].

Необходимо найти значение неизвестных коэффициентов $\lambda_1, \lambda_2, \ldots, \lambda_n$, являющихся решением системы

$$\begin{cases} f(v(S_1))>0, \\ \cdots \\ f(v(S_{m_1}))>0, \\ f(v(S_{m_1+1}))<0, \\ \cdots \\ f(v(S_m))<0. \end{cases} \tag{1}$$

Если система (1) совместна, то достаточно найти произвольное ее решение относительно неизвесных $\lambda_1, \lambda_2, \ldots, \lambda_n$.

Заранее неизвестно, совместна данная система или нет. Обычно, вследствие различных проблем моделирования, система (1) оказывается несовместной, т.е. объекты обучающей выборки невозможно безошибочно разделить гиперплоскостью. В этом случае находится некоторое обобщенное решение $\bar{\lambda}_1, \bar{\lambda}_2, \ldots, \bar{\lambda}_n$

системы (1) — решение ее максимальной совместной подсистемы, либо любое другое решение, удовлетворяющее ЛПР.

Пусть получено обобщенное решение системы (1) и, соответственно, коэффициенты $\bar{\lambda}_1, \bar{\lambda}_2, \ldots, \bar{\lambda}_n$ линейной функции $f(v)$, определяющей положение разделяющей гиперплоскости.

Сформулируем и решим следующую задачу: минимально исправить положение объектов таким образом, чтобы можно было провести новую разделяющую гиперплоскость, по возможности близкую к исходной и строго разделяющую области.

Задача P_r:

$$\sum_{j=1}^{n} |\lambda_j - \bar{\lambda}_j| \to \min, \quad \sum_{i=1}^{m}\sum_{j=1}^{n} e_{i,j}^2 \to \min,$$

$$\begin{cases} \lambda_1(v_{1,1}+e_{1,1}) + \ldots + \lambda_n(v_{1,n}+e_{1,n}) + 1 = \delta_1, \\ \ldots \\ \lambda_1(v_{m_1,1}+e_{m_1,1}) + \ldots + \lambda_n(v_{m_1,n}+e_{m_1,n}) + 1 = \delta_{m_1}, \\ \lambda_1(v_{m_1+1,1}+e_{m_1+1,1}) + \ldots + \lambda_n(v_{m_1+1,n}+e_{m_1+1,n}) + 1 = \delta_{m_1+1}, \\ \ldots \\ \lambda_1(v_{m,1}+e_{m,1}) + \ldots + \lambda_n(v_{m,n}+e_{m,n}) + 1 = \delta_m, \end{cases}$$

где $v_{i,j} = v_j(S_i)$, $\delta_i = \begin{cases} \varrho_i, \text{ если } \varrho_i > 0, \\ \sigma \text{ в противном случае,} \end{cases}$ для $i = 1, 2, \ldots, m_1$, $\delta_i = \begin{cases} \varrho_i, \text{ если } \varrho_i < 0, \\ -\sigma \text{ в противном случае,} \end{cases}$ для $i = m_1+1, m_1+2, \ldots, m$, $\varrho_i = \bar{\lambda}_1 v_{1,1} + \bar{\lambda}_2 v_{1,2} + \ldots + \bar{\lambda}_n v_{1,n} + 1$, $\sigma > 0$ — некоторое число, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$.

Или в матричной форме:

$$\|\lambda - \bar{\lambda}\|_1 \to \min, \quad \|E\|^2 \to \min, \quad (V+E)\cdot\lambda = p,$$

где $p = \begin{bmatrix} \delta_1 - 1 \\ \vdots \\ \delta_m - 1 \end{bmatrix}$.

Исходную задачу можно переформулировать: из всех возможных матриц $H = \begin{bmatrix} D_1 & -D_2 \end{bmatrix}$ найти матрицу H^* с минимальной евклидовой нормой такую, чтобы задача линейного программирования (ЛП)

$$(A + H^*)x = b, \quad x \geqslant 0, \quad c^T x \to \max, \tag{2}$$

оказалась собственной, где $x = \begin{bmatrix} x_{(1)} \\ x_{(2)} \end{bmatrix}$, $A = \begin{bmatrix} V & -V \end{bmatrix}$, $c = -l_{2n}$, $c, x \in R^N$, $b \in R^m$, $A, H \in R^{m\times N}$, $N = 2n + m$.

Учитывая тот факт, что коррекция допустимой области задачи ЛП без обеспечения непустоты допустимой области соответствующей двойственной задачи, не гарантирует ее собственность, получаем задачу оптимальной по минимуму евклидовой матричной нормы коррекции допустимой области задачи ЛП.

Отметим, что применение коррекции данных к задаче классификации исследовалось В.Л. Матросовым, В.А. Гореликом, С.А. Ждановым, О.В. Муравьевой (см. например, [2]). Данная задача сведена ими к задаче коррекции несовместной системы линейных неравенств. Основным отличием настоящей работы является возможность априори задавать примерное желаемое положение разделяющей гиперплоскости и возможность учета структуры некорректируемых координат исходной задачи.

В ходе вычислительного эксперимента данный метод был применен для решения задачи распознавания на модельных данных. После коррекции классы стали линейно разделимы.

Работа поддержана грантом РФФИ №18-31-00083.

[1] *Журавлев И.Ю., Рязанов В.В, Сенько О.В.* Распознавание. Математические методы. Программная система. Практические приложения.// Москва: Фазис, 2006.

[2] *Матросов В.Л., Горелик В.А., Жданов С.А., Муравъева О.В.* Коррекция данных в задаче классификации // Математические методы распознавания образов: Доклады XI Всероссийской конф. (ММРО-11). - М.: ВЦ РАН, 2003. - С. 136-137.

Matrix correction of restrictions of improper linear programming problems in the problem of pattern recognition with intersecting classes

Erokhin Vladimir[1] erohin_v_i@mail.ru
Krasnikov Alexander[2] askrasnikov@gmail.com
Volkov Vladimir[3]⋆ volkov@fizmat.net
[1]St. Petersburg, Mozhaisky Military Space Academy
[2]Moscow, Moscow Polytechnic University
[3]Borisoglebsk, Borisoglebsk branch of Voronezh State University

Present paper is devoted to the application of optimal matrix correction to the tasks of pattern recognition.

Let us consider the simplest formulation of recognition problem.

As the initial information in pattern recognition there are descriptions of objects in the form of vectors of attribute values for the objects: $S = (v_1(S), v_2(S), \ldots, v_n(S))$ and the values of the class $\Omega(S)$ of object S. The $\Omega(S)$ property can take a finite number of values.

Supposed that there is a functional relationship between the attributes v_j and the object class. The problem of pattern recognition is determining the value of the property $\Omega(S)$ of some object S using the information from training set $(S_1, S_2, \ldots, S_m)$.

Consider a recognition algorithm based on the construction of the dividing surfaces [1]. Let us consider the simplest case when there are only two classes of objects.

The problem of constructing a linear separating surface (hyperplanes) consists in calculating some linear function $f(v) = \lambda_1 v_1 + \lambda_2 v_2 + \ldots + \lambda_n v_n + 1 = \lambda^T v + 1$ and using the decision rule $\Omega(S) = \begin{cases} 1, \text{ if } f(S)>0, \\ \Delta, \text{ if } f(S)=0, \\ 0, \text{ if } f(S)<0. \end{cases}$ [1].

To solve the problem we need to find unknown coefficients $\lambda_1, \lambda_2, \ldots, \lambda_n$ being a solution of the system

$$\begin{cases} f(v(S_1))>0, \\ \cdots \\ f(v(S_{m_1}))>0, \\ f(v(S_{m_1+1}))<0, \\ \cdots \\ f(v(S_m))<0. \end{cases} \tag{1}$$

If the system (1) is consistent, then it is sufficient to find its arbitrary solution for unknown $\lambda_1, \lambda_2, \ldots, \lambda_n$.

However, it is unknown in advance whether this system is consistent or not. Usually, due to various modeling problems, the system (1) turns out to be inconsistent, i.e. training sets cannot be unmistakably divided by a hyperplane. If the system is inconsistent, there is some generalized solution $\bar{\lambda}_1, \bar{\lambda}_2, \ldots, \bar{\lambda}_n$ of the system (1) — the solution of some of its maximum joint subsystem, or any other solution satisfying the decision maker.

Let there are generalized solution of the system (1) and coefficients $\bar{\lambda}_1, \bar{\lambda}_2, \ldots, \bar{\lambda}_n$ of linear function $f(v)$, which determines the position of the dividing hyperplane, are known.

We formulate and solve the following problem: to correct (minimally in some sense) the position of objects in space so that it will be possible to draw a new strictly separating classes hyperplane close to the original hyperplane.

Problem P_r:

$$\sum_{j=1}^{n} |\lambda_j - \bar{\lambda}_j| \to \min, \quad \sum_{i=1}^{m}\sum_{j=1}^{n} e_{i,j}^2 \to \min,$$

$$\begin{cases} \lambda_1(v_{1,1} + e_{1,1}) + \ldots + \lambda_n(v_{1,n} + e_{1,n}) + 1 = \delta_1, \\ \ldots \\ \lambda_1(v_{m_1,1} + e_{m_1,1}) + \ldots + \lambda_n(v_{m_1,n} + e_{m_1,n}) + 1 = \delta_{m_1}, \\ \lambda_1(v_{m_1+1,1} + e_{m_1+1,1}) + \ldots + \lambda_n(v_{m_1+1,n} + e_{m_1+1,n}) + 1 = \delta_{m_1+1}, \\ \ldots \\ \lambda_1(v_{m,1} + e_{m,1}) + \ldots + \lambda_n(v_{m,n} + e_{m,n}) + 1 = \delta_m, \end{cases}$$

where $v_{i,j} = v_j(S_i)$, $\delta_i = \begin{cases} \varrho_i, \text{ if } \varrho_i > 0, \\ \sigma \text{ otherwise,} \end{cases}$ for $i = 1, 2, \ldots, m_1$, $\delta_i = \begin{cases} \varrho_i, \text{ if } \varrho_i < 0, \\ -\sigma \text{ otherwise,} \end{cases}$ for $i = m_1 + 1, m_1 + 2, \ldots, m$, $\varrho_i = \bar{\lambda}_1 v_{1,1} + \bar{\lambda}_2 v_{1,2} + \ldots + \bar{\lambda}_n v_{1,n} + 1$, $\sigma > 0$ – some number, $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, n$.

Or in matrix form

$$\|\lambda - \bar{\lambda}\|_1 \to \min, \quad \|E\|^2 \to \min, \quad (V + E) \cdot \lambda = p,$$

where $p = \begin{bmatrix} \delta_1 - 1 \\ \vdots \\ \delta_m - 1 \end{bmatrix}$.

Source problem can be rewritten as follows: from all possible matrices $H = \begin{bmatrix} D_1 & -D_2 \end{bmatrix}$ find the matrix H^* with the minimal Euclidean norm such that the linear programming problem

$$(A + H^*)x = b, \quad x \geqslant 0, \quad c^T x \to \max, \tag{2}$$

being proper. $x = \begin{bmatrix} x_{(1)} \\ x_{(2)} \end{bmatrix}$, $A = \begin{bmatrix} V & -V \end{bmatrix}$, $c = -l_{2n}$, $c, x \in R^N$, $b \in R^m$, $A, H \in R^{m \times N}$, $N = 2n + m$.

Considering the fact that the correction of an admissible domain of a linear programming problem without ensuring that the admissible domain of the corresponding dual problem is non-empty, does not guarantee that the problem is feasible, we obtain the problem of optimal data correction of the dual pair of improper linear programming problems with respect to the minimum of the Euclidean norm.

It should be noted that the applications of data correction to the classification problem was investigated by V.L. Matrosov, V.A. Gorelik, S.A. Zhdanov, O.V. Muravyova. (see, for example, [2]). They reduced this problem to the problem of correcting an inconsistent system of linear inequalities. The main feature of this paper is the ability to a priori set the approximate desired position of the separating hyperplane and the ability to take into account the structure of the initial problem.

In computational experiment we have used described method for solving pattern recognition problem for model data. After the correction classes became separable.

This research is funded by RFBR, grant 18-31-00083.

[1] *Zhuravlev Yu., Ryazanov V, Senko O.* Recognition. Mathematical methods. Software system. Practical application.// Moscow: Fazis, 2006. (in Russian)

[2] *Matrosov V., Gorelik V., Zhdanov S., Muravyova O.* Correction of data in the classification problem // Mathematical methods of pattern recognition: Reports of the XI All-Russian Conf. (MMPR-11), Moscow: CC RAS, 2003/ — p. 136–137. (in Russian)

Технология коррекции и обработки парных сравнений

Двоенко Сергей Данилович[1] sergdv@yandex.ru
Пшеничный Денис Олегович[1]⋆ denispshenichny@yandex.ru
[1]Тула, Тульский государственный университет

Результаты парных сравнений элементов множества (объекты или признаки) организованы в квадратную матрицу. Совокупность элементов образует конфигурацию в метрическом пространстве (гипотетическом, если реальных измерений признаков нет). Сквозная технология корректировки заключается в преобразовании произвольной матрицы парных сравнений к матрице взвешенных скалярных произведений, собственно корректировке одним из ранее разработанных нами методов, восстановлению исходной матрицы. Результатом коррекции является последовательность положительных значений главных миноров матрицы взвешенных скалярных произведений (это обеспечивает отсутствие отрицательных собственных чисел) при условии минимизации отклонений скорректированных значений от исходных [1].

Сквозная технология реализует систематический подход к обработке экспериментальных данных в виде парных сравнений и является основой для развития новых методов интеллектуального анализа данных. В рамках сквозной технологии разработан ряд процедур для улучшения результата коррекции (индивидуальная и групповая коррекции, локализация отрицательных собственных чисел, оптимальная обусловленность, вынос начала координат за пределы выпуклой оболочки и т.п.). На основе некоторых из них реализованы новые методы анализа данных (группировка признаков без построения факторов групп, метрическое согласование индивидуальных ранжирований). Развитие технологии предполагает разработку модификации метода Карунена-Лоэва при коррекции матриц парных сравнений и др.

Работа поддержана грантами РФФИ № 17-07-00319, 18-07-01087, 18-07-00942.

[1] *Dvoenko S.D., Pshenichny D.O.* On metric correction and conditionality of raw featureless data in machine learning // Pattern Recognition and Image Analysis, Pleiades Publishing, 2018. Vol. 28, No.4. p. 595–604.

The technology of correction and processing of pairwise comparisons

Dvoenko Sergey[1] sergdv@yandex.ru
Pshenichny Denis[1]⋆ denispshenichny@yandex.ru
[1]Tula, Tula State University

The results of paired comparisons of elements of the set (objects or features) are organized into a square matrix. The set of elements forms a configuration in metric space (hypothetical if there are no real measurements of features). The end-to-end correction technology consists in the transformation of an arbitrary matrix of paired comparisons to a matrix of weighted scalar products, the correction by one of the previously developed methods, the restoration of the original matrix. The result of the correction is a sequence of positive values of the principal minors of the matrix of weighted scalar products (this ensures the absence of negative eigenvalues) under condition that the deviations of the corrected values from the original ones are minimized [1].

The end-to-end technology implements a systematic approach to the processing of experimental data in the form of paired comparisons and is the basis for the development of new methods of intelligent data processing. Within the framework of this technology, a number of procedures have been developed to improve the correction result (individual and group correction, localization of negative eigenvalues, optimal conditionality, removal of the origin beyond the convex hull, etc.). On the basis of some of them, new methods of data analysis have been developed (grouping of features without group factors, metric concordance of individual rankings). It is supposed a modification of the Karhunen-Loeve method for the correction of matrices of paired comparisons, etc.

This research is funded by RFBR, grants 17-07-00319, 18-07-01087, 18-07-00942.

[1] *Dvoenko S.D., Pshenichny D.O.* On metric correction and conditionality of raw featureless data in machine learning // Pattern Recognition and Image Analysis, Pleiades Publishing, 2018. Vol. 28, No.4. p. 595–604.

Высокопроизводительный метод средних решающих правил для решения больших двухклассовых задач SVM в пространстве признаков

Курбаков Михаил Юрьевич[1]⋆ muwsik@mail.ru
Макарова Александра Игоревна[1] aleksarova@gmail.ru
Сулимова Валентина Вячеславовна[1] vsulimova@yandex.ru
[1]Тула, Тульский государственный университет

Важной тенденцией современных задач двухклассового распознавания является необходимость обучения в условиях большого объёма данных. В таких случаях многие хорошо зарекомендовавшие себя, методы, например, такие, как метод опорных векторов (SVM), оказываются неприменимыми из-за колоссальной трудоёмкости процедуры обучения.

В одной из предшествующих работ [1], нами был предложен относительно простой подход, позволяющий быстро найти достаточно близкое к точному решение задачи SVM в линейном признаковом пространстве, получивший название метод средних решающих правил (MDR). А в данной работе мы предлагаем его высокопроизводительную реализацию. При этом основной упор делается на повышение производительности этапа обучения.

Основная идея метода средних решающих правил заключается в усреднении частных решающих правил, построенных по случайным подвыборкам исходной обучающей совокупности.

В теории метод MDR обладает достаточно высокой степенью параллелизма по данным. Однако на практике повышению производительности может помешать использование распространенного формата хранения данных libsvm, а также классического способа работы с данными, подразумевающего единовременную загрузку всей обучающей совокупности в оперативную память. В данном случае невозможно быстро по номеру объекта определить его местоположение в файле, а время загрузки всех данных возрастает с ростом обучающей совокупности, которая, к тому же, может не поместиться целиком в оперативную память компьютера.

Для решения указанной проблемы в данной работе предложена двухэтапная стратегия работы с данными.

Первый этап заключается в осуществлении предварительной разметки исходного файла для обеспечения возможности быстрого поиска нужных объектов при формировании случайных подвыборок.

Разметка предполагает деление исходного файла на большое число фрагментов, независимое определение количества объектов каждого класса в каждом фрагменте и последующее объединение соответствующей информации. Результатом разметки является единый массив, который хранит диапазоны номеров объектов каждого класса, принадлежащих отдельным фрагментам файла.

При этом вместо традиционных операций чтения файла на диске предлагается использовать более быстрые операции отображения файла в оперативную память процесса.

На втором этапе информация о диапазонах номеров объектов используется для формирования случайных подвыборок. При этом подвыборки формируются непосредственно в процессе обучения по мере необходимости, что обеспечивает существенную экономию оперативной памяти.

В силу независимости большинства операций этап разметки допускает достаточно эффективную параллельную реализацию.

Этап формирования подвыборок также обладает высокой степенью параллелизма по данным, поскольку все подвыборки формируются независимо друг от друга.

Кроме того в целом предлагаемый подход не требователен к памяти как на этапе разметки, так и на этапе обучения, поскольку в силу особенностей метода средних решающих правил нет необходимости загружать все подвыборки одновременно. Дополнительно объем используемой оперативной памяти может быть уменьшен за счет уменьшения числа фрагментов отображения исходного файла в память и уменьшения размера подвыборок.

Таким образом, можно считать, что предлагаемый подход не имеет ограничения на объем обучающей совокупности.

Единственный недостаток такого подхода заключается в необходимости поиска и чтения объектов посредственно в процессе обучения. Однако это компенсируется за счёт возможности параллельной обработки.

Эксперименты на реальном крупном наборе данных из репозитория libsvm показывают, что разработанный подход имеет квазилинейное ускорение этапа обучения и совместно с предложенной стратегией работы с данными позволяет повысить производительность вычислений по сравнению с классическими способами работы с данными в том же формате.

Исследование выполнено при финансовой поддержке РФФИ в рамках научных проектов 17-07-00436, 18-07-00942, 18-07-01087.

[1] *Макарова А. И., Сулимова В. В.* Метод средних решающих правил для быстрого двухклассового обучения в пространстве, порожденном потенциальной функцией // Сборник трудов V международной конференции и молодежной школы "Информационные технологии и нанотехнологии"(ИТНТ-2019), Том 4. Науки о данных – Самара: Новая техника, 2019, С. 25–34.

High-performance MDR method for solving large two-class SVM problems in the feature space

Mikhail Kurbakov[1]⋆ muwsik@mail.ru
Alexandra Makarova[1] aleksarova@gmail.ru
Valentina Sulimova[1] vsulimova@yandex.ru
[1]Tula, Tula State University

An important trend in modern problems of two-class recognition is the need for training in a big data sets. In such cases, many well-proven methods, for example, such as the Support Vector Machines (SVM), are not applicable due to the extremely high complexity of the training procedure.

In one of our previous works [1], we proposed a relatively simple approach that allows you to quickly find a solution that is quite close to the exact solution of the SVM problem in a linear feature space. We named it the Mean Decision Rules method (MDR). In this work, we offer its high-performance implementation. At that the main emphasis is on increasing the productivity of the training stage.

The main idea of the MDR is to average particular decision rules constructed from random subsamples of the initioal training set.

In theory, the MDR method has a fairly high degree of data parallelism. However, in practice, performance can be hindered by the use of 1) the common data storage format libsvm due to in this case it is impossible to quickly determine an object's location in the file by its number and 2)the classic way of working with data, which implies a one-time loading of the entire training set into RAM, because in this case the loading time of full data set increases with the growth of the amount of data, which, moreover, may not fit into the RAM.

To solve this problem, a two-stage data strategy is proposed in this paper.

The first stege is to pre-mark the source file to provide the ability to quickly search for the desired objects when forming random subsamples.

The markup involves dividing the source file into a large number of fragments, independently determining the number of objects of each class in each fragment, and then combining the corresponding information. The result of the markup is a single array that stores the ranges of the object numbers of each class belonging to individual file fragments.

At the same time, instead of the traditional operations of reading a file on a disk, it is proposed to use faster operations of mapping a file into the RAM.

At the second stage, information on the object's numbers ranges is used to form random subsamples. In this case, subsamples are formed directly in the training process as necessary, which provides significant savings of the RAM.

Due to the independence of most operations, the markup stage allows a fairly efficient parallel implementation.

The forming subsamples stage also has a high degree of data parallelism, since all subsamples are formed independently of each other.

In addition, the proposed approach does not requere a lot of memory at the both stages, due to small enough the markup size and since there is no need to load all subsamples at the same time. More over, the amount of RAM used can be additionally reduced by reducing the number of file fragments and reducing the subsamples size.

Thus, we can consider that the proposed approach has no restrictions on the training set size.

The only drawback of this approach is the need to search and read objects in the traning process. However, this is offset by the possibility of parallel processing.

Experiments on a big data set from the libsvm repository show that the proposed approach has a quasilinear acceleration of the training stage and, together with the proposed data strategy, can improve computing performance compared to traditional ways of working with data in the same format.

The reported study was funded by RFBR according to the research projects 17-07-00436, 18-07-00942, 18-07-01087.

[1] *Makarova A. I., Sulimova V. V.* Fast approximate decision of two-class SVM problem for big training sets // Proceedings of the V interantional conference on Information technologies and nanotechnologies. (ITNT-2019), Vol. 4. Data Sciences – Samara: Novaya tekhnika, 2019, pp. 25–34.

О теоретико-информационной нижней границе вероятности ошибки классификации

Ланге Михаил Михайлович[1]⋆ lange_mm@ccas.ru
Ганебных Сергей Николаевич[1] sng@ccas.ru
Ланге Андрей Михайлович[1] lange_am@main.ru
[1]Москва, Федеральный исследовательский центр «Информатика и управление» РАН

В теории кодирования источников с допустимой погрешностью Шенноном введена функция «скорость – погрешность», которая базируется на вероятностной модели кодирования. Эта функция дает нижнюю границу скорости кодирования при фиксированной погрешности, либо нижнюю границу погрешности при фиксированной скорости кода Для детерминированной модели аналогичная функция известна как ε-энтропия Функция «скорость – погрешность» не зависит от выбранного алгоритма кодирования. Поэтому эффективность любого конкретного кода может быть оценена избыточностью его скорости или погрешности от соответствующих нижних значений, определяемых функцией «скорость – погрешность».

Следуя функции «скорость – погрешность» для кодирования источников, вводится аналогичная функция «взаимная информация – вероятность ошибки» , которая дает нижнюю границу вероятности ошибки классификации на заданном множестве данных для любого решающего алгоритма с известным набором разделяющих функций.

Пусть $\Omega = \{\omega_i, i = 1, \ldots, c\}$, $c \geqslant 2$, – множество меток классов с априорными вероятностями $P(\omega_i)$, $i = 1, \ldots, c$; $\mathbf{X}$ – множество объектов с известными условными по классам распределениями вероятностей $\{P(\mathbf{x}|\omega_i), \mathbf{x} \in \mathbf{X}\}$ для $i = 1, \ldots, c$; $\widehat{\Omega} = \{\omega_j, j = 1, \ldots, c\}$ – множество решений по классам с неизвестными условными распределениями $\{Q(\omega_j|\mathbf{x}), j = 1, \ldots, c\}$ для каждого $\mathbf{x} \in \mathbf{X}$. Используя введенные распределения, вероятность ошибки классификации задается средней погрешностью решений, измеряемой в метрике Хемминга на $\Omega \times \widehat{\Omega}$. Функция «взаимная информация – вероятность ошибки» определяется минимумом средней взаимной информацией $I(\mathbf{X}; \widehat{\Omega})$ между $\mathbf{X}$ и $\widehat{\Omega}$ по всевозможным условным распределениям $\{Q(\omega_j|\mathbf{x}), j = 1, \ldots, c\}$ для всех $\mathbf{x} \in \mathbf{X}$ при ε ограничении сверху на вероятность ошибки. Указанная функция аналогична функции «скорость – погрешность» для кодирования источника при наличии шума [*Добрушин Р.Л., Цыбаков Б.С.* Передача информации с дополнительным шумом, 1963].

Исследуются нижние границы функции «взаимная информация – вероятность ошибки», построенные на множествах данных от различных источников и на ансамбле данных от источников различной модальности [1]. Для ансамбля $\mathbf{X}_1 \ldots \mathbf{X}_M$ размера $M \geqslant 2$ полученная граница имеет вид

$$R_L(\varepsilon) = I(\mathbf{X}_1 \ldots \mathbf{X}_M; \Omega) - h(\varepsilon - \varepsilon_{\min}) - (\varepsilon - \varepsilon_{\min}) \ln(c - 1), \qquad (1)$$

где $\varepsilon_{\min} \leqslant \varepsilon \leqslant \varepsilon_{\max}$, $h(z) = -z\ln(z) - (1-z)\ln(1-z)$, $0 \leqslant z \leqslant 1$, $R_L(\varepsilon_{\min}) = I(\mathbf{X}_1, \ldots, \mathbf{X}_M; \Omega)$, $R_L(\varepsilon_{\max}) = 0$; $I(\mathbf{X}_1 \ldots \mathbf{X}_M; \Omega) = H(\Omega) - H(\Omega|\mathbf{X}_1 \ldots \mathbf{X}_M)$ – средняя взаимная информация между $\mathbf{X}_1 \ldots \mathbf{X}_M$ и Ω; $H(\Omega)$ и $H(\Omega|\mathbf{X}_1 \ldots \mathbf{X}_M)$ – энтропия и условная энтропия на множестве Ω. Показано, что $\varepsilon_{\min} \geqslant 0$ определяется величиной $H(\Omega|\mathbf{X}_1 \ldots \mathbf{X}_M) \geqslant 0$. При равномерном априорном распределении классов $\varepsilon_{\max} = (c-1)/c$. Для заданного множества объектов $\mathbf{X}$ граница $R_L(\varepsilon)$ сохраняет форму (1) с заменой $\mathbf{X}_1 \ldots \mathbf{X}_M$ на $\mathbf{X}$.

Граница (1) является обобщением нижней границы Шеннона для функции «скорость – погрешность» в случае источника независимых символов из алфавита $\mathbf{X}$ размера $c \geqslant 2$ и хемминговой метрики погрешности. В границе Шеннона $H(\Omega|\mathbf{X}) = 0$ и, следовательно, $\varepsilon_{\min} = 0$. В общем случае, при заданном ансамбле $\mathbf{X}_1 \ldots \mathbf{X}_M$, равенство $I(\mathbf{X}_1 \ldots \mathbf{X}_M; \widehat{\Omega}) = R_L(\varepsilon)$ дает нижнюю границу ε для вероятности ошибки любого решающего алгоритма с разделяющими функциями, обеспечивающими среднюю взаимную информацию $I(\mathbf{X}_1 \ldots \mathbf{X}_M; \widehat{\Omega}) \leqslant I(\mathbf{X}_1 \ldots \mathbf{X}_M; \Omega)$.

Численные реализации границы (1) получены для множеств полутоновых изображений лиц и подписей, а также для ансамбля этих данных. Множества данных содержат объекты (образы) от 25 персон (классов) по 40 объектов в каждом классе. Условные по классам распределения вероятностей образов построены с использованием метрики на соответствующих множествах представлений. Функции $R_L(\varepsilon)$, вычисленные для заданных множеств лиц, подписей и ансамбля этих множеств, представлены кривыми на рис. 1 и демонстрируют существенное уменьшение вероятности ошибки $\varepsilon_{\min}$ на ансамбле по сравнению с аналогичными значениями для множества подписей и множества лиц.

Используя разделяющие функции различных решающих алгоритмов, планируется построить условные распределения решений с варьируемым параметром. Распределение объектов на множестве $\mathbf{X}$ совместно с условными распределениями решений на множестве $\widehat{\Omega}$ позволит найти зависимости средней взаимной информации $I(\mathbf{X}; \widetilde{\Omega})$ от вероятности ошибки исследуемых алгоритмов. Сравнение таких зависимостей с нижней границей $R_L(\varepsilon)$ позволит оценить избыточность вероятностей ошибки алгоритмов.

Работа частично поддержана грантами РФФИ № 18-07-01231 и № 18-07-01385.

[1] *Ланге М.М.* О сравнительной эффективности схем классификации данных от ансамбля источников с использованием средней взаимной информации // Информатика и ее применения, 2019, 13(4) (в печати).

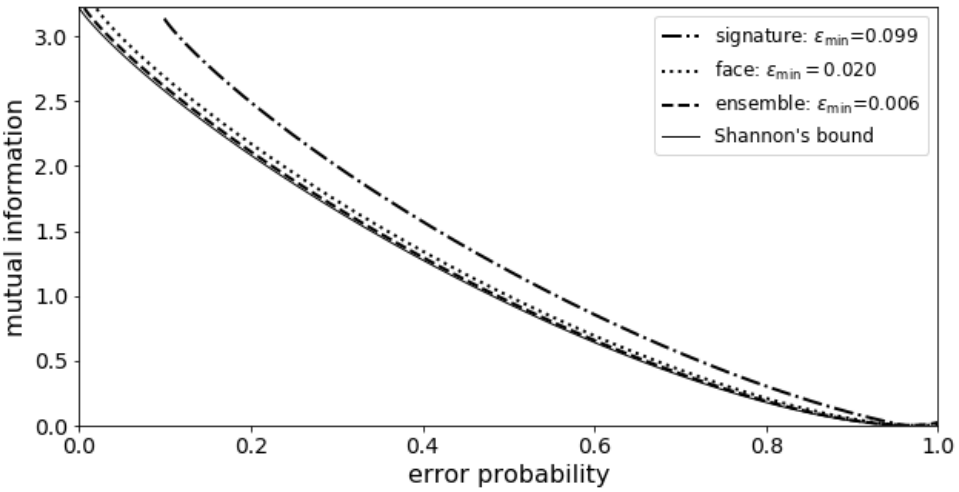


Рис. 1. Функции $R_L(\varepsilon)$ для множеств биометрических объектов.

On an information-theoretical lower bound to a classification error probability

Mikhail Lange[1]⋆ lange_mm@ccas.ru
Sergey Ganebnykh[1] sng@ccas.ru
Andrey Lange[1] lange_am@mail.ru
[1]Moscow, Federal Reseach Center "Computer Science and Control" of RAS

In the theory of source coding with a given fidelity, Shannon introduced the rate distortion function based on a probabilistic coding model. This function yields the lower bound to a code rate subject to a fixed distortion value or the lower bound to a distortion value subject to a fixed code rate In determinate coding model, the similar function is known as the ε-entropy The rate distortion function is independent on a coding algorithm. So, an efficiency of any coding algorithm can be evaluated by a redundancy of the corresponding code rate or the distortion value relative to the appropriate lower values satisfying the rate distortion function.

Following the rate distortion function, we introduce the similar "mutual information – error probability" function that yields a lower bound to a classification error probability in a given dataset for any decision algorithm with an appropriate collection of the discriminant functions.

Let $\Omega = \{\omega_i, i = 1, \ldots, c\}$, $c \geqslant 2$, be a set of the class labels of the prior probabilities $P(\omega_i)$, $i = 1, \ldots, c$; $\mathbf{X}$ be a set of the objects with known class-conditional probability distributions $\{P(\mathbf{x}|\omega_i), \mathbf{x} \in \mathbf{X}\}$ for each $i = 1, \ldots, c$; and $\widehat{\Omega} = \{\omega_j, j = 1, \ldots, c\}$ be a set of the class-label decisions with unknown conditional probability distributions $\{Q(\omega_j|\mathbf{x}), j = 1, \ldots, c\}$ for each $\mathbf{x} \in \mathbf{X}$. Using the above distributions, the classification error probability is given as the average distortion of the class-label decisions by the Hamming metric in $\Omega \times \widehat{\Omega}$. Like the rate distortion function for source coding with an additional noise [*Dobrushin R., Tsybakov B.* Information transmission with additional noise, 1963], for object classification in a given set $\mathbf{X}$, we define the "mutual information – error probability" function $R(\varepsilon)$ as the minimum of the average mutual information $I(\mathbf{X}; \widehat{\Omega})$ between $\mathbf{X}$ and $\widehat{\Omega}$ that is taken over the distributions $\{Q(\omega_j|\mathbf{x}), j = 1, \ldots, c\}$ for all $\mathbf{x} \in \mathbf{X}$ subject to ε error probability constraint from above.

For both the datasets of the individual sources and an ensemble of the different modality sources, the lower bounds to the function $R(\varepsilon)$ have been investigated [1]. Given ensemble $\mathbf{X}_1 \ldots \mathbf{X}_M$ of size $M \geqslant 2$, we obtained the lower bound as follows

$$R_L(\varepsilon) = I(\mathbf{X}_1 \ldots \mathbf{X}_M; \Omega) - h(\varepsilon - \varepsilon_{\min}) - (\varepsilon - \varepsilon_{\min}) \ln(c - 1). \qquad (1)$$

Here, $\varepsilon_{\min} \leqslant \varepsilon \leqslant \varepsilon_{\max}$, $h(z) = -z \ln(z) - (1 - z) \ln(1 - z)$, $0 \leqslant z \leqslant 1$, $R_L(\varepsilon_{\min}) = I(\mathbf{X}_1 \ldots \mathbf{X}_M; \Omega)$, $R_L(\varepsilon_{\max}) = 0$, $I(\mathbf{X}_1 \ldots \mathbf{X}_M; \Omega) = H(\Omega) - H(\Omega|\mathbf{X}_1 \ldots \mathbf{X}_M)$ is the average mutual information between $\mathbf{X}_1 \ldots \mathbf{X}_M$ and Ω; $H(\Omega)$ and $H(\Omega|\mathbf{X}_1 \ldots \mathbf{X}_M)$ are the entropy and the conditional entropy of the set Ω. It has been shown

that $\varepsilon_{\min} \geqslant 0$ is defined by $H(\Omega|\mathbf{X}_1 \ldots \mathbf{X}_M) \geqslant 0$. In case of the uniform prior distribution of the classes, $\varepsilon_{\max} = (c-1)/c$. For a source dataset $\mathbf{X}$, the bound $R_L(\varepsilon)$ retains the form (1) with substituting $\mathbf{X}$ for $\mathbf{X}_1 \ldots \mathbf{X}_M$.

The bound (1) is the generalization of the Shannon's lower bound to the rate distortion function for the source independent symbols belonging to the alphabet $\mathbf{X}$ of size $c \geqslant 2$. In the Shannon's bound, $H(\Omega|\mathbf{X}) = 0$ and therefore $\varepsilon_{\min} = 0$. Generally, for the ensemble $\mathbf{X}_1 \ldots \mathbf{X}_M$, the equality $I(\mathbf{X}_1 \ldots \mathbf{X}_M; \widehat{\Omega}) = R_L(\varepsilon)$ yields the lower bound ε to a classification error probability of any decision algorithm whose discriminant functions provide the average mutual information $I(\mathbf{X}_1 \ldots \mathbf{X}_M; \widehat{\Omega}) \leqslant I(\mathbf{X}_1 \ldots \mathbf{X}_M; \Omega)$.

The bounds of the form (1) have been calculated for the datasets of grayscale faces and signatures as well as for the ensemble of these sources. The source datasets contain the patterns taken from 25 persons (classes) by 40 patterns per each class. The lower bounds have been calculated in the space of the tree-structured pattern representations. The corresponding class-conditional distributions are constructed using the appropriate metrics in the datasets of the representations. The numerical bounds are shown in Fig. 1 and the curves demonstrate essential decreasing the minimal error probability $\varepsilon_{\min}$ in the ensemble as compared with the similar values in the datasets of the individual sources.

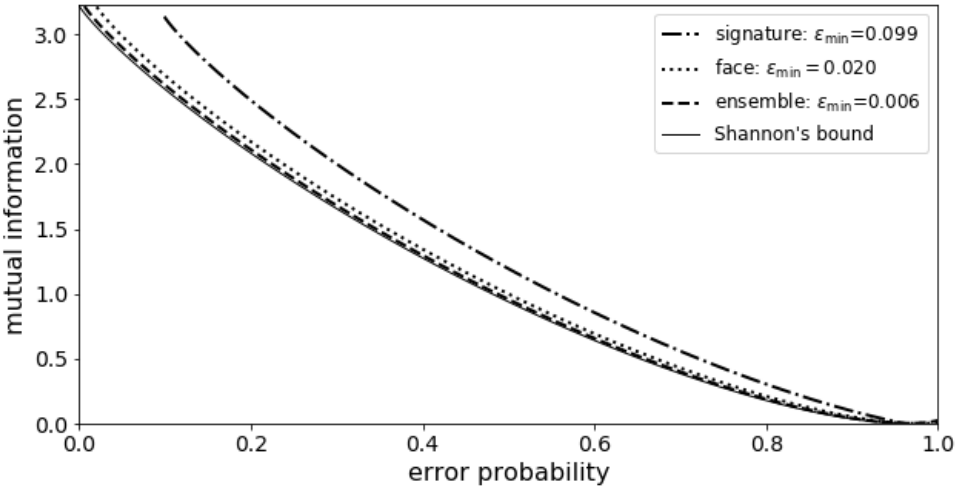


Fig. 1. The functions $R_L(\varepsilon)$ for the biometrical datasets.

Using the discriminant functions for the different decision algorithms, we plan to construct the conditional distributions of the class-label decisions with a variable parameter. Then, the probability distribution of the objects in $\mathbf{X}$ together with

the conditional distributions of the decisions in $\widehat{\Omega}$ will allow us to calculate the average mutual information $I(\mathbf{X};\widetilde{\Omega})$ as the function of the error probability for the investigated algorithms. A comparison of these functions with the corresponding lower bound $R_L(\varepsilon)$ will yield the redundancy of the algorithm error probabilities relative to the potentially possible values. Also, we plan to perform the similar experiments in the ensembles of the biometrical datasets.

The research is partially supported by RFBR, grants 18-07-01231 and 18-07-01385.

[1] *Lange M. M.* On a comparative efficiency of classification schemes in an ensemble of data sources using the average mutual information // Informatics and applications, 2019, 13(4) (in press).

Метод средних решающих правил для быстрого двухклассового обучения в пространстве, порожденном потенциальной функцией

Макарова Александра Игоревна[1]⋆ aleksarova@gmail.com
Сулимова Валентина Вячеславовна[1] vsulimova@yandex.ru
[1]Тула, Тульский государственный университет

Задача двухклассового обучения распознаванию объектов той или иной природы является одной из широко распространенных задач анализа данных, возникающей при решении многих прикладных задач.

В данной работе мы ориентируемся на применение метода опорных векторов — удобного и хорошо зарекомендовавшего себя подхода к решению задачи двухклассового распознавания, адаптированного для работы в пространстве, порожденном потенциальной функцией (Kernel-based Support Vector Machines, Kernel-based SVM), что позволяет строить нелинейные границы, разделяющие объекты пары классов, повышая качество распознавания.

Следует отметить, что особенностью многих современных задач анализа данных является большой размер обучающей совокупности, который может стать серьезным препятствием для применения Kernel-based SVM ввиду высокой вычислительной сложности этапа обучения.

Повышению производительности решения данной задачи посвящена целая серия работ [1, 2, 3, 4]. Однако, несмотря на это, исследования в данной области до сих пор являются актуальными.

В предыдущей работе [5] нами был предложен метод быстрого приближенного решения задачи SVM в линейном признаковом пространстве, названный методом средних решающих правил (Mean Decision Rules, MDR). В данной работе мы адаптируем данный подход для обучения в пространстве, порожденном потенциальной функцией.

Основная идея предложенного метода заключается в усреднении решающих правил SVM, построенных по небольшим случайным подвыборкам исходного обучающего множества объектов.

При увеличении числа случайных подвыборок усредненное решающее правило стабилизируется и перестает вести себя как случайная величина. На рисунке 1 представлены примеры решающих правил, найденных при помощи библиотеки libSVM по полной обучающей совокупности и решающих правил, найденных при помощи метода средних решающих правил для нескольких наборов модельных данных.

Особенностью, связанной с обучением в пространстве потенциальной функции, является невозможность (в общем случае) явного вычисления и хранения направляющего вектора оптимальной разделяющей гиперплоскости. В результате оказывается невозможным и его непосредственное усреднение, в отличие от случая обучения в линейном признаковом пространстве.

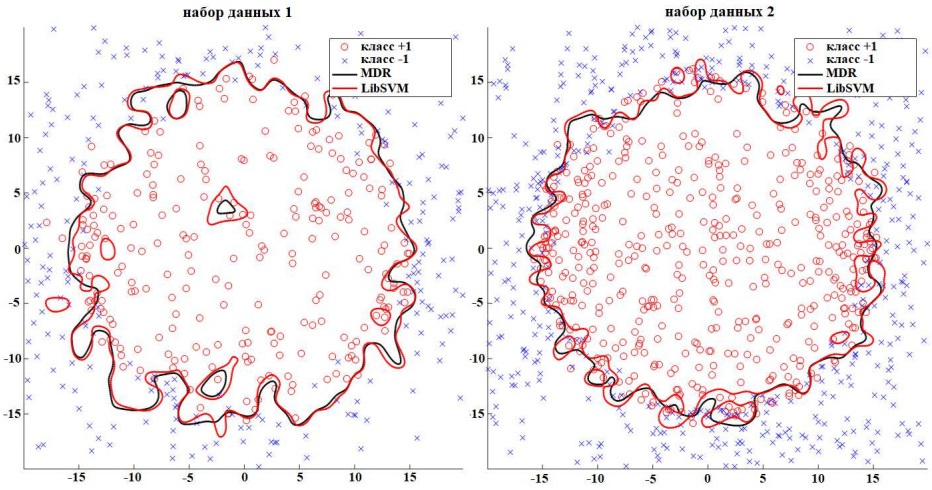


Рис. 1. Решающие правила, полученные MDR и библиотекой LibSVM

Учет указанной особенности требует использования большего количества оперативной памяти и большего объема вычислений по сравнению с признаковым случаем. Однако, несмотря на это, предложенный подход, по-прежнему, не имеет теоретического ограничения на размер обучающей выборки, поскольку не требует одновременного нахождения в памяти всех объектов, что обеспечивает возможность его применения для обучения даже на одной вычислительной машине. Кроме того, нетрудно увидеть, что метод средних решающих обладает высокой степенью параллелизма по данным, что обеспечивает возможность его эффективной реализации с применением технологий параллельных и распределенных вычислений.

Экспериментальные исследования на модельных и реальных данных показали, что, как и в случае обучения в линейном пространстве, предлагаемый подход позволяет достаточно быстро найти приближенное, но не сильно отличающееся от точного решение задачи SVM в пространстве, порожденном потенциальной функцией.

Кроме того, следует отметить, что в ряде случаев решение, найденное при помощи MDR, позволяет получить меньший процент ошибок на контроле по сравнению с результатами, полученными при помощи точных методов решения задачи SVM, в частности, с использованием библиотеки LibSVM.

Исследование выполнено при финансовой поддержке РФФИ в рамках научных проектов № 18-07-01087, № 18-07-00942, № 17-07-00436.

[1] *Zhao H. X., Magoules F.* Parallel support vector machines on multi-core and multiprocessor systems // 11th International Conference on Artificial Intelligence and Applications (AIA 2011), IASTED, 2011.

[2] *Agrwal A. et al.* A reliable effective terascale linear learning system // The Journal of Machine Learning Research. 2014—T. 15, No.1. — P. 1111-1133.

[3] *Joachims T.* Training linear SVMs in linear time // Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2006 — P. 217-226.

[4] *Chu C. T. et al.* Map-reduce for machine learning on multicore // Advances in neural information processing systems. 2007 — P. 281-288.

[5] *Макарова А. И., Сулимова В. В.* Быстрое приближенное решение двухклассовой задачи SVM для больших обучающих совокупностей // СБОРНИК ТРУДОВ ИТНТ-2019, 2019 — С. 25-34.

Method of mean decision rules for fast two-class learning in the space generated by a potential function

Alexandra Makarova[1]⋆ aleksarova@gmail.com
Valentina Sulimova[1] vsulimova@yandex.ru
[1]Tula, Russia, TulSu

The two-class pattern recognition problem is one of the widespread problems of data analysis that arises in solving many applied problems.

This paper focuses on the application of the Support Vector Machines to solving the two-class recognition problem in a linear space, produced by a potential function (Kernel-based SVM).

Kernel-based SVM is a convenient and well-proven approach, which allows to obtain non-linear decisions in a linear space, increasing the quality of recognition.

It should be noted that a feature of many modern data analysis tasks is the large size of the training set. It can become a serious obstacle to the use of Kernel-based SVM due to the high computational complexity of the training.

A series of works is devoted to increasing the productivity of SVM problem solving [1, 2, 3, 4]. However, despite this, research in this area is still relevant.

In the previous work [5], we proposed a simple approach to finding a fast approximate solution of the SVM problem in a linear feature space, named the Mean Decision Rules (MDR) method. In this paper, we adapt this approach for training in a space of a potential function.

The main idea of the proposed method is to average SVM decision rules, which are constructed for small random subsamples of the initial training set.

The mean decision rule stabilizes and ceases to behave as a random variable with an increase in the number of random subsamples.

Figure 1 shows examples of decision rules found using the state of the art LibSVM library and the proposed MDR method for several model data sets. As we can see from the figure, the obtained decisions are close enough.

A feature associated with the training in a potential function space is the impossibility (in the general case) of explicitly calculating the directional element of the optimal separating hyperplane. As a result, its direct averaging is impossible, in contrast to the case of training in a linear feature space.

To overcome this obstacle, it is required to use a larger amount of RAM and a greater amount of computation in comparison with training in the linear feature space. However, despite this, the proposed approach has no theoretical limit on the training sample size. This is due to that it does not require simultaneous storage in the memory of all objects and so it can be used for training even on one computer.

In addition, it is easy to see that the MDR method has a high degree of data parallelism and can be effectively implemented using parallel and distributed computing technologies.

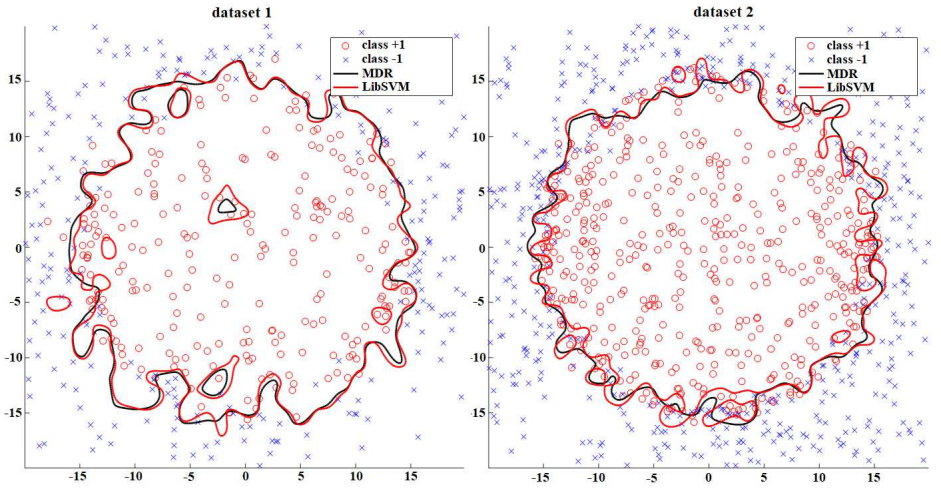


Fig. 1. Decision rules founded by MDR and LibSVM

Experimental research have shown that as in the case of training in a linear feature space, this approach allows to quickly find an approximate, but not very different from the exact solution of the SVM problem in the space generated by a potential function.

Moreover, in a number of cases it allows to reach a lower error rate in the test set in contrast to results obtained using the state of the art LibSVM library.

The reported study was funded by RFBR according to the research projects 18-07-01087, 18-07-00942,17-07-00436.

[1] *Zhao H. X., Magoules F.* Parallel support vector machines on multi-core and multiprocessor systems // 11th International Conference on Artificial Intelligence and Applications (AIA 2011), IASTED, 2011.

[2] *Agrwal A. et al* A reliable effective terascale linear learning system // The Journal of Machine Learning Research. 2014 T.15. — P. 1111-1133.

[3] *Joachims T.* Training linear SVMs in linear time // Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2006 — P. 217-226.

[4] *Chu C. T. et al* Map-reduce for machine learning on multicore // Advances in neural information processing systems. 2007 — P. 281-288.

[5] *Makarova A. I., Sulimova V. V.* Fast approximate two-class SVM learning for large training sets // ITNT-2019, 2019 — P. 25–34.

Определение сложности выборки с помощью универсальной аппроксимирующей модели

Малиновский Григорий Станиславович[1]⋆ grigoriy.malinovskiy@phystech.edu
Гадаев Тамаз Тезикоевич[1] gadaev.tt@phystech.edu
Стрижов Вадим Викторович[1,2] strijov@phystech.edu
[1]Москва, Московский физико-технический институт
[2]Москва, Вычислительный центр им. А.А. Дороницына ФИЦ ИУ РАН

Исследуются свойства обучающей выборки в задачах классификации и регресии. Сложность обучающей выборки может быть определена с помощью различных критериев таких как принцип минимальной длины описания и статистическая сложность. Предлагается вычислять сложность с помощью аппроксимирующих моделей. Данный подход является эвристическим. В данной работе предложен метод определения сложности обучающей выборки с помощью двухслойной полносвязной нейронной сети. Согласно теореме Цыбенко нейронная сеть прямой связи с одним скрытым слоем аппроксимирует любую непрерывную функцию многих переменных с заданной точностью. Для определения сложности используется количество нейронов внутреннего слоя нейронной сети.

В работе исследованы свойства этого метода. Для анализа метода поставлен вычислительный эксперимент на синтетических данных с различными конфигурациями мультиколинеарных признаков и уровнем шума. Также были проведены эксперементы на датасетах Boston Housing и Cleveland Heart Disease для задач регресии и бинарной классификации.

Работа поддержана РФФИ, проекты No 17-20-01212, 19-07-0885.

[1] *Aduenko A. A., Motrenko A. P, Strijov V. V.* Object selection in credit scoring using covariance matrix of parameters estimations // Annals of Operations Research, 2018, 260(1-2) : 3-21.

Determination of data complexity using a universal approximating model

Grigory Malinovsky[1]⋆ grigoriy.malinovskiy@phystech.edu
Tamaz Gadaev[1] gadaev.tt@phystech.edu
Vadim Strijov[1,2] strijov@gmail.com
[1]Moscow, Moscow Institute of Physics and Technology
[2]Moscow, FRCCSC of the Russian Academy of Sciences

The properties of the training dataset in classification and regression problems are investigated. The complexity of the training dataset can be determined using various criteria such as the principle of minimum description length and statistical complexity. It calculates the complexity using approximating models. This is a heuristic approach. In this paper, we propose a method for determining the complexity of the training data using a two-layer fully connected neural network. According to Tsybenko's theorem, a neural network of direct communication with one hidden layer approximates any continuous function of many variables with a given accuracy. The number of neurons in the inner layer of the neural network is used to determine the complexity.

The properties of this method are investigated. To analyze the method, a computational experiment is performed on synthetic data with different configurations of multicolinear features and noise level. Experiments were also conducted on Boston Housing and Cleveland Heart Disease datasets for regression and binary classification problems.

This research is funded by RFBR, projects 17-20-01212, 19-07-0885.

[1] *Aduenko A. A., Motrenko A. P, Strijov V. V.* Object selection in credit scoring using covariance matrix of parameters estimations // Annals of Operations Research, 2018, 260(1-2) : 3-21.

Сравнение двух подходов к разложению критериев качества решающих функций

Неделько Виктор Михайлович[1]⋆ nedelko@math.nsc.ru

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

Проводится сравнительный анализ двух подходов к разложению критерия качества решающих функций. Первый подход – разложение на смещение и разброс (bias-variance decomposition). Второй подход (предложен в статье Г. С. Лбов, Н. Г. Старцева, 1989) – разложение на меру адекватности и меру статистической устойчивости.

Идея второго подхода состоит в том, чтобы разложить ошибку прогноза на погрешность аппроксимации и статистическую погрешность.

Мера адекватности характеризует погрешность аппроксимации и представляет собой разность между асимптотическим средним риском и байесовским. Данная мера показывает, насколько хорошее решение метод мог бы дать в случае неограниченной выборки. Мера статистической устойчивости есть разность между фактическим средним риском и асимптотическим.

В настоящей работе предлагается метод статистического оценивания компонент обоих разложений на реальных данных и проводится сравнение зависимостей этих компонент от сложности решающей функции. В качестве универсальной меры сложности используется ненормированный отступ.

Результаты исследования показывают значительное качественное сходство между поведением смещения и меры адекватности и между разбросом и мерой статистической устойчивости.

Вместе с тем, между рассмотренными разложениями имеется принципиальное различие, в частности, при увеличении сложности мера адекватности не может увеличиваться, в то время как смещение сначала уменьшается, однако при очень больших значениях сложности обычно начинает расти.

Работа поддержана грантом РФФИ № 18-07-00600-а.

[1] *Неделько В. М.* Некоторые вопросы оценивания качества методов построения решающих функций // Вестник Томского государственного университета. Управление, вычислительная техника и информатика, Томск: ТГУ, 2013. № 3 (24) — С. 123–132.

Comparison of two approaches to decomposition of quality criteria of decision functions

Victor Nedel'ko[1]⋆ nedelko@math.nsc.ru

[1]Sobolev Institute of Mathematics

A comparative analysis of two approaches to the decomposition of quality criterion of decision functions is carried out. The first approach is the bias-variation decomposition. The second approach (proposed in the paper G. S. Lbov, N. G. Startseva, 1989) is a decomposition into a measure of adequacy and a measure of statistical stability.

The idea of the second approach is to decompose the prediction error into approximation error and statistical error.

The adequacy measure characterizes the approximation error. It is the difference between the asymptotic mean risk and the Bayesian mean risk. This measure shows how good a solution the method could give in the case of unlimited sample. A measure of statistical robustness is the difference between the actual average risk and the asymptotic one.

In this paper we propose a method of statistical estimation of the components of both decompositions on real data and compare the dependencies of these components on the complexity of the decision function. Non-normalized margin is used as a universal measure of complexity.

The results of the study show significant qualitative similarities in behavior of the bias and the adequacy measure and between the variance and the statistical stability measure.

At the same time, there is a fundamental difference between the considered decompositions, in particular, with increasing complexity, the measure of adequacy cannot increase, while the bias first decreases, but at very high values of complexity usually begins to grow.

This research is funded by RFBR, grant 18-07-00600-a.

[1] *Nedel'ko V.* On estimation of quality of decision functions construction methods // Tomsk State University Journal of Control and Computer Science, 2013. No. 3(24) — p. 123–132. (In Russian).

Машинное обучение на основе анализа выпуклых оболочек классов

Немирко Анатолий Павлович[1]⋆ apn-bs@yandex.ru
[1]Санкт-Петербург, СПбГЭТУ "ЛЭТИ"

В работе рассмотрены методы машинного обучения при использовании аппарата вычислительной геометрии. Применение выпуклых оболочек множеств в многомерном признаковом пространстве позволило реализовать алгоритмы визуализации области пересечения классов и алгоритмы многоклассового распознавания.

С помощью методов вычислительной геометрии дан пример отображения на плоскость области пересечения двух выпуклых оболочек, описывающих классы в многомерном признаковом пространстве. Предложен способ измерения близости выпуклой оболочки к испытуемой точке при расположении точки внутри выпуклой оболочки. Он основан на оценке направленной глубины проникновения данной точки в рассматриваемую выпуклую оболочку.

Задача решается при проектировании данной точки и вершин выпуклой оболочки на вектор направления от данной точки к центроиду класса. Описан алгоритм классификации ближайшей выпуклой оболочки, основанный на предложенном упрощенном способе оценки близости к выпуклой оболочке [1]. Проведено сравнение классификаторов ближайшей выпуклой оболочки и классификаторов ближайших соседей (kNN). Приведены результаты экспериментальных исследований на синтезированных числовых данных и на реальных данных задачи диагностики рака груди.

Работа поддержана грантами РФФИ № 19-29-01009, 18-07-00264.

[1] *Nemirko A. P.* Lightweight nearest convex hull classifier // Pattern Recogn. Image Anal., 2019. vol. 29, No 2 — P. 360–365.

Machine learning based on the analysis of convex hulls of classes

Anatoliy Nemirko[1]⋆ apn-bs@yandex.ru
[1]Saint Petersburg, ETU "LETI"

The paper considers the methods of machine learning using the apparatus of computational geometry. The use of convex hulls of sets in a multidimensional feature space made it possible to implement visualization algorithms for the class intersection area and algorithms for multiclass recognition.

Using methods of computational geometry, an example is given of mapping onto the plane the region of intersection of two convex hulls that describe classes in a multidimensional feature space. A method is proposed for measuring the proximity of a convex hull to a test point when the point is located inside the convex hull. It is based on the estimation of the directional penetration depth of a given point into the considered convex hull.

The problem is solved when designing a given point and vertices of a convex hull on a direction vector from a given point to the centroid of the class. The classification algorithm for the nearest convex hull is described, based on the proposed simplified method for assessing the proximity to a convex hull [1]. The classifiers of the nearest convex hull and the classifiers of nearest neighbors (kNN) are compared. The results of experimental studies on synthesized numerical data and on real data of the problem of diagnosing breast cancer are presented.

This research is funded by RFBR, grants 19-29-01009, 18-07-00264.

[1] *Nemirko A.* Lightweight nearest convex hull classifier // Pattern Recogn. Image Anal., 2019. 29(2) — p. 360–365.

Методы машинного обучения на основе минимизации сглаженных оценок средних, нечувствительных к выбросам

Шибзухов Заур Мухадинович[1,2]⋆ intellimath@mail.ru

[1]Москва, Институт математики и информатики МПГУ

[2]Нальчик, Институт прикладной математики и автоматизации КБНЦ РАН

Многие задачи машинного обучения обычно сводятся к задаче минимизации среднего арифметического от конечного набора параметризованных функций:

$$\mathcal{Q}(\mathbf{w}) = \tfrac{1}{N}\sum_{k=1}^{N}\ell_k(\mathbf{w}).$$

Однако, если эмпирическое распределение $\{\ell_1(\mathbf{w}),\dots,\ell_N(\mathbf{w})\}$ содержит значительный объем выбросов, то минимизация $\mathcal{Q}(\mathbf{w})$ приведет к смещению $\mathbf{w}^*$ из-за чувствительности среднего арифметического к выбросам. Одно из решений этой проблемы основано на использовании робастных дифференцируемых оценок среднего значения.

Большинство известных оценок среднего можно представить как M-средние [1,2]:

$$\bar{z}_\rho = \mathsf{M}_\rho\{z_1,\dots,z_N\} = \arg\min_u \sum_{k=1}^{N}\rho(z_k-u),$$

где $\rho(r)$ – выпуклая функция. Если ρ дважды дифференцируема, то M_ρ имеет частные производные:

$$\frac{\partial \mathsf{M}_\rho}{\partial z_k} = \frac{\rho''(z_k-\bar{z}_\rho)}{\rho''(z_1-\bar{z}_\rho)+\dots+\rho''(z_N-\bar{z}_\rho)}.$$

Например, при $\rho(r)=\sqrt{\varepsilon^2+r^2}-\varepsilon$, M-среднее представляет собой сглаженный вариант медианы.

Более робастный вариант среднего арифметического можно получить, используя сглаженный вариант винзоризированного среднего:

$$\mathsf{WM}_\rho\{z_1,\dots,z_N\} = \frac{1}{N}\sum_{z_k\leqslant\bar{z}_\rho}^{N} z_k + \frac{m}{N}\bar{z}_\rho,$$

где m – количество $z_k > \bar{z}_\rho$. Его частные производные имеют вид:

$$\frac{\partial \mathsf{WM}_\rho}{\partial z_k} = \begin{cases} \frac{1}{N}+\frac{m}{N}\frac{\partial \mathsf{M}_\rho}{\partial z_k}, & \text{если } z_k \leqslant \bar{z}_\rho \\ \frac{m}{N}\frac{\partial \mathsf{M}_\rho}{\partial z_k}, & \text{если } z_k > \bar{z}_\rho \end{cases}$$

Потенциально несмещенные оценки $\mathbf{w}^*$ ищутся путем минимизации дифференцируемой робастной оценки среднего от параметризованных функций:

$$\mathbf{w}^* \in \arg\min_{\mathbf{w}} \mathsf{M}\{\ell_1(\mathbf{w}),\dots,\ell_N(\mathbf{w})\}.$$

Такая постановка задачи открывает возможность для преодоления проблемы выбросов, а также для их идентификации.

Вычисление $\mathbf{w}^*$ осуществляется на базе алгоритмической схемы итеративного перевзвешивания. В ней строится сгущающаяся последовательность $\{\mathbf{w}_t\}$, где

$$\mathbf{w}_{t+1} \in \arg\min_{\mathbf{w}} \sum_{k=1}^{N} v_{t,k}\ell_k(\mathbf{w}),$$

а $v_{t,k} = \nabla\mathsf{M}\{\ell_1(\mathbf{w}_t), \ldots, \ell_N(\mathbf{w}_t)\}$.

На реальных примерах показывается нечувствительность предлагаемого подхода и алгоритмов к большому количеству выбросов (вплоть до 50%) при решении задач регрессии, классификации и кластеризации.

Работа выполнена при финансовой поддержке гранта РФФИ №18-01-00050.

[1] Shibzukhov Z.M. On the Principle of Empirical Rsik Minimization Based on Averagging Functions // Doklady Mathematics. 2017. Vol. 96, N. 2, PP. 494-497.

[2] Shibzukhov Z.M. Robust Neural Networks Learning: New Approaches. – In: Advanves in Neural Networks – ISNN 2018. Lecture Notes in Computer Sciences. Vol. 10878. 2018. PP. 247-255.

[3] *Shibzukhov Z.M., Kazakov M.A.* Clustering based on the principle of finding centers and robust averaging aggregation functions // Proceedings of V International Conference ITNT 2019. Journal of Physics: Conference Series. (в печати).

Machine learning based on minimizing smoothed estimates of averages, resistant to outliers

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

[1]Moscow, Institute Mathematics and Computer Sciences MPSU

[2]Nalchick, Institute of Applied Mathematics and Automation KBSC RAS

Many machine learning problems usually come down to the problem of minimizing the arithmetic mean of a finite set of parameterized functions:

$$\mathcal{Q}(\mathbf{w}) = \tfrac{1}{N}\sum_{k=1}^{N} \ell_k(\mathbf{w}).$$

However, if the empirical distribution $\{\ell_1(\mathbf{w}),\dots,\ell_N(\mathbf{w})\}$ contains a significant amount of outliers, then minimizing $\mathcal{Q}(\mathbf{w})$ will lead to a bias $\mathbf{w}^*$ due to the sensitivity of the arithmetic mean to outliers. One solution to this problem is based on the use of robust differentiable mean estimates.

Most known mean estimates can be represented as M-estimates [1,2]:

$$\bar{z}_\rho = \mathsf{M}_\rho\{z_1,\dots,z_N\} = \arg\min_u \sum_{k=1}^{N} \rho(z_k - u),$$

where $\rho(r)$ – convex function. If ρ is twice differentiable, then M_ρ has partial derivatives:

$$\frac{\partial \mathsf{M}_\rho}{\partial z_k} = \frac{\rho''(z_k-\bar{z}_\rho)}{\rho''(z_1-\bar{z}_\rho)+\dots+\rho''(z_N-\bar{z}_\rho)}.$$

For example, with $\rho(r) = \sqrt{\varepsilon^2+r^2}-\varepsilon$, M-mean is a smoothed version of the median.

A more robust version of the arithmetic mean can be obtained using a smoothed version of the winzorized mean:

$$\mathsf{WM}_\rho\{z_1,\dots,z_N\} = \frac{1}{N}\sum_{z_k\leqslant \bar{z}_\rho}^{N} z_k + \frac{m}{N}\bar{z}_\rho,$$

where m is the number $z_k > \bar{z}_\rho$. Its partial derivatives are of the form:

$$\frac{\partial \mathsf{WM}_\rho}{\partial z_k} = \begin{cases} \frac{1}{N} + \frac{m}{N}\frac{\partial \mathsf{M}_\rho}{\partial z_k}, & \text{if } z_k \leqslant \bar{z}_\rho \\ \frac{m}{N}\frac{\partial \mathsf{M}_\rho}{\partial z_k}, & \text{if } z_k > \bar{z}_\rho \end{cases}$$

Potentially unbiased estimates $\mathbf{w}^*$ are sought by minimizing a differentiable robust estimate of the mean of parameterized functions:

$$\mathbf{w}^* \in \arg\min_{\mathbf{w}} \mathsf{M}\{\ell_1(\mathbf{w}),\dots,\ell_N(\mathbf{w})\}.$$

Such a statement of the problem opens up the opportunity to overcome the problem of emissions, as well as to identify them.

The calculation of $\mathbf{w}^*$ is carried out on the basis of an iterative reweighting algorithmic scheme. It builds a thickening sequence $\mathbf{w}_t$, where

$$\mathbf{w}_{t+1} \in \arg\min_{\mathbf{w}} \sum_{k=1}^{N} v_{t,k}\ell_k(\mathbf{w})$$

and $v_{t,k} = \nabla\mathsf{M}\{\ell_1(\mathbf{w}_t), \ldots, \ell_N(\mathbf{w}_t)\}$.

To calculate $\mathbf{w}^*$, iterative weighting algorithms are used. Real examples show the insensitivity of the proposed approach and algorithms to a large number of outliers (up to 50%) for solving the problems of regression, classification and clustering.

This work was financially supported by the RFBR grant 18-01-00050.

[1] Shibzukhov Z.M. On the Principle of Empirical Rsik Minimization Based on Averagging Functions // Doklady Mathematics. 2017. Vol. 96, N. 2, PP. 494-497.

[2] Shibzukhov Z.M. Robust Neural Networks Learning: New Approaches. – In: Advanves in Neural Networks – ISNN 2018. Lecture Notes in Computer Sciences. Vol. 10878. 2018. PP. 247-255.

[3] *Shibzukhov Z.M., Kazakov M.A.* Clustering based on the principle of finding centers and robust averaging aggregation functions // Proceedings of V International Conference ITNT 2019. Journal of Physics: Conference Series. (in press).

Выбор структуры модели глубокого обучения субоптимальной сложности

Бахтеев Олег Юрьевич[1]⋆ 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 рассматривается задача выбор структуры модели глубокого обучения. Под моделью глубокого обучения понимается суперпозиция дифференцируемых по параметрам функций. Структура модели глубокого обучения задается графом, где ребрам графа соответствуют базовых нелинейные функции, а вершинам — промежуточные представления выборки под действием этих функций. Структурой модели назовем веса базовых функций. Для решения задачи выбора структуры модели вводятся вероятностные предположения о распределениях параметров и структуры модели. Задача выбора структуры модели рассматривается как двухуровневая оптимизация: нижний уровень оптимизации соответствует максимизации вариационной оценки обоснованности модели. Верхний уровень оптимизации соответствует оптимизации гиперпараметров модели. В качестве двухуровневой задачи оптимизации рассматривается обобщенная функция обоснованности модели. Показано, что данная задача соответствует оптимизации согласно ряду критериев: критерию максимального правдоподобия, максимальной апостериорной вероятности, максимальной обоснованности модели, а также позволяет производить оптимизацию параметров и структуры модели с последовательным увеличением и снижением сложности модели, а также с полным перебором структуры модели. Для анализа предлагаемой задачи оптимизации проводится вычислительный эксперимент на синтетических данных и выборке рукописных цифр MNIST.

Работа поддержана РФФИ, проекты No 17-20-01212, 19-07-0875, а также при поддержке Фонда содействия развитию малых форм предприятий в научно-технической сфере (проект 44116).

[1] *Бахтеев О.Ю., Стрижов В. В.* Выбор модели глубокого обучения субоптимальной сложности // Автоматика и телемеханика, Москва, 2018. — No. 8, С. 129–147.

Deep learning structure selection of suboptimal complexity

Oleg Bakhteev[1]⋆ bakhteev@phystech.edu
Vadim Strijov[1,2] strijov@phystech.edu
[1]Moscow, Moscow Institute of Physics and Technology
[2]Moscow, FRCCSC of the Russian Academy of Sciences

The paper is devoted to the problem of the selection of deep learning model structure. A deep learning model is a superposition of differential functions. A structure of a deep learning model is a graph, where the edges of the graph correspond to primitive nonlinear functions, and the vertices correspond to the intermediate representations of the data under the primitives functions. The structure of the model is the vector of weights of the primitive functions. In order to select suboptimal structure we introduce probabilistic assumptions about the distributions of parameters and structure of the model. The problem of the structure of the model is considered as bilevel optimization: the lower level of optimization corresponds to maximizing the evidence lower bound of the model. The upper level of optimization corresponds to the hyperparameter optimization. As a bilevel optimization problem, we consider the generalized model optimization problem. We show that the presented optimization allows to optimize model and its structure in accordance to a number of criteria: the maximum likelihood criterion, the maximum posterior probability criterion, evidence lower bound of the model. It also allows to optimize the deep learning model structure with an increase and decrease of model complexity, as well as with a complete exhaustive search of the model structure. The authors perform computational experiment on a synthetic dataset and on a dataset of handwritten digits MNIST.

This research is funded by RFBR, projects 17-20-01212, 19-07-0875 and FASIE project No.44116.

[1] *O. Yu. Bakhteev and V. V. Strijov* Deep Learning Model Selection of Suboptimal Complexity // Automation and Remote Control, Moscow, 2018. Vol. 79, No. 8, pp. 1474–1488.

Метод дифференциальной поэлементной кросс-валидации для выбора уровня сложности обобщенных линейных моделей зависимостей

Ангуло Бриан Флориан[1]⋆ brian.angulo@yandex.ru
Морозов Алексей Олегович[1] ao.morozov@phystech.edu
Моттль Вадим Вячеславович[2] vmottl@yandex.ru
[1]Москва, Московский физико-технический институт
[2]Москва, Вычислительный центр РАН

При решении задач восстановления зависимостей в классе моделей возрастающей сложности всегда необходимо выбирать значение структурного параметра, определяющего сложность модели. Наиболее популярный подход к выбору уровня сложности модели заключается в использовании принципа кросс-валидации, в частности метода скользящего контроля по отдельным объектам в составе единственной обучающей совокупности, имеющейся у наблюдателя. Недостатком данного метода является слишком высокая вычислительная сложность, связанная с необходимостью многократно повторять обучение, всякий раз удаляя из обучающей совокупности один очередной объект. В данной работе предложен метод дифференциальной беспереборной кросс-валидации для обобщенных линейных моделей произвольных зависимостей [1], позволяющий для каждого пробного значения структурного параметра проводить обучение лишь один раз на всей обучающей совокупности. Идея метода заключается в том, что объекты не удаляются полностью из обучения, а отбрасывается лишь бесконечно малая часть каждого объекта. Показатель качества модели зависимости при выбранном уровне ее сложности использует частные производные ошибок на объектах по весам их вхождения в обучающую совокупность. Процедура вычисления показателя качества модели практически не увеличивает вычислительную сложность обучения. Остается найти численный метод оптимизации качества модели при варьировании структурных параметров.

Работа поддержана грантом РФФИ № 19-37-90159.

[1] *V. Mottl, O. Krasotkina, V. Sulimova, at al.* Linear complexity algorithms for high dimensional SVM and regression problems with smart sparse regularization. Proceedings of the 15th Int. Conf. on Machine Learning and Data Mining MLDM 2019, Vol. I, pp. 412-430.

Method of differential leave-one-out cross validation for chosing the complexity level in generalized linear models of dependences

Brian Angulo[1] ao.morozov@phystech.edu
Alexey Morozov[1] ao.morozov@phystech.edu
Vadim Mottl[1] vmottl@yandex.ru
[1]Moscow Institute of Physics and Technology
[2]Moscow, Computing Center of the Russian Academy of Sciences

When solving dependence estimation problems using models of growing complexity, it is always required to choose the value of a structural parameter that controls the complexity level of the model. The most popular approach to choosing the complexity level of the model is the cross validation principle, in particular, the leave-one-out method, applied, in turn, to each single object in the available training set. The disadvantage of this method is too high computational complexity produced by the necessity to multiply repeat the training process, each time with one object deleted from the training set. In this work, we propose a method of differential non-enumrative leave-one-out cross validation for generalized linear models of arbitrary dependences [1], which requires only one run of the training process on the entire training set. The idea of the method is that only an infinitely small part of each single object is deleted from training instead of omitting the entire object. The proposed indicator of the model quality evaluates the partial derivatives of the loss function at single objects by the weights of their occurring in the training set. The procedure of computing the quality indicator practically does not increase the computational complexity of the training algorithmt. It remains only to choose the numerical method of model quality optimiztion by varying the structural parameters.

This research is funded by RFBR, grant No18-07-01087.

[1] *V. Sulimova, O. Krasotkina, at al.* Linear complexity algorithms for high dimensional SVM and regression problems with smart sparse regularization. Proceedings of the 15th Int. Conf. on Machine Learning and Data Mining MLDM 2019, Vol. I, pp. 412-430.

Последовательное восстановление обобщенных линейных моделей зависимостей по возрастающей обучающей совокупности

Морозов Алексей Олегович[1]⋆ ao.morozov@phystech.edu
Моттль Вадим Вячеславович[2] vmottl@yandex.ru
Сулимова Валентина Вячеславовна[3] vsulimova@yandex.ru
[1]Москва, Московский физико-технический институт
[2]Москва, Вычислительный центр РАН
[3]Тула, Тульский государственный университет

Принцип минимизации регуляризованного эмпирического риска в пределах обучающей совокупности приводит к оценке направляющего вектора линейной модели зависимости в виде линейной комбинации векторов признаков обучающих объектов [1]. Коэффициенты этого представления являются множителями Лагранжа при обобщенных линейных моделях обучающих объектов и играют роль аргументов в двойственной формулировке исходной задачи. Выпуклая двойственная задача имеет полиномиальную вычислительную сложность по числу множителей Лагранжа, т.е. по числу объектов, а сами искомые компоненты направляющего вектора определяются затем независимо друг от друга как линейные комбинации обучающих векторов признаков, т.е. с линейной вычислительной сложностью относительно числа признаков. В данной работе предлагается приближенная версия алгоритма решения двойственной задачи с линейной вычислительной сложностью по числу объектов. Эвристика заключается в том, что всякий раз при увеличении обучающей совокупности на один объект заново вычисляется лишь один новый множитель Лагранжа и один дополнительный коэффициент, общий для всех предыдущих множителей. Алгоритм пробегает по условно упорядоченой обучающей совокупности один раз, определяя приближенные коэффициенты представления направляющего вектора. Чем больше обучающая совокупность, тем точнее решение.

Работа поддержана грантом РФФИ № 19-37-90159.

[1] *V. Mottl, O. Krasotkina, A. Morozov, at al.* Linear complexity algorithms for high dimensional SVM and regression problems with smart sparse regularization. Proceedings of the 15th Int. Conf. on Machine Learning and Data Mining MLDM 2019, Vol. I, pp. 412-430.

On-line estimation of generalized linear dependence models from growing training sets

Alexey Morozov[1]⋆ ao.morozov@phystech.edu
Vadim Mottl[2] vmottl@yandex.ru
Valentina Sulimova[1] vsulimova@yandex.ru
[1]Moscow Institute of Physics and Technology
[2]Moscow State University

The principle of regularized empirical risk minimization within the bounds of the given training set results in estimating the direction vector of the linear model of the dependence as a linear combination of feature vectors representing the training objects [1]. Coefficients of this combination are Lagrange multipliers at generalized linear models of the training objects and occur as arguments in dual formulation of the primal problem. The convex dual problem has polynomial computational complexity relative to the number of Lagrange multipliers, i.e., in the nuber of training objects, whereas the sought-for components of the direction vector are then to be computed as the linear combinations of the training feature vectors, namely, with linear computational complexity in the number of features. In this work, we propose an algorithm for approximate solving the dual problem with linear computational complexity relative to the number of objects. When the training set is increased by one object, the heuristic step is to compute only one new Lagrange multiplier and one additional coefficient, common to all the previous Lagrange multipliers. The algorithm runs only once through the training set, and defines the coefficients of an approximate representation of the direction vector. The greater is the trainig set, the more precise is the solution.

This research is funded by RFBR, Grant No. 19-37-90159.

[1] *V. Mottl, O. Krasotkina, V. Sulimova, at al.* Linear complexity algorithms for high dimensional SVM and regression problems with smart sparse regularization. Proceedings of the 15th Int. Conf. on Machine Learning and Data Mining MLDM 2019, Vol. I, pp. 412-430.

Вычислительная сложность восстановления обощенных линейных моделей зависимостей

Моттль Вадим Вячеславович[1]⋆ vmottl@yandex.ru
Сулимова Валентина Вячеславовна[2] vsulimova@yandex.ru
Морозов Алексей Олегович[3] ao.morozov@phystech.edu
Пугач Илья Александрович[3] iliapugach@gmail.com
Татарчук Александр Игоревич[1] aitech@yandex.ru
[1]Москва, Вычислительный центр РАН
[2]Тула, Тульский государственный университет
[3]Москва, Московский физико-технический институт

Обычно, когда говорят о восстановлении зависимостей в больших массивах данных, то предполагают, что множество прецедентов не умещается в памяти одного компьютера, и необходимо использование технологии распределенных вычислений. Однако, даже если вся обучающая совокупность размещена в одном компьютере, остается вопрос о времени, необходимом для обучения. В данной работе мы опираемся на обобщенную линейную методологию восстановления зависимостей, покрыающую, в частности, оценивание регрессионных моделей и распознавание образов [1]. Предполагается, что обучающая информация имеет вид прямоугольной таблицы "объект-параметр". Мы рассматриваем два вида алгоритмов минимизации регуляризованного эмпирического риска, взаимно противоположных по их вычислительной сложности относительно двух размеров таблицы "объект-параметр". Вычислительная сложность первого алгоритма линейна по числу объектов и полиномиальна по числу признаков, а другой алгоритм, наоборот, полиномиален по числу объектов и линеен по числу признаков. Тот факт, что для любой комбинации размеров таблицы "объект-параметр" есть алгоритм, вычислительная сложность которого линейна по большему из двух размеров и полиномиальна по меньшему, особенно благоприятен в типичной ситуации, когда число признаков больше, чем число объектов.

Работа поддержана грантом РФФИ №18-07-01087.

[1] *V. Mottl, O. Krasotkina, V. Sulimova, at al.* Linear complexity algorithms for high dimensional SVM and regression problems with smart sparse regularization // Proceedings of the 15th Int. Conf. on Machine Learning and Data Mining MLDM 2019, Vol. I, pp. 412-430.

Computational complexity of dependence estimation in linear feature spaces

Vadim Mottl[1]⋆ vmottl@yandex.ru
Valentina Sulimova[2] vsulimova@yandex.ru
Alexey Morozov[3] ao.morozov@phystech.edu
Ilya Pugach[3] iliapugach@gmail.com
Alexander Tatarchuk[1] aitech@yandex.ru

[1]Moscow, Computing Center of the Russian Academy of Sciences
[2]Tula State University
[3]Moscow Institute of Physics and Technology

Usually, when speaking about dependence estimation in big sets of empirical data, it is adopted to suggest that the set of precedents does not fit in the memory of one computer, and some technology of distributed computing is required. However, even if the entire training set can be placed in one computer, the question remains how much time the training process will take. We keep here to the generalized linear methodology of dependence estimation, which covers, in particular, both regression estimation and pattern recognition [**?**]. It is assumed that the training information (empirical data set) is a rectangular objects/features table. We consider here two kinds of algorithms of regularized empirical risk minimization, which are mutually opposite in their computational complexity relative to the number of features and the number of objects, i.e., to the two sizes of the objects/features table. The computational complexity of one of them is linear with respect to the number of objects and polynomial relative to the number of features, whereas the other algorithm is of polynomial complexity in the number of features and linear in that of training objects. The fact, that for any combination of the two sizes of the objects/features table we have an algorithm whose computational complexity is linear relative to the greater and polynomial to the smaller of them, is especially favorable for the typical situation when the number of features is much greater than that of objects.

This research is funded by RFBR, Grant No. 18-07-01087.

bibitemmottl *V. Mottl, O. Krasotkina, V. Sulimova, at al.* Linear complexity algorithms for high dimensional SVM and regression problems with smart sparse regularization // Proceedings of the 15th Int. Conf. on Machine Learning and Data Mining MLDM 2019, Vol. I, pp. 412-430.

Методы достижения интепретируемости алгоритмов машинного обучения

Сенько Олег Валентинович [1] senkoov@mail.ru
Кузнецова Анна Викторовна[2]⋆ azfor@yandex.ru
[1]Москва, ФИЦ ИУ РАН
[2]Москва, Институт биохимической физики им. Н.М.Эмануэля

Несмотря на значительные успехи машинного обучения, достигнутые за последние годы и связанные прежде всего со значительным увеличением точности решения разнообразных задач распознавания, диагностики, прогнозирования, сдерживающим фактором использования этих технологий в различных областях является проблема непрозрачности, то есть недоступности для пользователей процесса принятия решений. Способом решения этой проблемы является выделение набора относительно простых и доступных для пользователя алгоритмов, которые в совокупности частично отображают процесс получения решения сложным, но обеспечивающим высокую точность алгоритмом машинного обучения. Прозрачность и интерпретируемость обучения может достигаться через графическое представление работы выделенных простых алгоритмов. При этом целесообразно добиваться по возможности полного представления представления всех значимых эффектов. Такую схему достижения прозрачности представляется более эффективной, если она применяется к ансамблям более простых алгоритма. В качестве примера можно привести решающие леса, генерируемые с помощью методов бэггинг или бустинг. Система обеспечения прозрачности несомненно будет заслуживать большего доверия пользователей, если она будет включать в рассмотрение только достоверные эффекты. Поскольку исходные алгоритмы строятся по данным, то для оценки значимости необходимо использовать средства статистической верификации. В условиях ограниченности данных верификацию закономерностей поиск закономерностей и ерификацию проводить по одной и той же обучающей выборке. Такая возможность предоставляют технологии ресэмплинга, включая перестановочные тесты. В качестве примера использования такого подхода можно предсавить метод Оптимальных достоверных разбиений. Следует также отметить, что верификации более сложных закономерностей должна обязательно состоять в вычислении нескольких p-значений, оценивающих значимость различных элементов. В противном случае на выходе оказывается избыточное, часто чрезвычайно большое число частично ложных закономерностей. Предполагается учёт при верификации проблемы множественного тестирования. Отличительной особенностью варианта ОДР, направленного на обеспечение прозрачности решающих лесов, является ранжирование достоверных закономерностей в соответствии со встречаемостью соответствующих сочетаний признаков в решающих деревьях, входящих в ансамбль. Настоящее исследование поддержано РФФИ, грант 17-07-01362.

[1] *Kuznetsova A.V.*, *Kostomarova I.V.*, *Senko O.V.*, Modification of the method of optimal valid patitioning for comparison of patterns related to the occurrence of ischemic stroke in two groups of patients. // Pattern Recognition and Image Analysis, 2014, v. 24, № 1, p. 114-123

The method of interpretability achieving in machine learning

Oleg Senko[1] senkoov@mail.ru
Anna Kuznetsova[2]⋆ azfor@yandex.ru
[1]Moscow, FRCCSC of the Russian Academy of Sciences
[2]Moscow, Emanuel Institute of biochemical physics of RAS

Despite the significant advances in machine learning in recent years and are primarily associated with a significant increase the accuracy of solving a variety of recognition, diagnostics, and prediction problems, a deterrent to the use of these technologies in different areas is the problem of opacity, that is, inaccessibility for users of the decision-making process. Way to solve this problem is the selection of a set of relatively simple and accessible to the user algorithms, which at least partially display the process of obtaining a solution by a complex, but providing high accuracy machine learning algorithm.

Transparency and interpretability of training can be achieved through a graphical representation of the work of selected simple algorithms. At the same time, it is advisable to achieve the fullest possible representation of the presentation of all significant effects. Such a scheme for achieving transparency seems more effective if it is applied to ensembles of simpler algorithms. An example is decisive forests generated using bagging or boosting methods. A transparency system will undoubtedly deserve more user confidence if it include only reliable effects. Since the original algorithms are built on the basis of data, it is necessary to use tools to assess significance statistical verification. In conditions of limited data, verification of patterns; search for patterns and Verification is carried out on the same training sample. This opportunity is provided by resampling technologies, including permutation tests. As an example of using this approach, we can present the Optimal reliable partitions method. It should also be noted that verification of more complex patterns must necessarily consist in calculating several $p-$values evaluate the significance of various elements. Otherwise, the output is redundant, often an extremely large number of partially false patterns. It is supposed to take into account when verifying the problem of multiple testing. A distinctive feature of the ODR option, aimed at ensuring transparency of decision forests, there will be a ranking of reliable patterns in accordance with the occurrence of relevant combinations of signs in the decisive trees included in the ensemble. This research is funded by RFBR, grant 17-07-01362.

[1] *Kuznetsova A. V., Kostomarova I. V., Senko O. V.* Modification of the method of optimal valid patitioning for comparison of patterns related to the occurrence of ischemic stroke in two groups of patients. // Pattern Recognition and Image Analysis, 2014, v. 24, No 1, Pp. 114-123

Метод генерации оптимальных ансамблей решающих деревьев

Медведев Дмитрий Владимирович[1] dm.medvedev97@gmail.com
Сенько Олег Валентинович[2]⋆ senkoov@mail.ru
[1]Москва, МГУ им. М.В. Ломоносова
[2]Москва, ФИЦ ИУ РАН

Эффективность методов распознавания, основанных на вычислении коллективных решений по наборам более простых алгоритмов, подтверждается результатами их применения при решении разнообразных прикладных задачах. Наибольшее распространение получили методы, использующие ансамбли решающих деревьев. Используются технологии генерации ансамблей, основанные на различных принципах, включая бэггинг, градиентный и адаптивный бустинг. В методе бэггинг каждое решающее дерево, добавляемое в ансамбль, строится по новой обучающей выборке, которая генерируется из исходной выборки с помощью бутстрэп. В градиентном бустинге на каждом шаге в ансамбль добавляется решающее дерево, аппроксимирующее зависимость от признаков так называемых псевдоостатков, представляющих собой первые производные квадратичного функционала потерь по прогнозам сделанным с помощью ансамбля, полученного на предыдущем шаге. В адаптивном бустинге на каждом шаге в ансамбль добавляется решающее дерево, которое строится исходя из условия увеличения вклада в обучение объектов, ошибочно классифицированных ансамблем, полученным на предыдущем шаге. Дополнительный способ генерации оптимальных ансамблей может быть выведен и разложения ошибки выпуклых комбинаций предикторов. Предположим, что бинарная индикаторная функция целевого класса Y прогнозируется по признакам $X_1, \ldots, X_n$ с помощью набора алгоритмов $A_1, \ldots, A_r$. Для ошибки выпуклой комбинации $A_{cc} = \sum_{i=1}^{r} c_i A_i$, где $\sum_{i=1}^{r} c_i = 1$ и $c_i \geqslant 0$ при $i = 1, \ldots, r$, справедливо следущее разложение [1]

$$\delta(A_{cc}) = \sum_{i=1}^{r} c_i \delta(A_i) - \frac{1}{2} \sum_{i=1}^{r} \sum_{j=1}^{r} c_i c_j \rho(A_i, A_j)^2, \tag{1}$$

где $\delta(A) = \mathsf{E}(Y - A)^2$ - ошибка некоторого алгоритма A при прогнозировании Y;
$\rho(A) = \mathsf{E}(A_i - A_j)^2$ - расхождение между алгоритмами A_i и A_j в пространстве прогнозов. Из разложения (1) хорошо видно, что величина ошибки коллективного решения существенно зависит как от ошибок отдельных алгоритмов, так и от расхождения между алгоритмами.Следовательно ансамбль должен генерироваться согласно обоим условиям.
Можно преложить схему генерации ансамбля, в которой на каждом шаге в него добавляется алглоритм, который одновременно

— по возможности точнее аппроксимирует связь Y с признаками $X_1, \ldots, X_n$ на обучающей выборке,
— по возможноти удалён от алгоритмов, ранее включённых в ансамбль.

Выполенение двух этих условий может быть достигнуто, когда построение нового добавляемого в ансабль алгоритма производится через максимизацию функционала

$$\mathbf{Q}_{full} = \mathbf{Q}_{approximation} + \gamma \mathbf{Q}_{divergence}, \tag{2}$$

где слагаемое $\mathbf{Q}_{approximation}$ характеризует качество аппроксимации, слагаемое $\mathbf{Q}_{divergence}$ характеризует характеризует расхождение прогнозо, который вычисляется добавляемым в ансамбль алгоритмов, и прогноза, вычисляемого сформированным на предыдущем шаге ансамблем, γ - задаваемая пользователем константа, характеризующая баланс двух слагаемых.

В методе решающих деревьев построение каждого нового узла в дереве производится через оптимизаци функционала $Q_T(\widetilde{S}_{inp}) = \nu_l H(\widetilde{S}_l) + (1-\nu_l) H(\widetilde{S}_r)$, где $\widetilde{S}_l$ и $\widetilde{S}_r$ являются двумя подвыборками, образующимися в результате применения привязанного к узлу правила к входной выбjрке $\widetilde{S}_{inp} = \widetilde{S}_l \cup \widetilde{S}_r$, $\nu_l = \frac{|\widetilde{S}_l|}{|\widetilde{S}_{inp}|}$, $H(\widetilde{S})$ - функционал неоднородности для выборки $\widetilde{S}$. Например, в качестве функционала неопределённости может выступить энтропийный индекс $H_e = -\sum_i^L \hat{p}_i \log \hat{p}_i$, где $\hat{p}_i$-доля объектов из класса K_i в выборке $\widetilde{S}$.Тогда Функционал $Q_{eT}(\widetilde{S}_{inp}) = = \nu_l H_e(\widetilde{S}_l) + (1-\nu_l) H_e(\widetilde{S}_r)$ может рассматриваться как слагаемое $\mathbf{Q}_{approximation}$. При этом в качестве слагаемого $\mathbf{Q}_{divergence}$ может рассматриваться функционал $Q^-_{eT}(\widetilde{S}_{inp}) = \nu_l H^-_e(\widetilde{S}_l) + (1-\nu_l) H^-_e(\widetilde{S}_r)$, $H^-_e(\widetilde{S}) = \sum_{i=1}^L \hat{p}^-_i \log \hat{p}^-{}_i$, а $\hat{p}^-_i$ является среднее значение оценок вероятности принадлежности классу K_i, вычисляемое ансамблем сгенерированном на предыдущем шаге. То есть $\hat{p}^-_i = \frac{1}{|\widetilde{S}|} \sum_{\mathbf{x}_j \in \widetilde{S}} \hat{p}^- -(i|\mathbf{x}_j)$, где оценка вероятности принадлежности объекта $\mathbf{x}_j$ классу K_i. Метод основанный на использовании функционала $Q_{eT}(\widetilde{S}_{inp}) + \gamma * Q^-_{eT}(\widetilde{S}_{inp})$ при построении каждого нового дерева, добавляемого в ансамбль был протестирован на ряде прикладных задач, представленных в известных репозиториях. При этом в качестве оценок за классы использовались усреднённые по онсамблю оценки вероятности.

Исследование показало, что метод демонстрирует эффективность, близкую к эффективности адаптивного бустинга. На некоторых задачах эффективность адаптивого бустинга была превышена. Проведеные исследования подтверждают перспективность рассмотренного подхода

Настоящее исследование поддержано РФФИ, грант 17-07-01362.

[1] *Докукин А.А.,Сенько О.В.* Оптимальные выпуклые корректирующие процедуры в задачах высокой размерности // Ж вычисл. матем. и матем. физ., М: Наука, 2011, 51:9 с. 1751–1760.

Method for generation of optimal ensembles of decision trees

Dmitriy Medvedev[1] dm.medvedev97@gmail.com
Oleg Senko[2]⋆ senkoov@mail.ru
[1]Moscow, Lomonosov Moscow State University
[2]Moscow, FRC Computer Science and Control of RAS

The effectiveness of recognition methods based on calculation of collective solutions by sets of simpler algorithms is confirmed in various applied problems. The most popular are ensembles of decision trees. Several ensemble generation technologies based on various principles, including bagging, gradient and adaptive boosting. In the bagging method, each decision tree that is added to an ensemble is constructed using a new training set, which is generated from the initial set by bootstrap technique. In gradient boosting a decision tree is added at each step to the ensemble, which approximates the dependence of the so-called pseudo-residues on the features. pseudo-residues are the first derivatives of the quadratic loss functional if the ensemble obtained at the previous step is used for predicting. In adaptive boosting, at each step, a new decision tree is built with greater contribution of training objects that were erroneously classified by the ensemble obtained in the previous step. An additional way of generating optimal ensembles can received from decomposition of convex combination error. Let binary target class Y is predicted by features $X_1, \ldots, X_n$ with the help of algorithms$A_1, \ldots, A_r$. Decomposition is true convex combination $A_{cc} = \sum_{i=1}^{r} c_i A_i$, where $\sum_{i=1}^{r} c_i = 1$, $c_i \geqslant 0$ if $i = 1, \ldots, r$ [1]:

$$\delta(A_{cc}) = \sum_{i=1}^{r} c_i \delta(A_i) - \frac{1}{2} \sum_{i=1}^{r} \sum_{j=1}^{r} c_i c_j \rho(A_i, A_j)^2, \tag{1}$$

where $\delta(A) = \mathsf{E}(Y - A)^2$ - error of some algorithm A when Y is predicted ;
$\rho(A) = \mathsf{E}(A_i - A_j)^2$ - divergence between algorithms A_i in A_j forecasts space. . From the expansion (1) it is clearly seen that the magnitude of the collective decision error depends significantly on both the errors of individual algorithms and the discrepancy between the algorithms. Therefore, the ensemble must be generated according to both conditions. We can provide an ensemble generation scheme in which at each step an algorithm is added to ensemble, which simultaneously begin itemize item as much as possible approximates the relationship Y with the features X_1, $ldots, X_n$ at the training set, item as much as possible diverged from algorithms previously included in the ensemble. end itemize The fulfillment of these two conditions can be achieved when the construction of a new algorithm added to the ensemble is done through maximizing the functional

$$\mathbf{Q}_{full} = \mathbf{Q}_{approximation} + \gamma \mathbf{Q}_{divergence}, \tag{2}$$

where the term $\mathbf{Q}_{approximation}$ characterizes the quality of approximation, the term $\mathbf{Q}_{divergence}$ describes the discrepancy between the forecast calculated by the algorithm that added to the ensemble and the forecast calculated by the ensemble formed

in the previous step, γ - a user-defined constant that characterizes the balance of two terms.

In the decision tree method, each new node in the tree is constructed through optimization of the functional $Q_T(\widetilde{S}_{inp}) = \nu_l H(\widetilde{S}_l) + (1-\nu_l)H(\widetilde{S}_r)$, where $\widetilde{S}_l$ and $\widetilde{S}_r$ are two subsamples resulting from applying a node-bound rule to the input set $\widetilde{S}_{inp} = \widetilde{S}_l \cup \widetilde{S}_r$,$\nu_l = \frac{|\widetilde{S}_l|}{|\widetilde{S}_{inp}|}$, $H(\widetilde{S})$- heterogeneity functional for the sample $\widetilde{S}$. For example, the entropy index $H_e = -\sum_i^L \hat{p}_i \log \hat{p}_i$ may be used as heterogeneity functional,where $\hat{p}_i$ is the fraction of K_i inside set $\widetilde{S}$ Then the functional $Q_{eT}(\widetilde{S}_{inp}) = = \nu_l H_e(\widetilde{S}_l) + (1-\nu_l)H_e(\widetilde{S}_r)$ can be discussed as a term $\mathbf{Q}_{approximation}$. At that the functional $Q_{eT}(\widetilde{S}_{inp}) = \nu_l H_e(\widetilde{S}_l) + (1-\nu_l)H_e(\widetilde{S}_r)$ can be considered as the term $\mathbf{Q}_{divergence}$, and $\hat{p}_i^-$ is the average value of estimates the probability to belonging to K_i that is calculated by the ensemble generated in the previous step. That is, $\hat{p}_i^- = \frac{1}{|\widetilde{S}|}\sum_{\mathbf{x}_j \in \widetilde{S}} \hat{p}^-(i|\mathbf{x}_j)$, where $\hat{p}^-(i|\mathbf{x}_j)$ -the estimate of the probability of belonging of the object $\mathbf{x}_j$ to the class K_i. A method based on the use of the functional $Q_{eT}(\widetilde{S}_{inp}) + \gamma * Q_{eT}^-(\widetilde{S}_{inp})$ when constructing each new tree added the ensemble was tested on a number of applied tasks presented in well-known repositories. Moreover, as grades for classes ensemble-averaged probability estimates were used.

The study showed that the method demonstrates efficiency close to the effectiveness of adaptive boosting. On some tasks, the effectiveness of adaptive boosting has been exceeded. The conducted studies confirm the prospects of the considered approach.

This research is funded by RFBR, grant 17-07-01362.

[1] *Dokukin A. A.,Senko O. V.*Dokukin Optimal convex correcting procedures in high dimensional tasks // Comput. Math. Math. Phys., 2011, 51:9 c. 1644–1652.

Верификация и оптимизация регрессионных моделей на панелях экономических данных с использованием методов Монте Карло

Кирилюк Игорь Леонидович[1]⋆ igokir@rambler.ru
Сенько Олег Валентинович[2] senkoov@mail.ru
[1]Москва, Институт экономики РАН
[2]Москва, Федеральный государственный центр "Информатика и управление" РАН

В предыдущих исследованиях [1] нами были разработаны методы для оценивания статистической достоверности регрессионных моделей, описывающие мезоэкономические производственные функции для регионов Российской Федерации. Предложенные методы позволяют не только оценивать статистическую достоверность моделей, полученных по доступным панельным данным, с учётом возможной нестационарности экономических процессов, но и оценивать значимость отдельных регрессоров. Интересной особенностью в исследованиях производственных функций является то, что для их формализации предложено множество альтернативных вариантов моделей. В рамках научного направления «селекция моделей» предлагаются методы выбора моделей оптимальной сложности. В наших исследованиях достоверности выявляемых закономерностей определяются с применением методов Монте-Карло (в том числе, не параметрических их вариантов, таких, как перестановочные тесты и бутстрапы) на основании временных рядов данных.

Представляет интерес не только исследование значимости влияния на целевые переменные отдельных переменных и их совокупности, но и оценка выполнимости более сложных налагаемых на систему условий. В развитие предыдущих исследований с помощью вышеупомянутых методов нами предложен метод оценки достоверности ограничения, часто налагаемого на коэффициенты производственных функций — постоянства отдачи от масштаба производства. Для производственной функции Кобба-Дугласа, описываемой формулой

$$Y = AK^{\alpha}L^{\beta}, \tag{1}$$

где Y, K, L – величины, в разных случаях имеющие разную интерпретацию, условно обозначаемые как выпуск, капитал и труд, α и β – вычисляемые коэффициенты, постоянная отдача определяется условием $\alpha + \beta = 1$. Нарушение этого условия может интерпретироваться как свидетельство эмерджентных свойств системы, обусловленных тем, что подсистемы в ней активно взаимодействуют между собой, и в результате мешают друг другу, или, наоборот, кооперируются, увеличивая совокупный выпуск продукции.

Интересно, что для некоторых данных результаты моделирования дают отрицательные значения коэффициентов α или β. Некоторые исследователи делают из этого вывод о неадекватности модели данным, другие пытаются давать

полученному результату некоторые интерпретации. Проверка достоверности подобных результатов также представляет значительный практический интерес.

Моделирование производственных функций по временным рядам сталкивается с проблемой, заключающейся в том, что используемые временные ряды не стационарны по своей природе, что снижает достоверность результатов по сравнению с достоверностью результатов, которые были бы получены при исследовании стационарных временных рядов аналогичной длины. Кроме того, при работе с временными рядами с использованием метода наименьших квадратов достоверность результатов моделирования может снижаться также вследствие мультиколлинеарности (при нехватке данных). Надёжные методы оценки достоверности закономерностей должны учитывать возможность нестационарности временных рядов и эффекты корреляции между регрессорами.

В дополнение к расчётам для временных рядов нами используются расчёты для «пространственных данных» (в нашем случае, вычисляются производственные функции по данным для всех регионов за конкретные годы, а не по данным для временных рядов каждого региона по отдельности). Таким образом, вычисляются обобщенные по регионам производственные функции, но при этом, по крайней мере, проблема нестационарности данных во времени не возникает.

Отдача от масштаба, как и коэффициенты детерминации, вычисляемые по временным рядам (или по пространственным данным), по сути являются функционалами от наборов данных (для более точного исследования свойств системы используется разнообразный ассортимент других подобных функционалов). В наших исследованиях свойства совокупности регионов, различия между ними, возможность их разбиения на классы, анализируются с применением таких методов, как кластерный анализ и метод оптимально достоверных разбиений.

[1] *Кирилюк И. Л.*, *Сенько О. В.* Исследования соотношений между нестационарными временными рядами на примере производственных функций // Машинное обучение и анализ данных. Том 4, № 3, 2018. — с. 142–151.

Verification and optimization of regression models at panels of economical data with the help of Monte-Carlo techniques.

Igor Kirilyuk[1]⋆ igokir@rambler.ru
Oleg Senko[2] senkoov@mail.ru
[1]Moscow, Institute of Economics of RAS
[2]Moscow, FRC "Informatics and Control" of RAS

In previous studies [1] we developed methods aimed to evaluate statistical validity of regression models describing mesoeconomic production functions for Russian Federation regions. The proposed methods allow not only to evaluate the statistical reliability of the models obtained from available panel data, taking into account the possible non-stationarity of economic processes, but also to evaluate the significance of individual regressors. An interesting feature in the study of production functions is that for their formalization proposed many alternative models. Within the scientific direction. Selection of models. methods for selecting models of optimal complexity are proposed. In our studies, the reliability of the revealed patterns is determined using Monte Carlo methods (including non-parametric variants of them, such as permutation tests and bootstraps) on based on time series data.

It is of interest not only to study the significance of the influence of individual variables and their combination on target variables, but also to assess the feasibility of more complex conditions imposed on the system. In the development of previous studies, we proposed technique aimed to evaluate reliability of the constraint that is often imposed on the coefficients of production functions - the constancy of returns to scale of production. For the Cobb-Douglas production function described by the formula

$$Y = AK^{\alpha}L^{\beta}, \tag{1}$$

where Y, K, L are quantities that in different cases have different interpretations, conventionally referred to as output, capital and labor, α andβ are calculated coefficients, constant return to scale is determined by the condition $\alpha + \beta = 1$. Violation of this condition can be interpreted as evidence of the emergent properties of the system, due to the fact that the subsystems in it actively interact with each other, and as a result interfere with each other, or, conversely, cooperate, increasing the total output.

Interestingly, for some data, the simulation results give negative values of the coefficients α or β. Some researchers conclude from this that the model is inadequate for the data, while others try to give some interpretations to the result. Validation of such results is also of significant practical interest.

Modeling production functions by time series faces the problem that the time series used are not stationary in nature, which reduces the reliability of the results compared with the reliability of the results that would be obtained by studying stationary time series of a similar length. In addition, when working with time series

using the least squares method, reliability of modeling is decreased in presence of multicollinearity in data (with insufficient data). Reliable methods for assessing the validity of patterns should take into account the possibility of non-stationaryty of time series and the effects of correlation between regressors.

In addition to calculations for time series, we use calculations for "spatial data" (in our case, production functions are calculated from data for all regions for specific years, and not according to data for time series of each region separately). Thus, production functions generalized by region are calculated, but at the same time, at least, the problem of non-stationary data in time does not arise.

The returns to scale, as well as the determination coefficients calculated from time series (or spatial data), are essentially functionals of data sets (for a more accurate study of the properties of the system, a diverse assortment of other similar functionals is used).In our studies, the properties of the totality of regions, the differences between them, the possibility of dividing them into classes, are analyzed using methods such as cluster analysis and the method of optimally reliable partitions.

[1] *Kirilyuk I.*, *Senko O.* Studies of the relationship between non-stationary time series on the example of production functions // Journal of Machine Learning and Data Analysis, 2018.

Повышение детализации трехмерных моделей местности с использованием генеративных состязательных сетей

Визильтер Юрий Валентинович[1] viz@gosniias.ru
Горбацевич Владимир Сергеевич[1] gvs@gosniias.ru
Мельнеченко Михаил Александрович[1]⋆ mmelnich@gosniias.ru
[1]г. Москва, ФГУП ГосНИИАС

В работе рассматривается проблема повышения качества грубых моделей земной поверхности. Трехмерную модель ландшафта удобно представлять в виде карт высот(Heightmap) – матрицы содержащей абсолютные значения высоты местности. Можно представить низкокачественную трехмерную модель ландшафта как Heightmap высокой размерности но с малым количеством заданных опорных значений. Высококачественная модель ландшафта соответвует полностью заданной карте высот. Такое представление(в виде матрицы) позволяет использовать для обработки 3-мерной модели классические CNN и во многом эквивалентно воксельному представлению широко используемому при обработке 3D-моделей алгоритмами на основе CNN. Таким образом исходная задача может быть представлена как задача восстановления плотной карты высот по разреженной.

В качестве базового алгоритма используется метод Pix2Pix. Такой подход позволяет успешно решать различные задачи схожие с рассматриваемой. В качестве входных данных для генератора используются спутниковые снимки и разреженные карты высот. Объединения данных производится через конкатенацию. В качестве генератора используется ГКНС U-Net. В качестве Дискриминатора PatchGAN.Для обучения ГКНС был синтезирован набор данных состоящий из 150 тысяч изображений различных территорий. Тестирование на реальных данных показало, что предлагаемый метод позволяет генерировать высококачественные 3-мерные модели земной поверхности с высокой точностью и скоростью.

[1] *Визильтер Ю. В. и др.* Повышение детализации трехмерных моделей местности с использованием генеративных состязательных сетей // ВКИТ, Город: Москва Издательство Спектр, 2019. (В печати)

3D Terrain Model Enhancing Using Generative Adversarial Network

Yuri Vizilter[1] viz@gosniias.ru
Vladimir Gorbatsevich[1] gvs@gosniias.ru
Mikhail Melnechenko[1]⋆ mmelnich@gosniias.ru
[1]Moscow, GosNIIAS

The paper addresses the problem of low quality 3D terrain models enhancement in automatic mode. To solve this problem, we propose an approach based on convolutional neural networks (CNN). 3D terrain model can be represented as a heightmap – the 2D matrix that contains surface elevation data. Therefore, using this approach, we can represent the low quality 3D terrain model as the high dimensional heightmap with small number of non-zero values (sparse matrix), whilst the high quality 3D terrain model as the dense matrix. Such type of representation allows us to use classical CNNs for 3D models processing, and it is largely equivalent to voxel representation, which is widely used for CNN-based algorithms of 3D models analysis. As a result, our task can be transformed to the task of the dense heightmap restoration using the sparse heightmap. The algorithm is based on Pix2Pix method that uses generative adversarial networks for similar problems. Satellite images and the sparse heightmap are used as input data for the Generator CNN. Data fusion is made by concatenation procedure. The Generator architecture is U-Net, the Discriminator architecture is PatchGAN. The training is performed on synthetic dataset that includes 150000 images and heightmaps of different landscapes. Tests on real data have shown that the algorithm can generate high quality 3D terrain models at high processing speed and with high accuracy.

[1] *Vizilter Yu. et al* 3D Terrain Model Enhancing Using Generative Adversarial Network // VKIT, City: Moscow, Spektr 2019. (In printing)

Алгоритм мимикрии с использованием генеративных состязательных сетей для задач обнаружения объектов

Визильтер Юрий Валентинович[1] viz@gosniias.ru
Горбацевич Владимир Сергеевич[1]⋆ gvs@gosniias.ru
Финогеев Евгений Сергеевич[1] finogeev@gosniias.ru
Моисеенко Анастасия Сергеевна[1] moiseenkoas@gosniias.ru
[1]г. Москва, ФГУП ГосНИИАС

На сегодняшний день существует множество практических приложений основанных на использовании глубоких сверточных (конволюционных) нейронных сетей(ГКНС). При этом одним из основных недостатков алгоритмов на основе ГКНС является их высокая вычислительная сложность, которая делает затруднительным их использование во встраиваемых системах. Несмотря на существенный прогресс в области специализированных нейроускорителей с низким потреблением энергии (например Google TPU Edge, Nvidia Xavier и др) данная проблема до сих пор актуальна. Для ее решения используются так называемые "мобильные" архитектуры ГКНС(MobileNet,ShuffleNet) – их отличительной особенностью является минимальное количество операций, необходимых для прямого прохода. К сожалению практическое использование данных архитектур является достаточно сложным т.к. они более зависимы от значений гиперпараметров по сравнению с обычными ГКНС.

В работе предлагается оригинальный алгоритм мимикрии ГКНС для задачи обнаружения объектов. Основной идеей предлагаемого подходя является использование генеративных состязательных сетей для повышения качества мимикрии. В качестве базового алгоритма обнаружения использовался алгоритм SSD, однако подход обобщается на любой другой алгоритм обнаружения объектов. Тестирование по публичным базам данных Pascal VOC 2007 и MS Coco показали, что использование предлагаемого подхода позволяет монотонно улучшить качество обнаружения.

[1] *Визильтер Ю. В. и др.* Алгоритм мимикрии с использованием генеративных состязательных сетей для задач обнаружения объектов // ВКИТ, Город: Москва Издательство Спектр, 2019. (В печати)

Knowledge distillation using GANs for object detection

Yuri Vizilter[1] viz@gosniias.ru
Vladimir Gorbatsevich[1]⋆ gvs@gosniias.ru
Eugeni Finogeev[1] finogeev@gosniias.ru
Anastasiia Moiseenko[1] moiseenkoas@gosniias.ru
[1]Moscow, GosNIIAS

There are many practical solutions available nowadays that are based on convolutional neural networks (CNN). The main drawback of CNN is a very high computational cost that makes it quite difficult to use CNN based applications on embedded systems. Despite the substantial advance in neural processing units of low power consumption such as Google TPU or NVIDIA Xavier, this problem is still acute. To counter this problem, special "mobile" CNN architectures have been developed (e.g. MobileNet, ShuffleNet). Compared to regular CNNs, these CNNs are very computational effective in inference (in terms of floating point operations), but their practical use is hampered by more dependence on hyperparameters selection.

In this paper, we propose a new CNN knowledge distillation (mimic) algorithm for object detection. In our approach, we use generative adversarial networks to improve mimic quality. Although in this work we used SSD as a basic algorithm, the proposed approach is general. Testing results on Pascal VOC 2007 and MS Coco datasets have shown that the algorithm provides detection quality for both datasets.

[1] *Vizilter Yu. et al* Knowledge distillation using GANs for object detection // VKIT, City: Moscow, Spektr 2019. (In printing)

Двухшаговый алгоритм семантического обнаружения на основе ГКНС

Визильтер Юрий Валентинович[1] viz@gosniias.ru
Горбацевич Владимир Сергеевич[1]⋆ gvs@gosniias.ru
Моисеенко Анастасия Сергеевна[1] moiseenkoas@gosniias.ru
[1]г. Москва, ФГУП ГосНИИАС

На сегодняшний день различные техники сравенения изображений активно используются при решении многих практических задач. К сожалению классические алгоритмы сравнения изображений сильно не устойчивы к существенным изменениям в условиях съемки. С другой стороны современные алгоритмы обнаружения объектов на основе ГКНС лишены этих недостатков. К сожалению они способны обнаруживать объекты только фиксированных классов, представленных в обучающей выборке. Однако для ряда практических задач необходимо проводить обнаружение объектов, основываясь на одном или нескольких примерах. Такая задача известна как задача семантического обнаружения. Семантический детектор принимает в качестве входных данных изображение по которому проводится поиск, а также одно или несколько изображений содержащих изображение объекта искомого класса.

В данной работе предлагается новый двухшаговый алгоритм семантического обнаружения.В отличии от однопроходных алгоритмов духшаговые позволяют добится более высокого качества обнраужения. Обнраужение проводится в два этапа. На первом этапе генерируются пердположения о положении объекта(Body CNN), на втором проводится их проверка(Head CNN). Предлагаемый алгоритм построен на схожих принципах. На первом этапе также генерируются предположения о положении объекта, ГКНС Head CNN проводит сравнение соответсвующих предположениям глубоких признаков с глубокими признаками полученными из изображения-запроса. В работе предлагается алгоритм обучения таких ГКНС и приведены результаты тестов по публичным БД.

[1] *Визильтер Ю. В. и др.* Двухшаовый алгоритм семантического обнаружения // ВКИТ, Город: Москва Издательство Спектр, 2019. (В печати)

Region proposal CNN based semantic matcher

Yuri Vizilter[1] viz@gosniias.ru
Vladimir Gorbatsevich[1]⋆ gvs@gosniias.ru
Anastasiia Moiseenko[1] moiseenkoas@gosniias.ru
[1]Moscow, GosNIIAS

Image matching techniques are widely used in various practical applications e.g. stereo vision, 3D reconstruction, structure-from-motion, SLAM landmark detection, object tracking and so on. However, classical image matching techniques cannot match images with essential shape or pose inter-frame changes. On the other hand, object detection techniques presume the detection and localization of all objects of some given class. Currently, all modern object detectors are based on deep convolutional neural networks (CNN). Unfortunately, there is an important limitation of CNN-based detectors: they can detect only the objects of already "seen" classes, i.e. presented at CNN training datasets. However, sometimes we need to detect objects of previously unseen classes based on just one or few sample images available at the execution stage. This unseen object detection problem is known as semantic matching problem. So semantic matcher takes two images as input – request image and test image. Request image represents object class needed to be found on test image.

In this paper, we propose a new region proposal based semantic matcher. In comparison to single shot detectors region based or two stage detectors provides better quality with lower speed. The two-stage detector divides the task into two steps: the first step (Body-CNN) generates many proposals(regions of interest - ROIs), and the second step (Head-CNN) focuses on the recognition of the proposals. In our region based semantic matcher we use same ideas. Our Body CNN also generates proposals like in classical Faster R-CNN, and Head-CNN compares proposals with request descriptor, extracted from request image. To extract features from request image we also use CNN – Request descriptor CNN. Training algorithm and testing results on public databases are provided.

[1] *Vizilter Yu. et al* Region proposal CNN based semantic matcher // VKIT, City: Moscow, Spektr 2019. (In printing)

Об алгебраических свойствах операций, используемых при построении современных свёрточных нейронных сетей

Фадеев Егор Павлович[1]⋆ fadeevegor@yandex.ru
Зубюк Андрей Владимирович[1] zubuk@cmpd2.phys.msu.ru
[1]Москва, МГУ имени М.В.Ломоносова, физический факультет

Сегодня для решения задач компьютерного зрения, распознавания речи и др. широко используются искусственные нейронные сети (ИНС). Многие архитектуры ИНС включают блоки, состоящие из свёрточных слоёв, активаций и пулингов. Известно, что одни виды этих операции позволяют достигать бóльших показателей качества, чем другие (например, ИНС с max-пулингами, как правило, превосходят ИНС с усредняющими пулингами). Однако, обоснованного объяснения этому не дано.

В большинстве случаев свёртки, активации и пулинги рассматриваются как манипуляции над пикселями изображений (отсчётами, если речь идёт об иных данных): свёртки — как выделение локальных «признаков» (например, границ объектов, характеризующихся больши́ми градиентами яркостей), пулинги — как выделение «наиболее интенсивных» признаков и отсечение «менее интенсивных» и т. д.

В настоящей работе предлагается рассматривать свёртки, активации и пулинги как операторы, действующие в евклидовых векторных пространствах, элементами (далее — *векторами*) которых являются изображения и др. многомерные массивы. Авторы полагают, что исследование алгебраических свойств этих операторов позволит лучше понять отличия различных видов используемых при построении ИНС операций (например, max-пулингов и усредняющих пулингов), а также определить, какие их свойства ключевым образом влияют на качество ИНС в целом.

С предложенной точки зрения свёрточный слой представляет собой линейный оператор, его сингулярными базисами являются базисы Фурье.

Активация — нелинейная функция одной переменной, применяется к каждой координате вектора в *каноническом* базисе в отдельности (каноническим будем называть ортонормированный базис, каждый вектор которого есть массив, состоящий из единственной единицы и остальных нулей). Одной из наиболее часто используемых в современных ИНС функцией активации является ReLU, определённая как $\mathrm{ReLU}(x) = \max\{0, x\}$. Применение ReLU к каждой координате вектора в каноническом базисе есть не что иное, как *проецирование* на неотрицательный ортант канонического базиса — коническую оболочку, натянутую на векторы канонического базиса.

Пулинг — линейный или нелинейный оператор, понижающий размерность [2, 3]. Композиция активации ReLU и одного из наиболее часто применяемых пулингов — max-пулинга — также может быть выражена через оператор *проецирования* [1, 4]. Действительно, рассмотрим действие такой композиции на

фрагмент f изображения, попавший в одно «окно пулинга» размером $n \times n$. Пусть $e_{ij},\ i, j = 1, \ldots, n,$ — векторы, образующие канонический базис (для фрагмента изображения). Пусть A_{ReLU} — оператор покоординатного применения активации ReLU, и $P_{\max}$ — оператор max-пулинга. Тогда $P_{\max} A_{\mathrm{ReLU}} f$ — норма проекции f на множество $\bigcup_{i,j=1}^{n} \{\alpha e_{ij},\ \alpha \geqslant 0\}$, т. е. на объединение «рёбер» неотрицательного ортанта канонического базиса.

Как видно из вышеизложенного, блоки типа «свёртка — активация — пулинг» с алгебраической точки зрения представляют собой линейное преобразование с последующим проецированием на нелинейное множество, тесно связанное с базисом, отличным от сингулярных базисов линейного преобразования. В работе исследована зависимость качества ИНС в задаче классификации изображений из базы CIFAR10 от выбора сингулярных базисов линейного преобразования и базиса, определяющего множества, на которые осуществляется проецирование. Показано, что в ряде случаев замена рассмотренных выше базисов Фурье и канонического базиса на отличные от них базисы не снижает качество ИНС. Это позволяет выдвинуть гипотезу о том, что критическое влияние на качество ИНС оказывают отмеченные выше общие алгебраические свойства рассмотренных операций, а не их конкретный вид. Однако, эта гипотеза требует дальнейшего изучения.

Работа поддержана грантом РФФИ No17-07-00832.

[1] *Зубюк А. В., Байков В.Г.* Идемпотентная нейронная сеть как реализация морфологического метода узнавания объектов по изображениям // Москва: Техническое зрение в системах управления - 2018, 2018 — p. 53–54.

[2] *Passricha V., Aggarwal R.K.* A comparative analysis of pooling strategies for convolutional neural network based Hindi ASR // J Ambient Intell Human Comput, Berlin: Springer Berlin Heidelberg, 2019. — p. 5–25.

[3] *Rippel O., Snoek J, Adams R.P.* Spectral Representations for Convolutional Neural Networks // In: Proceedings of the 28th international conference on neural information processing systems (NIPS'15), vol 2, 2015 — p. 2449–2457.

[4] *Jin L., Li S., Hu B, Liu M.* A survey on projection neural networks and their applications // Applied Soft Computing, 2019. —vol. 76, p. 533–544.

On the algebraic properties of operations constituting modern convolutional neural networks

Egor Fadeev[1]⋆ fadeevegor@yandex.ru
Andrew Zubuk[1] zubuk@cmpd2.phy.msu.ru
[1]Moscow, Faculty of Physics, M.V.Lomonosov Moscow State University

Artificial neural networks (ANN) are widely used in areas like computer vision, speech recognition, etc. Many ANN architectures include units consisting of convolutional layers, activations and pooling layers. Some kinds of these operations are known to allow get better quality then others (e.g. ANN with max-pooling usually outperform ANN with average pooling). However, there isn't reasonable explanation for this phenomenon.

Convolutions, activations and poolings are usually considered as manipulations with image pixels. Convolutions are considered as extraction of local features (e.g. object border characterized by high value gradients), poolings are considered as extraction most intense features, etc.

In this article convolutions, activations and poolings are proposed to be considered as operators which act on elements (images or other high dimensionals arrays) of euclidian vector spaces. Authors suppose that study of algebraic properties of these operators will allow us to better understand differences between various operations being used in ANN (e.g. max and average poolings) and to determine which properties of them have crucial impact on entire ANN quality.

From proposed point of view convolution layer is linear operator, singular bases of which are Furier bases.

Activation is nonlinear function of one variable, which is applied to each vector coordinate in canonical basis independently (we will call orthonormal basis canonical if each basis vector is array consisting of single one and zeros). One of the most used activations in modern ANN is ReLU defined as $\mathrm{ReLU}(x) = \max\{0, x\}$. Applying ReLU to each vector coordinate in canonical basis is projecting on nonnegative orthant of canonical basis (conical hull of basis vectors).

Pooling is linear or nonlinear downsampling operator. Composition of ReLU and max-pooling may be represented as projection operator. Indeed, lets consider action of such composition on subarea f which is matching with one "pooling window" size of which is $n \times n$. Let $e_{ij},\ i, j = 1, \ldots, n,$ be vectors forming canonical basis for this fragment. Let A_{ReLU} be operator of elementwise applying ReLU activation and $P_{\max}$ be max-pooling operator. Then $P_{\max} A_{\mathrm{ReLU}} f$ is norm of projection f on set $\bigcup_{i,j=1}^{n} \left\{\alpha e_{ij},\ \alpha \geqslant 0\right\}$, i.e. union of ribs of nonnegative orthant of canonical basis.

As it is seen from foregoing, units consisting of convolution, activation and pooling are linear transformation with projecting on nonlinear set which are closely related with basis different from singular basis of linear transformation. In this work dependency of ANN quality in classification problem from choice of singular basis of linear transformation and basis determing sets on which projection is performing.

It is shown, that replacement Furier and canonical bases on different ones doesn't reduce quality of ANN in some cases. It allows us to propose hypothesis that crucial influence on ANN quality are made by common algebraic properties of considered above operations, not by specific type of these operations. However, this hypothesis requires furthermore study.

This research is funded by RFBR, grant 17-07-00832.

[1] *Zubyuk A. V., Baykov V. G.* Idempodent neural network as realization of morphological method of object recognition in images Moscow: Technical vision in control systems – 2018 — p. 53–54 (in russian)

[2] *Passricha V., Aggarwal R.K.* A comparative analysis of pooling strategies for convolutional neural network based Hindi ASR // J Ambient Intell Human Comput, Berlin: Springer Berlin Heidelberg, 2019. — p. 5–25.

[3] *Rippel O., Snoek J, Adams R.P.* Spectral Representations for Convolutional Neural Networks // In: Proceedings of the 28th international conference on neural information processing systems (NIPS'15), 2015. — vol. 2, p. 2449–2457.

[4] *Jin L., Li S., Hu B, Liu M.* A survey on projection neural networks and their applications // Applied Soft Computing, 2019. — vol. 76, p. 533–544.

Новый тип вейвлет-нейронных сетей

Ефиторов Александр Олегович[1]⋆ a.efitorov@sinp.msu.ru
Доленко Сергей Анатольевич[1] dolenko@sinp.msu.ru
[1]Москва, Научно-исследовательский институт ядерной физики имени Д. В. Скобельцына

Вейвлет-преобразование – метод спектрального анализа сигналов, который использует в качестве базиса особый тип функций, обладающих свойствами локализации и ограниченности. Эти свойства позволяют проводить эффективный анализ нестационарных сигналов. Стандартный подход (например, дискретное вейвлет-преобразование) с фиксированными коэффициентами масштаба и сдвига может быть неоптимальным в контексте решения обратной задачи (ОЗ) ($X = f^{-1}(Y)$), так как процедура разложения сигнала по базису не предполагает наличия обратной связи с целевой функцией данной задачи. Поскольку в случае решения ОЗ обратное вейвлет-преобразование не требуется, значения коэффициентов масштаба и сдвига могут быть определены во время обучения сети, а окна, соответствующие различным положениям вейвлет-функций, могут перекрываться. В данном исследовании мы предлагаем новый тип вейвлет-нейронных сетей - ВНС с адаптивным окном (АОВНС), предназначенные для решения ОЗ, в качестве входных данных принимающие нестационарный сигнал. Процедура обучения сети представляется итеративной процедурой выбора оптимальных коэффициентов сдвига (положения окна) и масштаба (ширины окна) для выбранной вейвлет-функции. Две модификации этого нового типа ВНС сравнивались с линейной регрессионной моделью и многослойным персептроном на примере задачи Mackey-Glass.

[1] *Efitorov A. and Dolenko S.* A New Type of a Wavelet Neural Network // Optical Memory and Neural Networks, 2018. —V.27 p. 152–160. https://doi.org/10.3103/S1060992X18030050.

A New Type of a Wavelet Neural Network

Alexander Efitorov[1]⋆ a.efitorov@sinp.msu.ru
Sergey Dolenko[1] dolenko@sinp.msu.ru
[1]Skobeltsyn Institute of Nuclear Physics

Wavelet transformation uses a special basis widely known for its unique properties, the most important of which are its compactness and multiresolution (wavelet functions are produced from the mother wavelet by transition and dilation). Wavelet neural networks (WNN) use wavelet functions to decompose the approximated function. However, for a standard wavelet basis with fixed transition and dilation coefficients, the decomposition may be not optimal. If no inverse transformation is needed, the values of transition and dilation coefficients may be determined during network training, and the windows corresponding to various wavelet functions may overlap. In this study, we suggest a new type of a WNN—Adaptive Window WNN (AWWNN), designed primarily for signal processing, in which window positions and wavelet levels are determined with a special iterative procedure. Two modifications of this new type of WNN are tested against linear model and multi-layer perceptron on Mackey-Glass benchmark problem.

[1] *Efitorov A. and Dolenko S.* A New Type of a Wavelet Neural Network // Optical Memory and Neural Networks, 2018. —V.27 p. 152–160. https://doi.org/10.3103/S1060992X18030050.

Алгоритмы планирования в системе поддержки процессов принятия решений для задач логистики

Власов Сергей Евгеньевич[1] vlasov@niisi.ru
Старостин Николай Владимирович[2]⋆ nvstar@iani.unn.ru
Тимофеев Алексей Евгеньевич[2] alexey.timofeev@itmm.unn.ru
[1]Москва, НИИ Системных Исследований Российской Академии Наук
[2]Нижний Новгород, НИИ Механики ННГУ им. Н.И. Лобачевского

Рассматриваются задачи логистики в области железнодорожного транспорта, направленные на обеспечение организации движения подвижного состава в рамках железнодорожной инфраструктуры. Планирование и организация движения поездов является ключевым фактором обеспечения безопасного и эффективного функционирования всей системы железнодорожного транспорта. Применяемый в текущее время подход, основанный на нормативных графиках, не всегда успешно работает на практике, особенно в случае нештатных ситуаций, связанных с изменением пропускных способностей железнодорожной инфраструктуры. В работе строится математическая модель логистики движения подвижного состава с учетом всех ключевых особенностей железнодорожной инфраструктуры. В рамках построенной математической модели ставятся оптимизационные задачи построения графиков работы локомотивов, графиков движения поездов на заданный интервал планирования. Для поставленных задач приводятся алгоритмы решения с оценкой временной сложности.

Работа поддержана федеральной целевой программой «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2014—2020 годы» в рамках контракта No. 14.578.21.0246 (уникальный идентификатор RFMEFI57817X0246).

Planning algorithms in the decision-making support system for logistic problems

Sergey Vlasov[1] vlasov@niisi.ru
Nokolay Starostin[2]⋆ nvstar@iani.unn.ru
AlekseiTimofeev[2] alexey.timofeev@itmm.unn.ru
[1]Moscow, Scientific Research Institute for System Analysis Russian Academy of Sciences
[2]Nizhny Novgorod, Research Institute for Mechanics Lobachevsky State University

Railway transport logistic problems aimed at providing the organization of work of rolling stock in the framework of a railway infrastructure are considered. Planning and organization of railway traffic is a key factor to providing safe and effective functioning of the entire railway transportation system. The currently used approach, based on target plans, is not always successful in practice, especially in emergencies resulting from the fluctuation of traffic capacities of a railway infrastructure. A mathematical model of rolling stock logistics is constructed, accounting for all the key specific features of the railway infrastructure. In the framework of the constructed mathematical model, optimization problems for scheduling the traffic of locomotives and trains for an assigned planning interval are formulated. Solution algorithms, assessing time complexity, are presented for all the formulated problems.

The work is financially supported by the Federal Targeted Program for Research and Development in Priority Areas of Development of the Russian Scientific and Technological Complex for 2014-2020 under the contract No. 14.578.21.0246 (unique identifier RFMEFI57817X0246).

Кластерные срезы в модели ограниченного окружения

Бекларян Армен Левонович[1]⋆ abeklaryan@hse.ru

[1]Москва, Национальный исследовательский университет “Высшая школа экономики”

Рассматривается модель Шеллинга принятия решения по выбору агентом района проживания в зависимости от предпочтений по отношению к своему окружению. Изучается поведение двух групп агентов и возможность достижимости в модели устойчивого положения равновесия. В зависимости от начальных условий и вида кривых распределения порогов толерантности агентов разных типов исследуется возможность перехода в состояние равновесия и возникающие кластерные срезы. Модель реализована в системе имитационного моделирования AnyLogic. Показано, что привнесение ресурсной компоненты меняет саму геометрию кластеров и их метрические характеристики, но свойство устойчивости этих характеристик по начальным данным остается в силе.

[1] *Бекларян А. Л.* Кластерные срезы в модели ограниченного окружения // Машинное обучения и анализ данных, 2019.

Cluster slices in the bounded-neighborhood model

Armen Beklaryan[1]⋆ abeklaryan@hse.ru

[1]Moscow, National Research University Higher School of Economics

We consider the Schelling model of decision making on the agent's choice of the area of residence, depending on preferences in relation to his surroundings. The behavior of two groups of agents and the feasibility of attainability of a stable equilibrium position are studied. Depending on the initial conditions and the type of distribution curves of tolerance thresholds for agents of different types, the possibility of transition to an equilibrium state and arising cluster slices is investigated. The model is implemented in the AnyLogic simulation system. It is shown that the introduction of the resource component changes the geometry of the clusters and their metric characteristics, but the stability of these characteristics according to the initial data remains in force.

[1] *Beklaryan A.* Cluster slices in the bounded-neighborhood model // Machine Learning and Data Analysis, 2019.

Полиномиальная приближённая схема для задачи маршрутизации транспорта с неединичным делимым спросом и ограничением на временные промежутки обслуживания

Хачай Михаил Юрьевич[1,2,3]⋆ mkhachay@imm.uran.ru
Огородников Юрий Юрьевич[1,2] yogorodnikov@imm.uran.ru

[1]Екатеринбург, Институт математики и механики им. Н.Н.Красовского
[2]Екатеринбург, Уральский Федеральный Университет
[3]Омск, Омский Государственный Технический Университет

Задача маршрутизации транспорта с временными промежутками обслуживания (CVRPTW) является широко известной задачей комбинаторной оптимизации, имеющей огромное число приложений в исследовании операций. В отличие от классической постановки задачи CVRP, которая не учитывает временные промежутки обслуживания, аппроксимационные алгоритмы с гарантированными оценками точности для задачи CVRPTW до сих пор мало изучены, даже для случая евклидовой плоскости. В данной работе мы предлагаем, возможно, первую аппроксимационную схему для постановки задачи CVRPTW на плоскости с неединичным делимым спросом, сочетая хорошо известную схему декомпозиции задачи, разработанную A.Adamaszek et el., и квазиполиномиальную приближённую аппроксимационную схему (QPTAS), предложенную L.Song et al. Предложенная нами схема для любого $\varepsilon \in (0, 1)$ находит $(1 + \varepsilon)$-приближённое решение задачи за полиномиальное время при условии, что трудоёмкость q и число временных промежутков обслуживания p не превосходит $2^{\log^{\delta} n}$ для некоторого $\delta = O(\varepsilon)$. Также, данная схема является эффективной полиномиальной (EPTAS) для любых фиксированных значений параметров p и q с субквадратичной трудоёмкостью.

[1] *Khachay, M., Ogorodnikov, Y.* Approximation scheme for the Capacitated Vehicle Routing Problem with Time Windows and non-uniform demand // Mathematical Optimization Theory and Operations Research – 18th International conference (MOTOR 2019). Proceedings, Springer International Publishing, LNCS, V. 11548, 2019. — p. 297–315.

Polynomial Time Approximation Scheme for the CVRP with Time Windows and Non-Uniform Demand

Michael Khachay[1,2,3]⋆ mkhachay@imm.uran.ru
Yuri Ogorodnikov[1,2] yogorodnikov@imm.uran.ru
[1]Ekaterinburg, Krasovsky Institute of Mathematics and Mechanics
[2]Ekaterinburg, Ural Federal University
[3]Omsk,Omsk State Technical University

The Capacitated Vehicle Routing Problem with Time Windows (CVRPTW) is the well-known combinatorial optimization problem having numerous valuable applications in operations research. Unlike the classic CVRP (without time windows constraints), approximation algorithms with theoretical guarantees for the CVRPTW are still developed much less, even for the Euclidean plane. In this paper, perhaps for the first time, we propose an approximation scheme for the planar CVRPTW with non-uniform splittable demand combining the well-known instance decomposition framework by A. Adamaszek et el. and Quasi-Polynomial Time Approximation Scheme (QPTAS) by L. Song et al. Actually, for any $\varepsilon \in (0, 1)$ the scheme proposed finds a $(1 + \varepsilon)$-approximate solution of the problem in polynomial time provided the capacity q and the number p of time windows does not exceed $2^{\log^{\delta} n}$ for some $\delta = O(\varepsilon)$. For any fixed p and q the scheme is Efficient Polynomial Time Approximation Scheme (EPTAS) with subquadratic time complexity.

[1] *Khachay, M., Ogorodnikov, Y.* Approximation scheme for the Capacitated Vehicle Routing Problem with Time Windows and non-uniform demand // Mathematical Optimization Theory and Operations Research – 18th International conference (MOTOR 2019). Proceedings, Springer International Publishing, LNCS, V. 11548, 2019. — p. 297–315.

Вычислительные технологии для сверхбольших оптимизационных задач

Горнов Александр Юрьевич[1]⋆ gornov.a.yu@gmail.com
Аникин Антон Сергеевич[1] anton.anikin@gmail.com
Зароднюк Татьяна Сергеевна[1] tzarodnyuk@gmail.com
Сороковиков Павел Сергеевич[1] sorokovikov.p.s@gmail.com

[1]Иркутск, Институт динамики систем и теории управления СО РАН

Необходимость эффективного решения оптимизационных задач больших и сверхбольших размерностей в последнее время становится все более актуальной проблемой. Подобные huge-scale постановки возникают в самых различных прикладных областях человеческой деятельности, к ним относятся, например, транспортная логистика, создание новых материалов, фармацевтическая промышленность и другие.

В работе предлагаются вычислительные технологии, ориентированные на поиск глобального экстремума в задачах оптимизации сверхбольших размерностей. В качестве базовых алгоритмов, используемых в многометодных вычислительных схемах, используются модификации методов сопряженных градиентов (Хестенса-Штифеля, Флетчера-Ривса, Полака-Поляка-Рибьера и другие) и варианты методов LBFGS и MSBH. Реализованные на языке C/C++ параллельные версии рассматриваемых алго-ритмов направлены на эффективное использование вычислительных мощностей как современных графических ускорителей (GPU), так и традиционных процессоров (CPU).

Работоспособность и быстродействие предлагаемых реализаций были проверены на задачах нахождения PageRank-вектора с различными типами матрицы (диагональными, со случайной структурой, построенными из web-графов тестовой коллекции Стэнфордского университета), задачах поиска низкопотенциальных атомно-молекулярных кластеров Морса, Гупта, Саттона-Чена и других.

Работа поддержана грантом РФФИ №18-07-00587.

[1] *Sorokovikov P. S., Gornov A. Yu., Anikin A. S., Zarodnyuk T. S.* Computational technology for investigating low-potential Gupta clusters of extremely large dimensions // Open Computer Science, Berlin: De Gruyter, 2020 (in print).

Computing technology for huge-scale optimization problems

Alexander Gornov[1]⋆ gornov.a.yu@gmail.com
Anton Anikin[1] anton.anikin@gmail.com
Tatiana Zarodnyuk[1] tzarodnyuk@gmail.com
Pavel Sorokovikov[1] sorokovikov.p.s@gmail.com
[1]Irkutsk, Matrosov Institute for System Dynamics and Control Theory SB RAS

The large- and huge-scale optimization problems have recently become an increasingly urgent problem. Such statements arise in a wide variety of applied areas of human activity, such as, for example, transport logics, the creation of new materials, the pharmaceutical industry and others.

The report offers numerical technologies focused on the search for a global extremum in huge-scale optimization problems. As the basic algorithms used in multi-method computational schemes, modifications of conjugate gradient methods (Hestenes-Stiefel, Fletcher-Reeves, Polak-Polyak-Ribiere and others) and variants of the LBFGS and MSBH methods are implemented. Parallel versions of the considered algorithms realized in C/C++ are aimed at the efficient use of computing power of modern graphic accelerators (GPUs) and traditional processors (CPUs).

The operability and speed of the proposed algorithms were tested on the problems of finding a PageRank vector with various types of matrices (diagonal, with a random structure built from web graphs of the Stanford University test collection), problems of searching low-potential atomic-molecular clusters Morse, Gupta, Sutton-Chen, and others.

This research is funded by RFBR, grant 18-07-00587.

[1] *Sorokovikov P. S., Gornov A. Yu., Anikin A. S., Zarodnyuk T. S.* Computational technology for investigating low-potential Gupta clusters of extremely large dimensions // Open Computer Science, Berlin: De Gruyter, 2020 (in print).

Применение эволюционных методов в задаче распознавания периодических решений и резонансов динамических систем

Ручкин Константин[1]⋆ construchk@gmail.com

[1]Донецк, Донецкий Национальный Технический Университет

В данной статье рассматривается задача нахождения периодических решений и резонансов динамических систем. Исследование периодических решений динамических систем проводится путем анализа сечений Пуанкаре на плоскости на наличие замкнутых траекторий специального типа на этих сечениях [1]. В большинстве случаев исследуемые траектории представляют собой особые геометрических формы, такие как круги, эллипсы и др. Для распознавания таких фигур предлагается использовать эволюционный метод вычислительного интеллекта - адаптивный алгоритм бактериального поиска для задачи стохастической глобальной оптимизации. Разработанные за последние несколько десятилетий в компьютерном зрении подходы способны обнаруживать несколько одинаковых кругов на реальных изображениях, но часто не в состоянии обнаружить пересекающиеся и несовершенные формы.

В работе мы используем алгоритм бактериальной оптимизации бактериального (BFOA), для определения множественных форм. Предлагается адаптивная версия BFOA для поиска накрадывающихся и не накладывающихся форм по всему изображению. Каждая бактерия здесь моделирует пробную форму, и в области таких пробных форм была получена нечеткая целевая функция. Чем лучше тестовая окружность приближается к фактической окружности, тем меньше значение этой функции. Минимизация целевой функции с помощью BFOA в конечном итоге приводит к быстрому и надежному детектированию заданных форм на исходном изображении. Применение этого метода к исследованию динамических систем позволяет говорить об их регулярном или квазирегулярном поведении.

[1] *Ruchkin C.* The General Conception of the Intellectual Investigation of the Regular and Chaotic Behavior of the Dynamical System Hamiltonian Structure. // Applied Non-Linear Dynamical Systems. Springer Proceedings in Mathematics and Statistics, vol 93., Springer, Cham, 2014. *Ручкин К.А., Миньков О. В.* Разработка алгоритма распознавания сложных накладывающихся геометрических объектов // Информатика и кибернетика, Вып. №2 (8), 2017, с. 65 – 72

Application of evolutionary methods in the problem recognition of periodic solutions and resonances of dynamical systems

Constantin Ruchkin[1]⋆ construchk@gmail.com

[1]Donetsk, Donetsk National Technical University

In this paper we consider the problem of investigation periodic solutions and resonances of dynamical systems. The study of periodic solutions of dynamical systems is carried out by means of analysis of the Poincare sections on the plane for the presence of closed trajectories of a special type on these sections [1]. In most cases, the trajectories under investigation represent special geometric shapes like circles, an ellipse, etc. To solve the problem of recognizing such forms, it is proposed to use the evolutionary method of computational intelligence - the generalized algorithm of bacterial search for stochastic global optimization. The problem circle and ellipse detection from digital images by means computational intelligence have received considerable attention over the last few decades in computer vision. So exist approaches is capable of detecting multiple circles on real images but fails frequently to detect intersecting and imperfect shapes.

In this work, we use the bacterial bacterial optimization algorithm (BFOA) to determine multiple forms. An adaptive version of BFOA is offered to search for imperfect and non-overlapping shapes throughout the image. Each bacterium here models a test form, and a fuzzy objective function was obtained in the field of such test forms. The better the test form approaches the actual form, the lower the value of this function. Minimization of the objective function using BFOA ultimately leads to the rapid and reliable detection of given shapes in the original image. The article also discusses the conditions of applicability of this approach for the detection of various geometric objects on the Poincare section. Application of this method to the study of dynamical systems allows us to talk about their regular or quasiregular behavior.

[1] *Ruchkin C.* The General Conception of the Intellectual Investigation of the Regular and Chaotic Behavior of the Dynamical System Hamiltonian Structure. // Applied Non-Linear Dynamical Systems. Springer Proceedings in Mathematics and Statistics, vol 93., Springer, Cham, 2014.

[2] *Ruchkin C., Minkov O.* Development of an algorithm for recognizing complex overlapping geometric objects // Computer Science and Cybernetics, Vol. No 2(8), 2017, Pp. 65 – 72

Сложность вычисления: решённые задачи и открытые проблемы

Карацуба Екатерина Анатольевна[1]⋆ ekaratsuba@gmail.com

[1]Москва, Вычислительный центр им. А. А. Дородницына ФИЦ ИУ РАН

Понятие сложности вычисления появилось в результате развития вычислительных методов и теории информации. Современные основы теории сложности вычислений в информатике были заложены работами Г. Найквиста [1] и Р. Хартли [2], с введением понятия меры информации. Первые постановки задач о битовой сложности вычислений (1956 г.) принадлежат А.Н. Колмогорову [3].

Далее будем считать, что числа записаны в двоичной системе счисления, знаки которой 0 и 1 называются битами.

Опр. 1. Запись знаков 0, 1 , плюс, минус, скобка; сложение, вычитание и умножение двух битов назовём одной элементарной или битовой операцией. Пусть вещественная функция $f(x)$ вещественного переменного x, $a \leqslant x \leqslant b$, удовлетворяет на (a, b) условию Липшица порядка α, $0 < \alpha < 1$, так что при $x_1, x_2 \in (a, b)$: $|f(x_1) - f(x_2)| \leqslant |x_1 - x_2|^\alpha$. Пусть n — натуральное число.

Опр. 2. Вычислить функцию $y = f(x)$ в точке $x = x_0 \in (a, b)$ с точностью до n знаков, значит найти такое число A, что $|f(x_0) - A| \leqslant 2^{-n}$.

Опр. 3. Количество битовых операций, достаточное для вычисления функции $f(x)$ в точке $x = x_0$ с точностью до n знаков посредством данного алгоритма, называется сложностью (битовой) вычисления $f(x)$ в точке $x = x_0$.

Таким образом, сложность вычисления $f(x)$ в точке $x = x_0$ есть функция n, а также $f(x)$ и $x = x_0$. Эту функцию обозначают символом $S_{f(n)} = S_{f,x_0(n)}$. Ясно, что S_f зависит также от алгоритма вычисления и при разных алгоритмах будет разной. Сложность вычисления непосредственно связана со временем, затрачиваемым компьютером на это вычисление и потому иногда в литературе (например, в книге "Искусство программирования на ЭВМ"Д. Кнута) обозначается «временной» функцией $T(n)$.

До сих пор построение теории (битовой) сложности вычислений (включая основные определения и понятия) не завершено. Кроме того, история показала, что некоторые естественные логические умозаключения в этой области не работают: в 50-х гг. А.Н. Колмогоров высказал гипотезу, что нижняя оценка сложности умножения $M(n)$ при любом методе умножения есть величина порядка n^2 («гипотеза n^2 Колмогорова»), на том основании, что все известные к тому времени методы умножения имеют сложность n^2, используются не менее 4-х тысячелетий, и если бы был более быстрый метод умножения, то он, вероятно, уже был бы найден. Тем не менее, в 1960 г. [4] был найден новый метод умножения двух n-значных чисел с оценкой сложности $M(n) = O(n^{\log_2 3}), \log_2 3 = 1,5849\ldots$, опровергая гипотезу n^2. С момента построения этого метода умножения началась теория быстрых вычислений, и было построено множество быстрых алгоритмов обычного и матричного умножений,

Фурье-преобразований и вычислений элементарных и высших трансцендентных функций и классических констант.

В 1991 автор построил [5] метод БВЕ (Быстрого Вычисления Е-функций) – метод быстрого суммирования специального вида рядов, который позволяет вычислить любую элементарную трансцендентную функцию для любого аргумента, классические константы e, π, постоянную Эйлера γ, постоянные Апери и Каталана, такие высшие трансцендентные функции, как гамма-функцию Эйлера, гипергеометрические функции, сферические функции, цилиндрические функции и т. д. для алгебраических значений аргумента и параметров, дзета-функцию Римана для целых значений аргумента, дзета-функцию Гурвица для целого аргумента и алгебраических значений параметра, а также такие специальные интегралы, как интеграл вероятности, интегралы Френеля, интегральную экспоненциальную функцию, интегральные синус и косинус и т. д. при алгебраических значениях аргумента с оценкой сложности вычисления, близкой к оптимальной, а именно $S_f(n) = O(M(n)\log^2 n)$. Дополнительным преимуществом метода является возможность распараллеливания основанных на БВЕ алгоритмов.

Построение алгоритмов вычисления широкого класса функций с оценкой битовой сложности, близкой к оптимальной, а также получение нетривиальных нижних оценок битовой сложности – основные задачи в той области вычислительной математики, которая называется быстрые алгоритмы или быстрые вычисления. В настоящее время (ноябрь 2019) здесь остаётся много нерешённых проблем, таких как

1. получение нетривиальной оценки снизу сложности умножения или сложности вычисления трансцендентных функций;
2. построение быстрых алгоритмов вычисления высших трансцендентных функций в трансцендентных точках;
3. построение быстрых алгоритмов вычисления таких констант, как константа Бруна, значения дзета-функции Римана в нецелых точках и т.д.
4. оценка сложности вычисления решений систем дифференциальных, интегро-дифференциальных, матричных и т.п. уравнений, когда решение не выписывается конечной комбинацией известных трансцендентных функций.

В то же время существуют и более общие проблемы теории сложности вычисления, скажем, как оценивать эффективно сложность вычисления решений задач посредством слабо-структурированных методов, при которых заранее нельзя гарантировать определённую точность вычисления (множество методов классификации и распознавания), или, как адаптировать теорию битовой сложности (предполагающей бесконечную память компьютера) на реальные технические ограничения.

[1] *H. Nyquist* Certain factors affecting telegraph speed // Bell System Technical Journal, 3, 324-346 (1924).

[2] *R. Hartley* Transmission of Information // Bell System Technical Journal, 7, 535-563 (1928).

[3] *Колмогоров А. Н.* О некоторых асимптотических характеристиках вполне ограниченных метрических пространств // Доклады Академии Наук СССР, 108:3, 385-388 (1956).

[4] *Карацуба А., Офман Ю.* Умножение многозначных чисел на автоматах // Доклады Академии Наук СССР, 145: 2, 293-294 (1962).

[5] *Карацуба Е. А.* Быстрое вычисление трансцендентных функций // Проблемы передачи информации, 27:4, 87-110 (1991).

[6] *Карацуба Е. А.* Быстрые аппроксимации некоторых теоретико-числовых констант // Доклады Академии Наук, 462:2, 137-140 (2015).

The complexity of the calculations: Solved problems and apened questions

Ekaterina Karatsuba[1]⋆ ekaratsuba@gmail.com

[1]Moscow, FRCCSC of the Russian Academy of Sciences

The concept of the complexity of computation appeared as a result of development of computational methods and information theory. The modern foundations of the theory of computational complexity in Computer Science were laid by works of H. Nyquist [1] and R. Hartley [2], with introduction of the idea of measure of information. The first statements of the problems of estimation of the bit complexity of computation (1956) belong to A.N. Kolmogorov [3].

Further, we assume that numbers are written in the binary notation, the signs of which 0 and 1 are called bits.

Def. 1. Writing a symbol 0, 1, plus, minus, bracket; addition, subtraction and multiplication of two bits will be called one elementary or bit operation. Let real function $y = f(x)$ of the real variable x, $a \leqslant x \leqslant b$ satisfy on (a, b) the Lipschitz condition of order α, $0 < \alpha < 1$, so for $x_1, x_2 \in (a, b)$: $|f(x_1) - f(x_2)| \leqslant |x_1 - x_2|^\alpha$. Let n be an integer, $n \geqslant 1$.

Def. 2. To compute the function $y = f(x)$ at the point $x = x_0 \in (a, b)$ up to n digits, means to find a number A such that $|f(x_0) - A| \leqslant 2^{-n}$.

Def. 3. The total number of bit operations sufficient to compute the function $f(x)$ at the point $x = x_0$ with accuracy up to n digits by using a certain algorithm is called the (bit) complexity of computation of $f(x)$ at the point $x = x_0$.

Thus, the complexity of computation of $f(x)$ at the point $x = x_0$ is a function of n, as well as $f(x)$ and $x = x_0$. This function is denoted by $S_{f(n)} = S_{f,x_0(n)}$. It is clear that S_f also depends on the computational algorithm and will be different for different algorithms. The complexity of computation is directly related to the time spent by the computer on this calculation and therefore sometimes in the literature (for example, in the book "The Art of Computer Programming" by D. Knuth) is denoted by the "time" function $T(n)$.

Till the present the construction of the theory of (bit) complexity of computation (including basic definitions and concepts) is still not completed. In addition, history has shown that some natural logical conclusions in this field do not work. In the 50s A.N. Kolmogorov hypothesized, that the lower bound for the complexity of multiplication $M(n)$ for any method of multiplication is of the order of n^2 ("Kolmogorov hypothesis n^{2}"), on the basis of such idea: since all the methods of multiplication known to that time have the complexity no better than n^2, and they are used at least 4 millennia, so that, if there was a faster method of multiplication, then it would probably have already been found.

However, in 1960 [4] a new method was found for multiplying two n -digital integers ($n \to +\infty$) with the complexity bound $M(n) = O(n^{\log_2 3})$, $\log_2 3 = 1,5849\ldots$, disproving the hypothesis n^2. From the moment of creation of this method of mul-

tiplication, the theory of fast computations began, and many fast algorithms of ordinary multiplication and of matrix multiplication, Fourier transforms and calculations of elementary and higher transcendental functions and classical constants were developed.

In 1991, the author constructed (see [5]) the FEE (Fast Evaluation of E-Functions) method – a method for fast summation of a special kind of series that allows to calculate any elementary transcendental function for any argument, the classical constants e, π, the Euler constant γ, the Apéry and Catalan constants, higher transcendental functions such as the Euler gamma function, hypergeometric functions, spherical functions, cylindrical functions, etc. for algebraic values of argument and parameters, the Riemann zeta function for integer argument, the Hurwitz zeta function for integer argument and algebraic parameter, as well as such special integrals as the probability integral, Fresnel integrals, the integral exponential function integral sine , integral cosine etc. for algebraic argument with the complexity bound, close to the optimal one, namely $S_f(n) = O(M(n)\log^2 n)$. An additional advantage of the method is the ability to parallelize FEE-based algorithms.

The construction of algorithms for computing a wide class of functions with near-optimal complexity bounds, as well as obtaining non-trivial lower bounds for the complexity of computation are the main problems in the field of the computational mathematics which is called Fast Algorithms or Fast Computations. At present (2019, November), many unsolved problems remain in this field, such as

1. obtaining a non-trivial lower bound for the complexity of multiplication or the complexity of computing a transcendental function;
2. constructing fast algorithms for calculation of a higher transcendental function at a transcendental point;
3. constructing fast algorithms for calculation constants such as the Brun constant, the values of the Riemann zeta function at non-integer points, etc.
4. estimating the complexity of computation of solutions of systems of differential, integro-differential, matrix, etc. equations, when it's impossible to derive the solution in the form of a finite combination of known transcendental functions.

At the same time, there are also more general unanswered questions in the theory of complexity of computation, for example, how to estimate effectively the complexity of calculating the solutions obtained by means of a weakly-structured methods for which there is no the advance guarantee of certain accuracy of computation (many methods of classification and recognition), or how to adapt the theory of bit complexity (involving infinite computer memory) for real technical limitations.

[1] *H. Nyquist* Certain factors affecting telegraph speed // Bell System Technical Journal, 3, 324-346 (1924).

[2] *R. Hartley* Transmission of Information // Bell System Technical Journal, 7, 535-563 (1928).

[3] *Kolmogorov A. N.* On some asymptotic characteristics of completely bounded metric spaces // Reports of the Academy of Sciences of the USSR, 108:3, 385-388 (1956).

[4] *Karatsuba A., Ofman Yu.* Multiplication of multivalued numbers on automata // Reports of the Academy of Sciences of the USSR, 145: 2, 293-294 (1962).

[5] *Karatsuba E. A.* Fast calculation of transcendental functions // Information Transmission Problems, 27:4, 87-110 (1991).

[6] *Karatsuba E. A.* Fast approximations of some number-theoretic constants // Reports of the Academy of Sciences of the USSR, 462:2, 137-140 (2015).

Неизученные задачи Data Mining: сложность и аппроксимируемость

Кельманов Александр Васильевич[1,2]⋆ kelm@math.nsc.ru
Пяткин Артем Валерьевич[1,2] artem@math.nsc.ru
Хандеев Владимир Ильич[1,2] khandeev@math.nsc.ru
[1]Новосибирск, Институт математики им. С.Л.Соболева
[2]Новосибирск, Новосибирский государственный университет

Рассматриваются несколько недавно выявленных задач дискретной оптимизации, которые индуцируются проблемой Data mining. Получены результаты о вычислительной сложности задач, раскрыты некоторые вопросы аппроксимируемости этих задач.

Выяснение структуры данных с помощью так называемого разведочного поиска подходящего (адекватного) описания (т.е. интерпретации) данных в виде модели порождения данных типично для прикладных проблем Data mining и математической статистики. В классической статистике, в отличие от Data mining, предполагается, что данные однородны, т.е. являются выборкой из одного распределения. Напротив, в Data mining предполагается, что данные неоднородны, т.е. являются выборкой из нескольких распределений, причем априорное соответствие данных распределениям неизвестно. Отсутствие этого соответствия обуславливает создание математических инструментов в виде эффективных алгоритмов решения необозримого множества задач разбиения данных с самой разнообразной структурой на однородные по какому-либо фиксированному критерию кластеры, а также инструментов в виде критериев проверки адекватности аппроксимационных моделей разбиения имеющимся данным. Например, чтобы выяснить, какая из сформулированных ниже задач (моделей аппроксимации) разбиения адекватна данным (входному множеству точек) или ни одна из них не адекватна данным, в первую очередь необходимы эффективные алгоритмы решения этих кластеризационных задач. Очевидно, что создание эффективных в вычислительном плане алгоритмов является одной из ключевых проблем для Data mining. В свою очередь, создание таких алгоритмов обуславливает исследование сложностного статуса задач разбиения. Приведенные замечания поясняют мотивацию настоящего исследования. Фактически, наша работа отвечает на вопрос — можно ли эффективно (за полиномиальное время) разбить имеющиеся данные в соответствии целевыми функциями сформулированных ниже задач.

Задача 1 (*Сбалансированное 2-разбиение по дисперсионному критерию*). *Дано*: N-элементное множество $\mathcal{Y}$ точек в евклидовом пространстве размерности d и некоторое вещественное число $\varepsilon > 0$. *Вопрос*: существует ли такое

разбиение множества $\mathcal{Y}$ на непустые кластеры $\mathcal{C}$ и $\mathcal{Y} \setminus \mathcal{C}$, что

$$\Big| \frac{1}{|\mathcal{C}|} \sum_{y \in \mathcal{C}} \|y - \overline{y}(\mathcal{C})\|^2 - \frac{1}{|\mathcal{Y} \setminus \mathcal{C}|} \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y - \overline{y}(\mathcal{Y} \setminus \mathcal{C})\|^2 \Big| \leqslant \varepsilon \,,$$

где $\overline{y}(\mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum_{y \in \mathcal{C}} y$ и $\overline{y}(\mathcal{Y} \setminus \mathcal{C}) = \frac{1}{|\mathcal{Y} \setminus \mathcal{C}|} \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} y$ — центроиды (геометрические центры) кластеров $\mathcal{C}$ и $\mathcal{Y} \setminus \mathcal{C}$ соответственно?

Задача 2 (*Сбалансированное 2-разбиение по критерию суммарного квадратичного разброса*). *Дано*: N-элементное множество $\mathcal{Y}$ точек в евклидовом пространстве размерности d и некоторое вещественное число $\varepsilon > 0$. *Вопрос*: существует ли такое разбиение множества $\mathcal{Y}$ на непустые кластеры $\mathcal{C}$ и $\mathcal{Y} \setminus \mathcal{C}$, что

$$\Big| \sum_{y \in \mathcal{C}} \|y - \overline{y}(\mathcal{C})\|^2 - \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y - \overline{y}(\mathcal{Y} \setminus \mathcal{C})\|^2 \Big| \leqslant \varepsilon \,?$$

Задача 3 (*Сбалансированное 2-разбиение по критерию мощ-ностно-взвешенного суммарного квадратичного разброса*). *Дано*: N-элементное множество $\mathcal{Y}$ точек в евклидовом пространстве размерности d и некоторое вещественное число $\varepsilon > 0$. *Вопрос*: существует ли такое разбиение множества $\mathcal{Y}$ на непустые кластеры $\mathcal{C}$ и $\mathcal{Y} \setminus \mathcal{C}$, что

$$\Big| |\mathcal{C}| \sum_{y \in \mathcal{C}} \|y - \overline{y}(\mathcal{C})\|^2 - |\mathcal{Y} \setminus \mathcal{C}| \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y - \overline{y}(\mathcal{Y} \setminus \mathcal{C})\|^2 \Big| \leqslant \varepsilon \,?$$

Доказано [1], что задачи 1–3 NP-полны. Из этого результата следует NP-трудность оптимизационных вариантов этих задач.

Задача 4 (*Quadratic* 1-*Mean and* 1-*Median* 2-*Clustering with the Constraints on the Cluster Sizes*). *Дано:* N-элементное множество $\mathcal{Y}$ точек в d-мерном евклидовом пространстве и натуральное число M. *Найти:* такие точку $x \in \mathcal{Y}$ и разбиение $\mathcal{Y}$ на кластеры $\mathcal{C}$ и $\mathcal{Y} \setminus \mathcal{C}$ размеров M и $N - M$ соответственно, что

$$\sum_{y \in \mathcal{C}} \|y - \overline{y}(\mathcal{C})\|^2 + \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y - x\|^2 \to \min \,.$$

Доказано [2], что задача NP-трудна в сильном смысле. Установлено, что для нее не существует полностью полиномиальная аппроксимационная схема, если P$\neq$ NP.

Очевидно, что рассмотренные задачи 1–3 можно обобщить на случай, когда число кластеров больше, чем 2. Ясно, что когда число кластеров является частью входа, эти обобщения тоже являются труднорешаемыми задачами. Вопрос о статусе сложности параметрического случая этих задач, когда число кластеров не является частью входа, остается открытым. Сказанное справедливо и для многокластерного обобщения задачи 4.

Построение эффективных приближенных алгоритмов с гарантированными оценками точности для задач 1–4 является делом ближайшей перспективы.

Работа выполнена при финансовой поддержке РФФИ, проекты 19-01-0030 и 18-31-00398, программы ФНИ РАН, проекты 0314-2019-0014, 0314-2019-0015, а также программы Top-5-100 Министерства образования и науки РФ.

[1] *Кельманов А.В., Пяткин А.В., Хандеев В.И.* NP-полнота некоторых задач разбиения конечного множества точек евклидова пространства на сбалансированные кластеры // Доклады РАН, 2019. Т. 488, № 1, с. 595–599.

[2] *Кельманов А.В., Пяткин А.В., Хандеев В.И.* NP-трудность квадратичной евклидовой задачи 2-кластеризации 1-Mean and 1-Median с ограничениями на размеры кластеров // Доклады РАН, 2019. Т. 489, № 4, с. 1–4 (accepted).

Some Unexplored Data Mining Problems: Complexity and Approximability

Kel'manov Alexander[1,2]⋆ kelm@math.nsc.ru
Pyatkin Artem[1,2] artem@math.nsc.ru
Khandeev Vladimir[1,2] khandeev@math.nsc.ru
[1]Novosibirsk, Sobolev Institute of Mathematics
[2]Novosibirsk, Novosibirsk State University

The paper introduces some unexplored Data mining problems and corresponding discrete optimization problems. The computational complexity of the problems is investigated. Some issues of approximability of these problems are considered.

Clarification of the data structure by means of the so-called exploratory search for a suitable (adequate) description (i.e. interpretation) of the data in the form of a data generation model is typical for applied problems of Data mining and mathematical statistics. In classical statistics, unlike Data mining, it is assumed that the data is homogeneous, i.e. it is a sample from a single distribution. In contrast, in Data mining it is assumed that the data is heterogeneous (is a sample from several distributions), and the a priori correspondence between the data and the distributions is unknown. The absence of this correspondence makes it necessary to create: (1) mathematical tools in the form of effective algorithms for solving an immense number of problems of partitioning data with the most diverse structure into clusters which are homogeneous by some fixed criterion; (2) tools in the form of criteria for checking the adequacy of the approximation partition models to the available data. For example, finding out which of the following partition problems (approximation models) is adequate to the data (input set of points) or proving that none of them is adequate definitely requires effective algorithms for solving these clustering problems. It is obvious that creation of computationally efficient algorithms is one of the key Data mining problems. This, in turn, requires studying the complexity status of the partition problems. These arguments motivate this research. In fact, our paper establishes whether it is possible to partition the available data according to the objective functions of the problems formulated below efficiently (in polynomial time).

The problems under consideration are as follows.

Problem 1 *(Balanced 2-partition by the criterion of the normalized by a cluster size sum of squared deviations from the mean). Given*: N-element set $\mathcal{Y}$ of points in d-dimensional Euclidean space and a real number $\varepsilon > 0$. *Question*: is there a partition of $\mathcal{Y}$ into non-empty clusters $\mathcal{C}$ and $\mathcal{Y} \setminus \mathcal{C}$ such that

$$\Big| \frac{1}{|\mathcal{C}|} \sum_{y \in \mathcal{C}} \|y - \overline{y}(\mathcal{C})\|^2 - \frac{1}{|\mathcal{Y} \setminus \mathcal{C}|} \sum_{y \in \mathcal{Y} \setminus \mathcal{C}} \|y - \overline{y}(\mathcal{Y} \setminus \mathcal{C})\|^2 \Big| \leqslant \varepsilon \,,$$

where $\overline{y}(\mathcal{C}) = \frac{1}{|\mathcal{C}|}\sum_{y\in\mathcal{C}} y$ and are the centroids (geometric centers) of the clusters $\mathcal{C}$ and $\mathcal{Y}\setminus\mathcal{C}$, respectively?

Problem 2 *(Balanced 2-partition by the criterion of the sum of squared deviations from the mean). Given*: N-element set $\mathcal{Y}$ of points in Euclidean space of dimension d and a real number $\varepsilon > 0$. *Question*: is there a partition of $\mathcal{Y}$ into non-empty clusters $\mathcal{C}$ and $\mathcal{Y}\setminus\mathcal{C}$ such that

$$\Big|\sum_{y\in\mathcal{C}} \|y-\overline{y}(\mathcal{C})\|^2 - \sum_{y\in\mathcal{Y}\setminus\mathcal{C}} \|y-\overline{y}(\mathcal{Y}\setminus\mathcal{C})\|^2\Big| \leqslant \varepsilon\,?$$

Problem 3 *(Balanced 2-partition by the criterion of the size-weighted sum of squared deviations from the mean). Given*: N-element set $\mathcal{Y}$ of points in Euclidean space of dimension d and a real number $\varepsilon > 0$. *Question*: is there a partition of $\mathcal{Y}$ into non-empty clusters $\mathcal{C}$ and $\mathcal{Y}\setminus\mathcal{C}$ such that

$$\Big||\mathcal{C}|\sum_{y\in\mathcal{C}} \|y-\overline{y}(\mathcal{C})\|^2 - |\mathcal{Y}\setminus\mathcal{C}|\sum_{y\in\mathcal{Y}\setminus\mathcal{C}} \|y-\overline{y}(\mathcal{Y}\setminus\mathcal{C})\|^2\Big| \leqslant \varepsilon\,?$$

We have proved [1] that Problems 1–3 are NP-complete. This result implies NP-hardness of optimization variants of these problems.

Problem 4 *(Quadratic* 1*-Mean and* 1*-Median* 2*-Clustering with the Constraints on the Cluster Sizes). Given*: N-element set $\mathcal{Y}$ of points in d-dimensional Euclidean space and a positive integer number M. *Find*: a point $x\in\mathcal{Y}$ and a partition of $\mathcal{Y}$ into clusters $\mathcal{C}$ and $\mathcal{Y}\setminus\mathcal{C}$ of sizes M and $N-M$, respectively, such that

$$\sum_{y\in\mathcal{C}} \|y-\overline{y}(\mathcal{C})\|^2 + \sum_{y\in\mathcal{Y}\setminus\mathcal{C}} \|y-x\|^2 \to \min\,.$$

We have proved [2] the strong NP-hardness of this problem and nonexistence of the fully polynomial-time approximation scheme unless P=NP.

Obviously, the considered Problems 1–3 can be generalized to the case when the number of clusters is greater than 2. It is clear that when the number of clusters is a part of the input, these generalizations are also intractable problems. The question of the complexity status of the parametric cases of these problems (when the number of clusters is not a part of the input) remains open. The same is true for a multicluster generalization of Problem 4.

The construction of efficient approximate algorithms with guaranteed accuracy bounds for Problems 1–4 will be addressed in the nearest future.

The research was supported by the Russian Foundation for Basic Research, projects 19-01-00308 and 18-31-00398, by the Russian Academy of Science (the Program of basic research), projects 0314-2019-0015 and 0314-2019-0014, and by the Russian Ministry of Science and Education under the 5-100 Excellence Programme.

[1] *Kel'manov A., Pyatkin A.V., Khandeev V.I.* NP-Completeness of Some Problems of Partitioning a Finite Set of Points in Euclidean Space into Balanced Clusters // Doklady Mathematics, 2019, Vol. 100, No 2, pp. 1–4 (accepted).

[2] *Kel'manov A., Pyatkin A.V., Khandeev V.I.* NP-Hardness of Quadratic Euclidean 1-Mean and 1-Median 2-Clustering Problem with the Constraints on the Cluster Sizes // Doklady Mathematics, 2019 (accepted).

Задача минимизации суммы разностей взвешенных сверток и новый подход к обработке и анализу ECG- и PPG-сигналов

Кельманов Александр Васильевич[1,2]⋆ kelm@math.nsc.ru
Михайлова Людмила Викторовна[1] mikh@math.nsc.ru
Рузанкин Павел Сергеевич[1,2] ruzankin@math.nsc.ru
Хамидуллин Сергей Асгадуллович[1] kham@math.nsc.ru
[1]Новосибирск, Институт математики им. С.Л. Соболева
[2]Новосибирск, Новосибирский государственный университет

Рассматривается неизученная задача суммирования элементов числовых последовательностей Y длины N и U длины $q \leqslant N$. Задача индуцируется новым подходом к помехоустойчивой обработке квазипериодически повторяющихся последовательностей импульсов при наличии нелинейно-временных флуктуаций некоторого эталонного импульса в каждом повторе. Подобные изменчивые последовательности характерны, в частности, для ECG- и PPG-сигналов. Мы рассматриваем два варианта общей прикладной проблемы. Первый из них индуцирует следующую задачу [1].

Задача 1. *Дано:* числовые последовательности $Y = (y_1, \ldots, y_N)$, $U = (u_1, \ldots, u_q)$, и натуральные числа $T_{\max}$, ℓ. *Найти:* набор $\mathcal{M} = \{n_1, \ldots n_m, \ldots\}$ номеров последовательности Y, набор $\mathcal{P} = \{p^{(1)}, \ldots, p^{(m)}, \ldots\}$ натуральных чисел, набор $\mathcal{J} = \{J^{(1)}, \ldots, J^{(m)}, \ldots\}$ сжимающих отображений, в котором $J^{(m)} : \{1, \ldots, p^{(m)}\} \longrightarrow \{1, \ldots, q\}$, а также размерность M этих наборов, которые минимизируют целевую функцию

$$F(\mathcal{M}, \mathcal{P}, \mathcal{J}) = \sum_{m=1}^{M} \sum_{i=1}^{p^{(m)}} \{u^2_{J^{(m)}(i)} - 2y_{n_m+i-1} u_{J^{(m)}(i)}\},$$

при ограничениях

$$q \leqslant p^{(m)} \leqslant \ell \leqslant T_{\max} \leqslant N, \;\; m = 1, \ldots, M,$$
$$p^{(m-1)} \leqslant n_m - n_{m-1} \leqslant T_{\max}, \;\; m = 2, \ldots, M,$$
$$p^{(M)} \leqslant N - n_M + 1,$$

на элементы искомых наборов $\mathcal{M}$, $\mathcal{P}$, и при ограничениях

$$J^{(m)}(1) = 1, \quad J^{(m)}(p^{(m)}) = q,$$
$$0 \leqslant J^{(m)}(i) - J^{(m)}(i-1) \leqslant 1, \quad i = 2, \ldots, p^{(m)},$$
$$m = 1, \ldots, M,$$

на элементы искомых сжимающих отображений.

Второй вариант общей проблемы индуцирует **Задачу 2**, в которой на входе дополнительно задано натуральное число $M \leqslant \lfloor \frac{N}{q} \rfloor$, т.е., в отличие от задачи 1, в задаче 2 размерность искомых наборов задана на входе [2].

Основной математический результат работы:

Теорема 1. Существуют алгоритмы, которые находят точное решение задач 1 и 2 за время $\mathcal{O}(T_{\max}^3 N)$ и $\mathcal{O}(T_{\max}^3 MN)$.

Доказательство теоремы конструктивно: мы даем прямой вывод двух схем динамического программирования и доказываем, что они гарантируют отыскание точного решения каждой задачи за указанное в теореме полиномиальное время. При этом, если значение $T_{\max}$ ограничено константой (фиксировано), то время работы алгоритмов равно $\mathcal{O}(N)$ и $\mathcal{O}(MN)$.

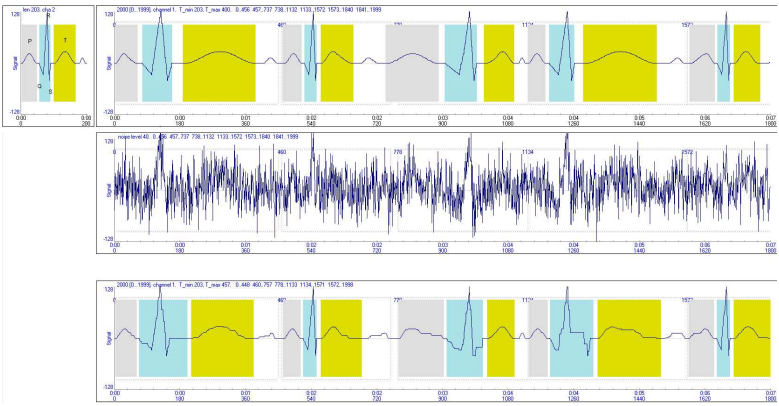


Рис. 1. Пример обработки ECG-подобного сигнала

На рис. 1 и 2 приведены примеры помехоустойчивой обработки смоделированных сигналов. В верхней части каждого из рисунков 1 и 2 изображены входная последовательность U (импульс) и недоступная для алгоритма смоделированная последовательность, соответствующая последовательности флуктуирующих импульсов. В средней части рисунков представлена вторая входная последовательность Y, т.е. доступный для обработки сигнал. В нижней части — последовательность, полученная в результате работы алгоритма, т.е. восстановленный сигнал.

Полученные результаты: показана полиномиальная разрешимость двух новых задач дискретной оптимизации, предложен новый подход к обработке флуктуирующих во времени биомедицинских импульсных сигналов.

Работа выполнена при финансовой поддержке РФФИ, проекты 19-07-00397 и 19-01-00308, программы ФНИ РАН, проект 0314-2019-0015, а также программы Top-5-100 Министерства образования и науки РФ.

[1] *Кельманов А. В., Михайлова Л.В., Рузанкин П.С., Хамидуллин С.А.* Задача минимизации суммы разностей взвешенных сверток // Журн. вычисл. математики и мат. физики, 2019 (accepted).

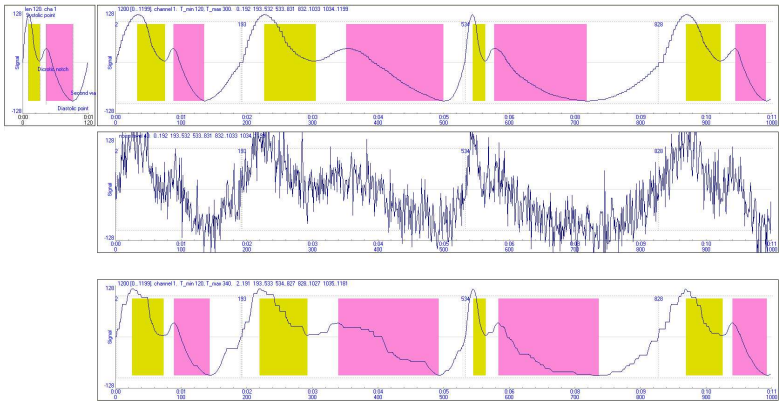


Рис. 2. Пример обработки PPG-подобного сигнала

[2] *Кельманов А. В., Михайлова Л.В., Рузанкин П.С., Хамидуллин С.А.* Задача минимизации суммы разностей взвешенных сверток, случай заданного числа элементов в сумме // Сиб. журн. вычисл. математики, 2019 (accepted).

A minimization problem for the sum of weighted convolutions' difference and a novel approach to the processing and analysis of ECG and PPG signals

Alexander Kel'manov[1,2]⋆ kelm@math.nsc.ru
Liudmila Mikhailova[1] mikh@math.nsc.ru
Pavel Ruzankin[1,2] ruzankin@math.nsc.ru
Sergey Khamidullin[1] kham@math.nsc.ru
[1]Novosibirsk, Sobolev Institute of Mathematics
[2]Novosibirsk, Novosibirsk State University

We consider an unexplored problem of summing the elements of numeric sequences Y and U of lengths N and $q \leqslant N$, respectively. The problem is induced by a novel approach to noise-robust processing of quasiperiodic pulse trains in the presence of nonlinear temporal fluctuations of a certain reference pulse in each repetition. Such quasiperiodic sequences are characteristic, in particular, for ECG and PPG signals. We consider two variants of the common application problem. The first variant induces the following problem [1].

Problem 1. *Given:* some numeric sequences $Y = (y_1, \ldots, y_N)$, $U = (u_1, \ldots, u_q)$, and natural numbers $T_{\max}$, ℓ. *Find:* a collection $\mathcal{M} = \{n_1, \ldots n_m, \ldots\}$ of indices of the sequence Y; a collection $\mathcal{P} = \{p^{(1)}, \ldots, p^{(m)}, \ldots\}$ of natural numbers; a collection $\mathcal{J} = \{J^{(1)}, \ldots, J^{(m)}, \ldots\}$ of contraction mappings, where $J^{(m)} : \{1, \ldots, p^{(m)}\} \longrightarrow \{1, \ldots, q\}$; and the length M of these collections; which minimize the objective function

$$F(\mathcal{M}, \mathcal{P}, \mathcal{J}) = \sum_{m=1}^{M} \sum_{i=1}^{p^{(m)}} \{u^2_{J^{(m)}(i)} - 2y_{n_m+i-1} u_{J^{(m)}(i)}\},$$

under the constraints

$$q \leqslant p^{(m)} \leqslant \ell \leqslant T_{\max} \leqslant N, \;\; m = 1, \ldots, M,$$

$$p^{(m-1)} \leqslant n_m - n_{m-1} \leqslant T_{\max}, \;\; m = 2, \ldots, M,$$

$$p^{(M)} \leqslant N - n_M + 1,$$

on the elements of the collections $\mathcal{M}$ and $\mathcal{P}$, and under the constraints

$$J^{(m)}(1) = 1, \quad J^{(m)}(p^{(m)}) = q,$$
$$0 \leqslant J^{(m)}(i) - J^{(m)}(i-1) \leqslant 1, \quad i = 2, \ldots, p^{(m)},$$
$$m = 1, \ldots, M,$$

on the contraction mappings.

The second variant of the general problem induces **Problem 2**, in which the length $M \leqslant \lfloor \frac{N}{q} \rfloor$ is defined at the input, in contrast to Problem 1 [2].

The main mathematical result of this work is the following theorem.

Theorem 1. There exist algorithms that find exact solutions to Problems 1 and 2 in time $\mathcal{O}(T_{\max}^3 N)$ and $\mathcal{O}(T_{\max}^3 MN)$, respectively.

We prove the theorem constructively. Namely, we explicitly derive two dynamic programming schemes and prove that they find exact solutions for the problems in the specified polynomial time. Herewith, if the value of $T_{\max}$ is bounded by a constant (or fixed), then the running times of the algorithms are $\mathcal{O}(N)$ and $\mathcal{O}(MN)$.

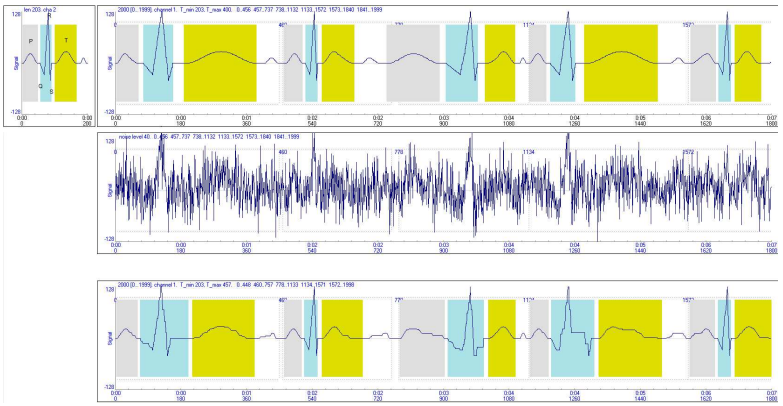


Fig. 1. An example of processing an ECG-type pulse train

Fig. 1 and 2 depict examples of noise-robust processing of modeled signals. In the top row of each figure, the input sequence U (pulse) and a modeled sequence, which is inaccessible to the algorithm, are depicted. The modeled sequence corresponds to the train of fluctuating pulses. In the middle row of each figure, the observable input sequence Y is shown. In the bottom rows of the figures, the recovered signal computed by the algorithm is shown.

Our obtained results are as follows. Firstly, we prove the polynomial-time solvability of two new discrete optimization problems, and secondly, we present a novel approach to processing of biomedical pulse signals fluctuating in time.

The study was supported by the Russian Foundation for Basic Research, projects 19-07-00397 and 19-01-00308, by the Russian Academy of Science (the Program of basic research), project 0314-2019-0015, and by the Russian Ministry of Science and Education under the 5-100 Excellence Programme.

[1] *Kel'manov A., Mikhailova L., Ruzankin P., Khamidullin S.* A minimization problem for sum of weighted covolutions' differences // Comp. Math. and Math. Phys. 2019 (accepted).

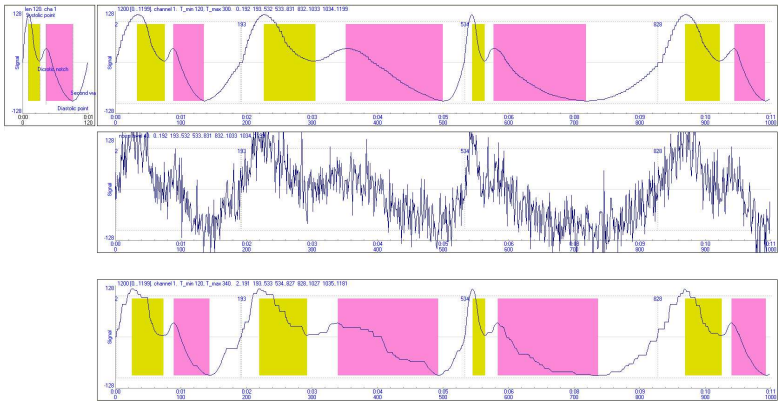


Fig. 2. An example of processing a PPG-type pulse train

[2] *Kel'manov A., Mikhailova L., Ruzankin P., Khamidullin S.* A minimization problem for sum of weighted covolutions' differences, the case of a given numbers of elements in the sum // Siberian J. Nun. Math. 2019 (accepted).

Применение нейронной сети Mask RCNN в задачах анализа пространственно- временных характеристик сердечного ритма модельного тест-объекта Daphnia magna

Заалишвили Никита Юрьевич[1]⋆ nikzasel@gmail.com
Каленков Георгий Сергеевич[2] kalenkov@mail.ru
Сарапульцева Елена Игоревна[3] helen-bio@yandex.ru
[1]Москва, Московский Политех
[2]Москва, ИДГ(РАН)
[3]Москва, ИАТЭ НИЯУ МИФИ

Daphnia является полупрозрачным низшим ракообразным, широко используемым в качестве ключевой модели в области экотоксикологии, где анализируется воздействие токсинов на человека и экосистемы. Daphnia как модельный тест-организм используется также для тестирования новых лекарственных средств в области фармакологии, поскольку структура генома Daphnia определена и имеет высокое сходство с геномом человека. Исследование частоты сердечных сокращений Daphnia является важным показателем жизнеобеспечения организма. Помимо повышения точности измерения, актуальность разработки методов интерферометрии живых систем, например, в радиологии, может быть обусловлена необходимостью регистрации ряда важных пространственно-временных характеристик сердца. К ним можно отнести форму, объем, амплитуду колебаний, аритмичность сокращения клапанов межкамеральных остий. Одна из проблем автоматизации измерений состоит в поиске и сегментации области изображения, содержащего сердечную мышцу животного [1]. Мы показали, что эту задачу можно решить, используя нейронную сеть Mask RCNN, которая является надстройкой Faster RCNN. Данные представляют собой набор из 40 изображений рачков, полученных под микроскопом, размером 2560x2048 пикселей, 8 бит на пиксель. Эксперимент проведён, с использованием перекрёстной валидации на трёх свёртках. Тренировочные данные были аугментированы в процессе обучения сети. Использовали следующие аугментации: вращение изображения, гауссовский шум, размытие по Гауссу, изменение контрастности.

Рис. 1. Изображение и объект распознавания. а)Целевое изображение, б)Исходный сигнал в)Увеличенный фрагмент изображения, содержащий предмет обучения – сердце дафнии (выделено желтым).

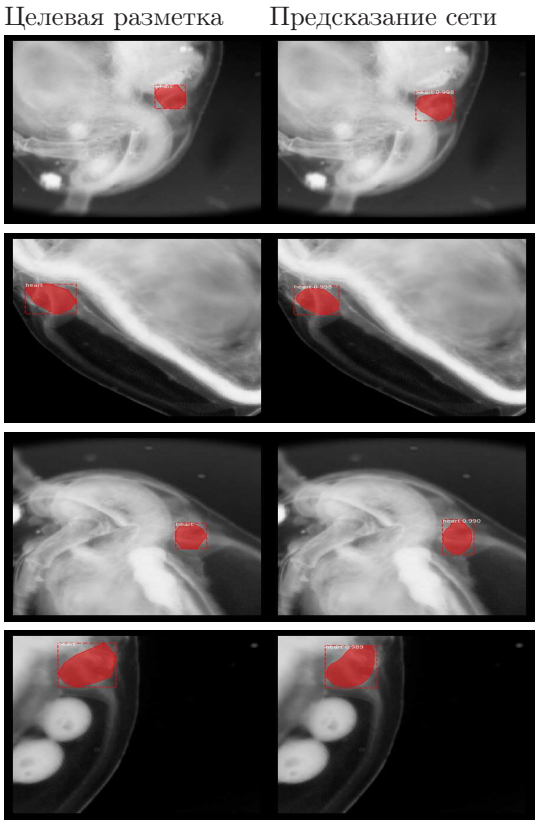


Рис. 2. Целевая разметка и предсказание сети.

Несмотря на то, что разметка в ручном режиме часто неточно определяла нужную область на изображении, сеть показала высокую обобщающую способность и корректно выделила область изображения, содержащего сердечную мышцу. В этой связи, для оценки достоверности результатов попиксельные метрики не использовали, параметров точности (Precision) и отзывчивости (Recall). Успешным предсказание считалось при значении метрики IoU (intersection over union) большим 0.5. Результаты представлены на рис. 2. Используя вышеуказанные условия, мы получили высокую производительность сети в контексте выбранных метрик. Работа выполнена при поддержке РФФИ № 18-07-01403.

[1] *Sarapultseva E. I. et. al.* "Image correlation and low-coherence interferometry applied to Daphnia magna heartbeat counting". // 4th International Conference on Radioecology & Environmental Radioactivity, P8-16, 2017.

Application of the RCNN neural network mask for spatio-temporal characteristics analysis of the test model-object Daphnia magna heart beat counting

Nikita Zaalishvili[1]⋆ nikzasel@gmail.com
Grigoriy Kalenkov[2] kalenkov@mail.ru
Elena Sarapultseva[3] helen-bio@yandex.ru
[1]Moscow, Moscow Polytechnic University
[2]Moscow, Institute of Geosphere Dynamics
[3]Moscow, National Research Nuclear University MEPhI

Daphnia is a translucent inferior crustacean widely used as a key model in ecotoxicology, which analyzes the effects of toxins on humans and ecosystems. Daphnia as a model test organism It is also used to test new drugs in the field of pharmacology, since the structure of the Daphnia genome is defined and has a high similarity to the human genome. Daphnia Heart Rate Study is an important indicator of the body's life support. In addition to improving measurement accuracy, the relevance of developing methods for the interferometry of living systems, for example, in radiology, may be due to the need to register a number of important spatio-temporal characteristics of the heart. These include the shape, volume, amplitude of oscillations, arrhythmic contraction of the valves of the intercameral awns. One of the problems of measurement automation is the search and segmentation of the image area containing the animal's heart muscle [1]. We showed that this problem can be solved using the neural network Mask RCNN, which is a superstructure of Faster RCNN. The data is a set of 40 images of crustaceans obtained under a microscope, size 2560x2048 pixels, 8 bits per pixel. The experiment was conducted using cross-validation in three packages. Training data was augmented during training network. The following augmentations were used: image rotation, Gaussian noise, Gaussian blur, change in contrast.

Fig. 1. The image and object recognition. a) Target image, b) Original signal, c) An enlarged fragment of the image containing the subject of study - the heart of daphnia (highlighted in yellow).

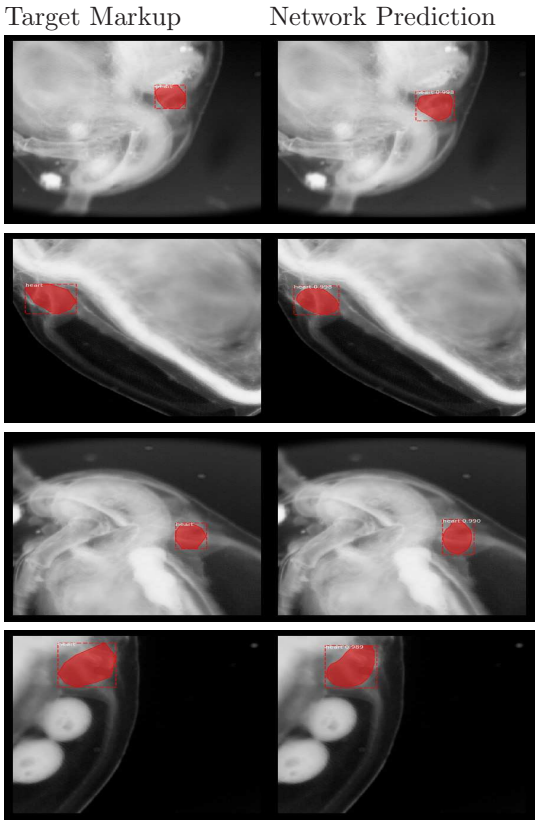


Fig. 2. Target Markup and network Prediction

Despite the fact that manual marking often inaccurately determined the desired area in the image, the network showed high generalizing ability and correctly selected the area of the image containing the heart muscle. In this regard, to evaluate the reliability of the results, pixel-by-pixel metrics were not used, parameters of accuracy (Precision) and responsiveness (Recall). The prediction was considered successful if the IoU (intersection over union) metric was greater than 0.5. The results are presented in Fig. 2. Using the above conditions, we have obtained high network performance in context of selected metrics. This research is funded by RFBR, grant 18-07-01403.

[1] *Sarapultseva E. I. et. al.* "Image correlation and low-coherence interferometry applied to Daphnia magna heartbeat counting". // 4th International Conference on Radioecology & Environmental Radioactivity, P8-16, 2017.

Метод сравнения бинарных растровых изображений, содержащих дыры, с учетом информации об осях симметрии

Федотова Софья Антоновна* fedotova.sonya@gmail.com
Середин Олег Сергеевич oseredin@yandex.ru
Кушнир Олеся Александровна kushnir-olesya@rambler.ru
Тула, Тульский государственный университет

В работе [1] рассматривается проблема сравнения бинарных растровых изображений. Для достижения инвариантности сдвигу, повороту и масштабированию предлагается метод сравнения двух изображений с использованием информации об осях симметрии фигур. Ось симметрии фигуры предлагается искать одним из ранее разработанных методов: основанном на скелетном представлении фигуры [2], уточнения скелетной оси или полным перебором [3]. В качестве меры сходства используется подобие Жаккара.

Важной особенностью предлагаемого подхода является то, что при сравнении форм учитывается не только внешний контур фигуры, но и внутренние контуры, то есть дыры в фигуре, при их наличии. Фигуры с дырами, как круглой формы, так и произвольной, встречаются во многих приложениях компьютерного зрения. Примером таких фигур могут послужить круг и кольцо, чьи внешние контуры будут одинаковыми, однако внутри эти объекты сильно различаются. При классификации таких объектов, если метод основывается только на анализе внешнего контура, кольцо будет ошибочно классифицировано как круг.

В случае если дыра является аномалией на изображении (шумы оцифровки, производственные дефекты на сравниваемых изделиях и т.п.), качество наложения и распознавания может ухудшиться в зависимости от размера, формы самой дыры. Однако, если объекты изначально должны иметь внутренние контуры, то данный подход позволит повысить качество распознавания для объектов такого класса и отделить их от объектов с похожим внешним контуром, но отличающейся внутренней структурой. Как альтернативу можно рассмотреть вариант, когда ось симметрии определяется непосредственно по изображению с дырами без предварительной заливки. Видимо, такой подход может быть состоятелен в случаях, когда выполняется сравнение с точным эталонным изображением, например, анализируется силуэт детали на конвейере с целью ее отбраковки.

Алгоритмы экспериментально исследованы на базах бинарных растровых изображений «Бабочки» и FLAVIA. Для определения качества работы алгоритмов на морфологически сложных объектах, содержащих внутренние контуры, на бинарных изображениях бабочек и листьев искусственно случайным образом создавались дыры. Предложенный подход позволяет классифицировать такие объекты с незначительной потерей качества.

Работа выполнена при поддержке РФФИ, гранты № 18-07-00942, № 18-07-01087.

[1] *Федотова С. А., Кушнир О. А., Середин О. С.* Сравнение бинарных изображений на основе меры Жаккара с использованием информации о симметрии // Статья находится на рецензировании на конференции VISAPP, 2020.

[2] *Kushnir O., Seredin O., Fedotova S., Karkishchenko, A.* Reflection symmetry of shapes based on skeleton primitive chains // International Conference on Analysis of Images, Social Networks and Texts, Cham: Springer, 2016. — P. 293–304.

[3] *Kushnir O., Seredin O., Fedotova S.* Algorithms for Adjustment of Symmetry Axis Found for 2d Shapes by the Skeleton Comparison Method // International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2019.

Comparison of binary images containing holes considering information about the axes of symmetry

Fedotova Sofia⋆ fedotova.sonya@gmail.com
Seredin Oleg oseredin@yandex.ru
Kushnir Olesia kushnir-olesya@rambler.ru
Tula, Tula State University

In this work [1] the problem of comparison of binary images is considered. We proposed a method of comparing two images using information about the axes of symmetry of the figures to achieve invariance to translation, rotation and scaling. The symmetry axis of the figure is found by one of the previously developed methods: based on the pair-wise comparison of sub-sequences of skeleton primitives [2], adjustment of the skeleton axis or pair-wise exhaustive search [3]. The Jaccard index is used as a measure of similarity.

An important feature of the proposed approach is that the comparison of images consider not only the external contour of the figure, but also the internal contours, that is, holes in the figure, if they exist. Shapes with holes, both circular and arbitrary, are found in many computer vision applications. An example of such shapes can serve as a circle and a ring, whose outer contours are the same, but inside these objects are very different. When classifying such objects, the ring will be falsely classified as a circle, if the method is based only on the analysis of the outer contour.

If the hole is an anomaly in the image (digitization noise, manufacturing defects on the compared workpieces, etc.), the quality of overlay and recognition may deteriorate depending on the size and shape of the hole. However, if the objects must initially have internal contours, this approach will improve the quality of recognition for objects of this class and separate them from objects with a similar external contour, but different internal structure. As an alternative, the axis of symmetry can be found on the image with holes without pre-filling. Apparently, this approach can be made in cases where a comparison is made with an accurate reference image, for example, the silhouette of a part on a conveyor is analyzed for its rejection.

The algorithms were experimentally studied on the bases of binary bitmaps "Butterflies" and FLAVIA. To determine the quality of the algorithms on morphologically complex objects containing internal contours, holes were artificially randomly created on binary images of butterflies and leaves. The proposed approach allows to classify such objects with a slight loss of quality.

This research is funded by RFBR, grants 18-07-00942, 18-07-01087.

[1] *Fedotova S., Kushnir O., Seredin O.* Comparison of binary images based on the Jaccard measure considering information about symmetry // The article is under review at the VISAPP conference, 2020.

[2] *Kushnir O., Seredin O., Fedotova S., Karkishchenko, A.* Reflection symmetry of shapes based on skeleton primitive chains // International Conference on Analysis of Images, Social Networks and Texts, Cham: Springer, 2016. — P. 293–304.

[3] *Kushnir O., Seredin O., Fedotova S.* Algorithms for Adjustment of Symmetry Axis Found for 2d Shapes by the Skeleton Comparison Method // International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2019.

Анализ и поиск видеоизображений по опорным кадрам с использованием гранично-скелетной модели формы

Аминова Ксения Владимировна[1] kz@pisem.net
Рейер Иван Александрович[1]⋆ reyer@forecsys.ru
[1]Москва, ФИЦ ИУ РАН

В докладе рассматривается задача анализа и поиска видеоизображений в базах графических данных по визуальным характеристикам (выделение различных структурных элементов в видеопоследовательности, индексирование и сравнение). При этом предполагается использовать вспомогательную информацию, уже полученную на этапе кодирования последовательности кадров видеоизображения специальными методами сжатия. Примерами такой информации являются коэффициенты дискретного косинусного преобразования опорных кадров, по которым можно быстро восстановить «грубую» копию изображения, и «векторы движения» — оценки предполагаемых движений элементов в соседних кадрах.

Ранее авторами было предложено непрерывное представление дискретного цветного изображения в виде разбиения плоскости непересекающимися многоугольными фигурами, соответствующими фрагментам растра с однородными цветовыми или текстурными характеристиками. Для построения множества многоугольников используется модификация алгоритма аппроксимации бинарного изображения разделяющими многоугольниками минимального периметра. Сегментация изображения основана на процедурах выделения суперпикселов — небольших однородных областей. Соответствующие суперпикселам многоугольники объединяются в относительно небольшое количество крупных фигур с использованием алгоритма иерархической кластеризации цветов суперпикселов и анализа количества элементов в выделенных кластерах и их соседства в гранично-скелетной модели.

Помимо границ многоугольных фигур модель сегментированного изображения также включает скелетную часть — «размеченные» скелеты многоугольников, описывающие изменение скелетного представления, и оценки значимости выпуклых особенностей границы, соответствующих вершинам многоугольников [1]. Оценки значимости вычисляются на основе анализа параметрического семейства гранично-скелетных моделей формы, порожденного многоугольной фигурой.

Полученные модели изображений сравниваются по форме и цвету многоугольных фигур. Для сравнения формы фигур в моделях можно использовать параметрический дескриптор, описывающий характер изменения числа существенных особенностей контура при росте величины точности аппроксимации; а также функцию медиальной ширины фигуры. Для сокращения количества сравнений при вычислении сходства изображений применяется аппарат M-деревьев.

В докладе предлагаются процедуры построения гранично-скелетных моделей для приближенно восстановленных опорных кадров. Рассматриваются стандарты сжатия видеоинформации, использующие разбиение кадра на пиксельные макроблоки и кодирование опорных кадров с помощью дискретного косинусного преобразования и его модификаций (MPEG-2, H.264, VP8 и т.п.). В случае независимого кодирования макроблоков опорного кадра для построения модели используется изображение с восстановленными усредненными значениями цвета макроблоков (DC-picture). Если же при кодировании применялись методы «пространственного предсказания» (spatial prediction), то «грубая» копия кадра строится на основе приближенно восстановленной разности между макроблоком и его предсказанным значением.

Работа поддержана грантом РФФИ № 17-07-01432.

[1] *Reyer I., Aminova K.* Parametric Shape Descriptor based on a Scalable Boundary-Skeleton Model // Communications in Computer and Information Science, 2019. (В печати).

Video analysis and retrieval by intra-frames with use of boundary-skeleton shape model

Ksenia Aminova[1] kz@pisem.net
Ivan Reyer[1]⋆ reyer@forecsys.ru
Moscow, FRC CSC RAS

An approach to content-based video retrieval and analysis is considered. A continuous model of a segmented raster frame consisting of a set of nonoverlapping polygonal figures is constructed. Each polygon from the set approximates a segmented raster region within the image, with polygons of two neighbor regions having common fragments of boundary. To obtain the set of polygons a modified algorithm for approximation of a binary image with polygons of minimal perimeter is used. The segmentation is based on superpixel extraction procedures. The polygons corresponding to superpixels are grouped into a relatively small number of large figures using an algorithm of hierarchical color clustering and analyzing the number and neighborhood of elements in the obtained clusters.

The model also includes marked skeletons of polygons describing changes of skeletal representation and significance estimations for boundary convexities corresponding to polygon vertices [1]. The estimations are calculated with use of a family of boundary-skeleton shape models generated by a polygonal figure.

Obtained image models are compared by shape and color of polygons. To estimate the shape similarity, integral morphological features are compared. To reduce the number of comparisons at the calculation of image similarity, M-trees are used.

The presented approach is applied to retrieval and analysis of video sequences in compressed domain. DCT-based video coding standards with and without intra-frame spatial prediction are considered. Procedures for constructing boundary-skeleton models from partially reconstructed intra-frames are suggested.

This research is funded by RFBR, grant 17-07-01432.

[1] *Reyer I., Aminova K.* Parametric Shape Descriptor based on a Scalable Boundary-Skeleton Model // Communications in Computer and Information Science, 2019. (in press).

Метод распознавания осевой симметрии объектов на цифровых изображениях

Местецкий Леонид Моисеевич[1]⋆ mestlm@mail.ru
Журавская Александра Валерьевна[1] a.v.zhuravskaya@gmail.com
[1]Москва, МГУ, ВМК

Симметричность объектов является важным классификационным признаком при распознавании объектов на цифровых изображениях. В частности, симметрию необходимо определять в задачах дешифрирования аэрокосмических снимков, при анализе изображений биологических объектов. При этом оценка симметричности объектов на цифровых изображениях часто оказывается затруднительной из-за шумов, окклюзий, а также ввиду низкого разрешения или слишком мелких размеров изображений. Вследствие этих причин на растровом изображении не существует идеально симметричных объектов. Поэтому необходимо приспособить геометрические критерии идеальной зеркальной симметрии для оценки степени симметричности реальных дискретных объектов на цифровых изображениях. При этом требования, предъявляемые к такому критерию, включают низкую алгоритмическую сложность вычисления, поскольку задача должна решаться в реальном времени систем компьютерного зрения. В данной работе предлагается количественная мера асимметричности произвольных односвязных объектов, представленных в формате растровых бинарных изображений. Мера асимметричности объекта определяется через дискретное преобразование Фурье последовательности точек контура дискретной границы объекта. На основе этого показателя строится критерий для определения симметричных объектов, выполняется ранжирование наблюдаемых объектов по степени асимметричности. Для вычисления предложенной меры асимметричности объекта разработан эффективный алгоритм. Алгоритм позволяет вычислить меру асимметричности и определить наиболее правдоподобно ось симметрии на основе решения задачи минимизации меры асимметричности объекта. Алгоритм имеет квадратичную сложность по количеству точек в контуре границы объекта. Практическая оценка работоспособности и эффективности алгоритма получена на примере классификации объектов при дешифрировании аэрокосмических снимков. Работа поддержана грантом РФФИ № 17-01-00917.

[1] *Mestetskiy L. Zhuravskaya A.* Method for assessing the symmetry of objects on digital binary images based on fourier descriptor // Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci. — XLII-2/W12. — 2019. — P. 143–148.

Method for recognition of axial symmetry of objects in digital images

Leonid Mestetskiy[1]⋆ mestlm@mail.ru
Aleksandra Zhuravskaya[1] a.v.zhuravskaya@gmail.com
[1]Moscow, MSU

The symmetry of objects is an important classification feature when recognizing objects in digital images. In particular, symmetry must be determined in the problems of decoding aerospace images, as well as in the analysis of images of biological objects. Moreover, assessing the symmetry of objects in digital images is often difficult due to noise, occlusions, and also due to low resolution or too small image sizes. For these reasons, there are no perfectly symmetrical objects on the bitmap. Therefore, it is necessary to adapt the geometric criteria of ideal mirror symmetry to assess the degree of symmetry of real discrete objects in digital images. Moreover, the requirements for such a criterion include low algorithmic computational complexity, since the problem must be solved in real time by computer vision systems. In this paper, we propose a quantitative measure of the asymmetry of arbitrary simply connected objects represented in the format of binary raster images. The measure of the asymmetry of an object is determined through the discrete Fourier transform of a sequence of points in the contour of the discrete boundary of the object. Based on this indicator, a criterion is constructed to determine symmetric objects, and the observed objects are ranked according to the degree of asymmetry. An effective algorithm has been developed to calculate the proposed measure of the asymmetry of the object. The algorithm allows us to calculate the measure of asymmetry and determine the most likely axis of symmetry based on the solution of the problem of minimizing the measure of asymmetry of an object. The algorithm has quadratic complexity in the number of points in the contour of the boundary of the object. A practical assessment of the efficiency and effectiveness of the algorithm is obtained by the example of the classification of objects during decoding of aerospace images.
This work was supported by a grant RFBR No 17-01-00917.

[1] *Mestetskiy L. Zhuravskaya A.* Method for assessing the symmetry of objects on digital binary images based on fourier descriptor // Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci. — XLII-2/W12. — 2019. — P. 143–148.

Метод графемного описания и распознавания букв на основе медиального представления

Местецкий Леонид Моисеевич[1]⋆ mestlm@mail.ru
Липкина Анна Львовна[1] lipkina96@mail.ru
[1]Москва, МГУ, ВМК

Понятие графемы является фундаментальным в письменности и в чтении. Графема представляет собой наиболее общую схему символа алфавита, и нарисовать её сможет любой грамотный человек, даже ребёнок. Школа учит читать и писать на основе графем. Однако компьютерные программы распознавания текста не используют это понятие в явном виде. Графемы используют филологи в своих теоретических построениях, а также дизайнеры при создании компьютерных шрифтов. И те и другие обходятся без строгого определения понятия графемы. Если попытаться создать алгоритмы распознавания символов алфавита на основе графем, то необходимо более строго определить это понятие и способы его описания и построения.

В этой статье мы делаем попытку определить схематические описания символов алфавита таким образом, чтобы их можно было получить из любого шрифта, и чтобы после этого можно было распознавать буквы во всех остальных шрифтах. Для решения этой задачи мы предлагаем метод получения графем в виде графов из цифровых изображений букв какого-либо шрифта и метод распознавания символов других шрифтов на основе сравнения с графемами. Основная гипотеза состоит в том, что для построения универсального набора графем достаточно одного типового шрифта, а остальные шрифты можно будет распознавать по этому набору. Таким образом, целью исследования является реализация и проверка графемного подхода в распознавании букв.

В статье предлагается концепция математической модели графемы символов и метод построения графем, основанный на непрерывном медиальном представлении букв в цифровых изображениях. Предлагается также метод распознавания изображения печатного текста на основе математической модели графемы, используемой при генерации признаков и для построения классификатора. Представлены результаты экспериментов, подтверждающих эффективность графемного подхода, высокое качество распознавания текста в разных вариантах шрифта и в разных качествах текстового изображения.

Работа поддержана грантом РФФИ № 17-01-00917.

[1] *Lipkina A. Mestetskiy L.* Grapheme Approach to Recognizing Letters based on Medial Representation // VISIGRAPP (4: VISAPP) 2019 — P. 351-358.

The method of grapheme description and recognition of letters based on the medial representation

Leonid Mestetskiy[1]⋆ mestlm@mail.ru
Anna Lipkina[1] lipkina96@mail.ru
[1]Moscow, MSU

The concept of grapheme is fundamental in writing and in reading. A grapheme is the most general outline of an alphabet symbol, and any person, even a child, can draw it. The school teaches reading and writing based on graphemes. However, computer text recognition programs do not use this concept explicitly. Graphemes are used by philologists in their theoretical constructions, as well as designers when creating computer fonts. Both of them do without a strict definition of the concept of grapheme. If you try to create algorithms for recognizing alphabet characters based on graphemes, you need to more strictly define this concept and how to describe and construct it.

In this article, we attempt to define schematic descriptions of alphabet characters in such a way that they can be obtained from any font, and so that after that letters can be recognized in all other fonts. To solve this problem, we propose a method for obtaining graphemes in the form of graphs from digital images of letters of a font and a method for recognizing characters of other fonts based on comparisons with graphemes. The main hypothesis is that to build a universal set of graphemes, one standard font is enough, and the rest of the fonts can be recognized by this set. Thus, the aim of the research is the implementation and verification of the grapheme approach in recognizing letters.

The article proposes the concept of a mathematical model of the grapheme of symbols and the method of constructing graphemes based on the continuous medial representation of letters in digital images. A method for recognizing images of printed text based on the mathematical model of the grapheme used to generate features and to construct a classifier is also proposed. The results of experiments confirming the effectiveness of the grapheme approach, the high quality of text recognition in different font types and in different qualities of the text image are presented.

This work was supported by a grant RFBR No 17-01-00917.

[1] *Lipkina A. Mestetskiy L.* Grapheme Approach to Recognizing Letters based on Medial Representation // VISIGRAPP (4: VISAPP) 2019 — P. 351-358.

Нейросетевые детекторы в задаче анализа предпочтений пользователя по фотографиям

Гречихин Иван Сергеевич[1]⋆ igrechikhin@hse.ru
Савченко Андрей Владимирович[1] avsavchenko@hse.ru

[1]Нижний Новгород, Национальный исследовательский университет Высшая школа экономики

В статье [1] описывается подход, в котором автоматически анализируется галерея фотографий для определения предпочтений пользователя. Предложен новый метод, в котором вначале для обнаружения объектов непосредственно на мобильном устройстве использованы вычислительно эффективные SSD-детекторы. Предполагается, что фото и видео пользователя с лицами его самого и его близких людей являются приватными. Для определения таких фото проводится детектирование лиц на фотографиях и видео; извлечённые с помощью предварительно обученной сверточной нейронной сети векторы признаков лиц кластеризуются для определения лиц пользователя и его близких родственников и друзей. Для определения кластеров лиц использовался метод иерархической кластеризации. Публичные фотографии, не содержащии их лиц, могут быть проанализированы на удалённом вычислительном сервере с помощью высокоточных детекторов на основе Faster R-CNN. Выполнено обучение нейросетевых детекторов на наборе изображений объектов, представляющих категории интеерсов из наборов данных MS COCO, ImageNet и OID v4. Лучшая модель имела среднюю полноту 0.75 для 79 категорий объектов при средней точности 0.663. Вычислительно эффективные SSD-модели оказались менее точными, по сравнению с серверными Faster R-CNN моделями, но намного более быстрыми.

Статья подготовлена в результате проведения исследования (№ 19-04-004) в рамках Программы «Научный фонд Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ)» в 2019 г. и в рамках государственной поддержки ведущих университетов Российской Федерации “5–100”.

[1] *Гречихин И.С., Савченко А.В.* Метод анализа предпочтений пользователя по фото- и видеоизображениям на мобильном устройстве на основе нейросетевых детекторов объектов на изображениях // Информационные технологии. 2019. Т. 25. No 9. С. 538-544

Neural-Network Object Detectors in Analysis of a Gallery of Photos for User Modeling

Ivan Grechikhin[1]⋆ igrechikhin@hse.ru
Andrey Savchenko[1] avsavchenko@hse.ru
[1]Nizhny Novgorod, National Research University Higher School of Economics

The article [1] describes an approach for extraction of user preferences based on the analysis of a gallery of photos. It is proposed to firstly use fast SSD-based methods in order to detect objects of interests in offline mode directly on mobile device. This is done to ensure the safety of personal user data. We assume that the photos or videos with faces of owner or close friends/relatives are private. Therefore, we perform facial analysis of all visual data: extract feature vectors from detected facial regions, cluster them and select public photos and videos which do not contain faces of friends and relatives. Hierarchical clustering is used to discover groups of faces that regularly appear in user's gallery. At the second stage, the public images without such faces are processed on the remote server using very accurate but rather slow object detectors based on Faster R-CNN.

Experimental study of several contemporary detectors is presented with the specially designed subset of MS COCO, ImageNet and Open Images datasets. Models for both online (on remote server) and offline use were trained. For the best online model, there are 79 categories of objects with average recall more than 0.75 and average precision 0.663. Quite obviously, light-weighted SSD-based models for offline object detection perform less acurrately, but work faster and take less memory.

This research was prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE University) in 2019 (grant No. 19-04-004) and within the framework of the Russian Academic Excellence Project "5–100".

[1] *Grechikhin I.S., Savchenko A.V.* Analysis of user preferences using photos and videos from mobile device based on object detection and neural networks, Informatsionnye tekhnologii 2019, vol. 25, no. 9, pp. 538–544 (in Russian)

Распознавание пола и возраста лица на видеоизображениях для мобильных платформ

Харчевникова Ангелина Сергеевна[1]⋆ angelina.kharchevnikova@gmail.com
Савченко Андрей Владимирович[1] avsavchenko@hse.ru
[1]Нижний Новгород, Национальный исследовательский университет Высшая школа экономики

В целях повышения точности классификации изображений в настоящее время во многих задачах в области компьютерного зрения применяют сверточные нейронные сети (СНС). Несмотря на широкое распространение программных продуктов по распознаванию пола и возраста, их внедрение все еще не становится повсеместным, так как эффективность предложенных решений не всегда приемлема для практического применения. Действительно, реализация глубоких нейросетевых архитектур является вычислительно дорогостоящей операцией, что ограничивает эксплуатацию системы на мобильных устройствах. В работе [1] исследуются способы преодоления проблемы высокой вычислительной сложности без потерь в точности принимаемых решений на основе применения специально созданных эффективных нейросетевых моделей.

Задача распознавания видеоизображения состоит в том, чтобы отнести вновь поступающую на вход последовательность $\{X(t)\}, t = \overline{1,T}$ из T кадров с изображением лица к одному из L классов из базы эталонов. При этом предполагается, что классы заданы с помощью множества прецедентов (эталонных изображений) $\{X_i\}, i = \{1, ..., M\}$, метка класса которых известна. Так, задача распознавания пола является примером бинарной классификации (число классов $L = 2$). Распознавание возраста обычно рассматривается как задача регрессии, однако, на практике наибольшая точность принимаемых решений достигается при ее сведении к задаче классификации с определением нескольких возрастных интервалов (типичны значения $L = 8$ или $L = 100$).

При обработке видеоизображений для каждого поступающего кадра $X(t)$ сначала может решаться обычная задача автоматического распознавания изображений с помощью СНС, после все индивидуальные решения комбинируются в одно общее для конкретной видеозаписи. В процессе распознавания характеристик пола и возраста на видео на вход СНС подается матрица пикселей изображения лица на каждом последовательном кадре $X(t)$. Выход нейросетевой модели обычно получается в слое Softmax, который выдает оценки апостериорных вероятностей принадлежности t-го кадра к каждому l-му классу:

$$P(l \mid X(t)) = \mathrm{softmax} z_l(t) = \frac{\exp z_l(t)}{\sum\limits_{j=1}^{L} \exp(z_j(t))}, l = 1, 2, ..., L, \quad (1)$$

где $z_l(t)$ - выходы l-го нейрона на предпоследнем слоей сети (logits). Решение для каждого кадра принимается в пользу класса с максимальной апостериорной вероятностью:

$$l^*(t) = \underset{l=1,2,\ldots,L}{\arg\max} P(l \mid X(t)) \tag{2}$$

Вследствие влияния различных внешних факторов, как слабое разрешение видеокамеры, недостаток освещения, быстрая смена ракурса и др., принятие решения в пользу класса с максимальной апостериорной вероятностью для каждого кадра обычно отличается низкой точностью. В связи с этим в данной работе распознавание пола и возраста на видео рассматривается как задача выбора наиболее надежного решения из нескольких на основе коллективов решающих правил (КРП, комитетов классификаторов). В частности, исследуется применение теории Демпстера-Шафера, обычно использующейся для построения ансамблей классификаторов. Также рассматриваются такие методы КРП, как простое голосование, вычисление среднего арифметического, среднего геометрического и математического ожидания для задачи определения возраста.

Для проведения экспериментальных исследований было разработано специальное мобильное приложение для платформы Android. Для реализации СНС использовалась библиотека TensorFlow. В экспериментальных исследованиях для обработки каждого кадра использовались традиционные СНС Gender_net и Age_net, VGG-16, а также обученная нами MobileNet с двумя выходами, соответствующими предсказанию пола и возраста. Базовая часть этой сети обучалась для задачи идентификации лиц с использованием набора данных VGGFace2. Далее добавлялись новые слои для классификации пола и возраста, которые оюучались с использованием наборов данных IMDB-Wiki and Adience. Кроме того, исследовалась дообученная (fine-tuned) версия последней модели (MobileNet_ft), в которой на втором этапе обучения предсказанию пола и возраста выполнялась подстройка *всех* весов СНС (включая веса базовой модели, первоначально обученной для идентификации лиц). В результате удалось на 1% и 2% повысить точность по сравнению с исходной моделью для распознавания пола и возраста, соответственно. В экспериментах использовались следующие наборы видеоданных: Eurecom Kinect, Indian Movie (IMFDB), IARPA Janus Benchmark A (IJB-A), EmotiW 2018.

Проведенное экспериментальное исследование продемонстрировало повышение точности распознавания при реализации КРП в сравнении с традиционным подходом принятия решения для единичного кадра. Представлен сравнительный результат реализации нескольких СНС архитектур: моделей Age_net и Gender_net, VGG-16 и специально обученных для одновременного распознавания пола и возраста на мобильном устройстве моделей MobileNet/MobileNet_ft, а также сжатой MobileNet. Метод агрегации Демпстера-Шафера и правило произведения оказались более точными в задаче классификации по полу. В то же

время наиболее надежный результат распознавания возраста достигается с применением алгоритма оценки математического ожидания с учетом количества классов возраста, включенных в результирующую формулу.

Статья подготовлена в результате проведения исследования (№ 19-04-004) в рамках Программы «Научный фонд Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ)» в 2019 г. и в рамках государственной поддержки ведущих университетов Российской Федерации "5-100".

[1] *Kharchevnikova A. S., Savchenko A. V.* Neural Networks in Video-Based Age and Gender Recognition on Mobile Platforms // Optical Memory and Neural Networks (Information Optics), Springer, 2018, vol. 27, no. 4, pp. 246–259.

Video-Based Age and Gender Recognition on Mobile Platforms

Angelina Kharchevnikova[1]⋆ angelina.kharchevnikova@gmail.com
Andrey Savchenko[1] avsavchenko@hse.ru
[1]Nizhny Novgorod, National Research University Higher School of Economics

The accuracy of traditional computer vision methods makes them impracticable. In particular, prediction of age and gender characteristics given a photo or video of a face sometimes cannot be applied in practice due to insufficient efficiency. In contrast to them convolutional neural networks (CNNs) specially trained for gender and age identification are now considered as a promising approach for effective decision making. Unfortunately, the CNN-based methods are characterized by considerable memory consumption and computational complexity. Reducing the amount of computations and memory consumption is still a challenge for developers of CNN-based software. Thus, the emphasis of the paper [1] is the improvement of computational efficiency of age and gender recognition without significant degradation of accuracy by using a specially trained modification of the MobileNet architecture with two outputs.

The goal of video recognition is to associate an input video frame sequence $\{X(t)\}, t = \overline{1,T}$ of T frames with one of L classes of reference images. The classes are defined by sets of examples (reference images) $\{X_i\}, i = \{1, ..., M\}$, which class labels are known. For example, gender recognition is an instance of binary classification (the number of classes $L = 2$). The age prediction is usually regarded as a regression problem; however, in practice the best results are achieved by reducing it to a classification task with a few predefined age intervals ($L = 8$ or $L = 100$ are typical).

In the video-based recognition system the processing of each input video frame $X(t)$ can be firstly done by feeding each video frame to a CNN. Then all CNN outputs are combined into a single solution for a particular video frame sequence. The images of the face from successive video frame $X(t)$ in the form of pixel arrays come to the CNN in the stage of gender and age identification. The output of the neural net is usually generated in the softmax layer, which estimates the posterior probabilities $P(l \mid X(t))$ of affiliation of the $t-th$ frame to the $l-th$ class:

$$P(l \mid X(t)) = softmax z_l(t) = \frac{\exp z_l(t)}{\sum\limits_{j=1}^{L} \exp(z_j(t))}, l = 1, 2, ..., L, \quad (1)$$

where $z_l(t)$ is the output of the l-th neuron at penultimate layer (logits). Each frame is inferred to belong to a class with the highest a-posteriori probability:

$$l^*(t) = \arg\max_{l=1,2,...,L} P(l \mid X(t)) \quad (2)$$

External factors such as insufficient camera resolution, poor lighting, quickly changing camera angles often causes the low accuracy of such approach. In the research we use an ensemble of decision rules is the most effective approach to increase stability and accuracy of classification. We investigate the popular fusion algorithms (classifier committees) where classifier predicts the probability of the input object belonging to a class. In particular, we investigate the theory of Demster-Shafer - a popular approach for generating classifier ensembles. Also, we consider the following fusion techniques: simple voting, arithmetic mean, geometric mean and mathematical expectation for the task of determining age.

A special mobile application for Android has been developed to carry out the experiments. The TensorFlow library was used to build a CNN. In the experiments such well-known CNNs as Gender_net, Age_net, VGG-16, and our MobileNet were used for the processing of each frame.Besides, a fine-tuned version of the latter CNN (MobileNet_ft) was investigated.

In this research we have investigated decision aggregating methods for gender and age video-frames identification. Several CNN models, including "lightweight" MobileNet models are considered. We developed a prototype gender and age recognition architecture for Android-based mobile systems, which uses the most reliable CNN and accurate decision aggregating method. The following video datasets were used in the experiments: Eurecom Kinect, Indian Movie (IMFDB), IARPA Janus Benchmark A (IJB-A), EmotiW 2018.

We carried out experiments to demonstrate that the use of sets of decision rules provides the increase in recognition reliability as compared with conventional decision-making algorithms designed for single frames. We compared different CNN architectures: Age_net, Gender_net, VGG-16 and mobile platforms-oriented model MobileNet specially trained for concurrent gender and age recognition. At first, the base part of the latter CNN is trained from scratch for face identification using VGGFace2 dataset. After that, we simultaneously train two heads (outputs) for age and gender classification using IMDB-Wiki and Adience datasets. Moreover, we fine-tuned this model (MobileNet_ft), in which the second stage included training of *all* weights of a CNN including weights of basic face identification model. We also tested the compressed version of MobileNet. The Dempster-Shafer aggregation method and product rule proved to be the most reliable solutions for the age classification problem. At the same time the most reliable results of age recognition are achieved with the aid of the expectation evaluating algorithm adjusted for the number of age classes included in the resulting formula.

This research was prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE University) in 2019 (grant No. 19-04-004) and within the framework of the Russian Academic Excellence Project "5-100".

[1] *Kharchevnikova A. S., Savchenko A. V.* Neural Networks in Video-Based Age and Gender Recognition on Mobile Platforms // Optical Memory and Neural Networks (Information Optics), Springer, 2018, vol. 27, no. 4, pp. 246–259.

Минимизация ошибки аппроксимации структурированного изображения кусочно–постоянными приближениями

Харинов Михаил Вячеславович khar@iias.spb.su

[1]Санкт–Петербург, Институт информатики и автоматизации Российской академии наук

В настоящее время возможности кластеризации пикселей изображения с реальной минимизацией *ошибки аппроксимации* E (суммарной квадратичной ошибки) еще далеко не исчерпаны. Для вычисления *оптимальных* приближений изображения с минимально возможной ошибкой E при данном числе кластеров пикселей, а также для формализации понятия объектов на изображении необходима точная постановка оптимизационной задачи и система методов минимизации E с учетом особенностей видеоданных. В докладе описываются методы минимизации ошибки E, которые выработаны путем получения и исследования последовательностей оптимальных приближений для конкретных изображений и опираются на понятие иерархически структурированного цветового изображения из N пикселей.

Изображение считается *структурированным*, если для него вычислена бинарная иерархия разбиений на кластеры пикселей и соответствующая последовательность кусочно–постоянных приближений, которые в зависимости от числа кластеров g описываются выпуклой последовательностью значений E. Бинарная иерархия кластеров пикселей допускается любой, что является главной особенностью разработки. Число кластеров $2N-1$ в бинарной иерархии почти вдвое больше, чем число пикселей в изображении. За счет операций с этими кластерами достигается реальная минимизация ошибки E и обеспечивается детектирование иерархии различаемых по цвету объектов из одного, нескольких или многих сегментов произвольной величины и формы.

Основным методом получения иерархически структурированного изображения является метод Уорда, который реализуется числом алгоритмов, сравнимым с N. Для определенности алгоритма, метод Уорда параметризуется *числом* g_0 *объектов* на изображении, совпадающим с числом кластеров, при котором значение E максимально приближается к оптимальному. Для снижения вычислительной сложности метод Уорда выполняется *по частям*, т. е. по g_0 кластерам некоторого начального разбиения изображения, которые обрабатываются как самостоятельные изображения. При рекурсивном ускорении метода Уорда по частям вычислительная сложность снижается с N^2 до $N^{\frac{4}{3}}$, $N^{\frac{16}{15}}$,$N^{\frac{256}{255}}$,...,N. Для корректного вычисления множества из $2N-1$ кластеров пикселей методом Уорда по частям, который завершается итеративным слиянием g_0 кластеров в один кластер оригинальным методом Уорда, достаточно, чтобы начальное разбиение изображения на g_0 кластеров нельзя было *улучшить*, снизив ошибку E посредством встречных операций разделения одного кластера надвое и слияния пары кластеров в один.

Указанное условие обеспечивается улучшением качества разбиения на g_0 кластеров CI(Clustering Improvement)–методом. CI–метод сводится к итеративному повторению встречных операций разделения/слияния кластеров, отвечающих максимальному снижениию E. CI–метод не изменяет разбиений, построенных методом Уорда, и особенно эффективен для улучшения грубых начальных приближений изображения.

CI–метод улучшения разбиения изображения при данном числе кластеров дополняется методом K–meanless (Двоенко С. Д., 2014). Метод K–meanless предназначен для снижения ошибки аппроксимации произвольного приближения изображения при неизменном числе структурированных кластеров пикселей. На выходе метода K–meanless получается приближение, которое нельзя улучшить по ошибке аппроксимации E за счет реклассификации той или иной предусмотренной части одного кластера пикселей в другой кластер. Метод K–meanless является модернизированной версией традиционных методов K–средних, которые при кластеризации пикселей, как правило, приводят к ложным минимумам E из-за выполнения операций с отдельными пикселями, огрубленного критерия их реклассификации и неблизкого к оптимальному начального приближения изображения. Метод K–meanless выполняет реклассификацию предусмотренных множеств пикселей с максимальным снижением E в стратегии от больших множеств к меньшим и обеспечивает эффективную *минимизацию* ошибки E для приближения изображения g_0 кластерами, которое настолько близко к оптимальному приближению, что не меняется при обработке CI–методом. При этом для приближений изображения g_0 кластерами из числа приближений, полученных методом Уорда, гарантируется, что ошибка E не превышает определенного порога.

В качестве *меры* H(Heterogeneity) неоднородности кластеров пикселей рассматривается абсолютная величина $\left|\frac{dE}{dg}\right|$ производной ошибки E по числу кластеров g, которая не убывает при укрупнении кластера.

Генерация и запоминание иерархии приближений изображения, а также скоростные операции с кластерами пикселей выполняются с помощью сети, «наброшенной» на пиксели изображения. Основная сеть кодирует иерархию разбиений изображения в паре массивов из N элементов. В одном массиве задается ациклический граф — дерево Слейтора-Тарьяна. Во втором массиве посредством циклического графа задается порядок установления дуг в дереве. По изображению и основной сети генерируется еще несколько десятков графов, систем указателей, массивов чисел и др. дополнительных компонентов из N элементов, которые позволяют работать с кластерами пикселей так же быстро, как с отдельными пикселями. При этом деревья Слейтора–Тарьяна, в отличие от традиционных деревьев(дендрограмм), с минимальными затратами памяти поддерживают произвольную бинарную иерархию кластеров пикселей.

В докладе вычислительная сеть для детектирования объектов сопоставляется с многослойной искусственной нейронной сетью (ИНС), что полезно для интерпретации зрительного восприятия.

Доклад актуален для совершенствования общеупотребительных методов кластерного анализа в инструментариях типа MatLab, а также создания программного обеспечения для детектирования объектов на изображении двухэтапной процедурой вычисления иерархии кластеров пикселей и ее преобразвания в приближение изображения с «объектами интереса» по пороговому значению параметра неоднородности H [1].

[1] *Харинов М. В.* Локализация объектов на цифровом изображении посредством кусочно–постоянных приближений // Известия ТулГУ. Технические науки, 2019. Вып. 6, Тула, 2019. — С. 160–169.

Minimization of the approximation error for describing of an image by piecewise constant approximations

Mikhail Kharinov[1]⋆ khar@iias.spb.su

[1]Saint–Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences

At present, the possibilities of clustering of image pixels with real minimization of the *approximation error* E (total squared error) are far from been exhausted. To calculate the *optimal* approximations of the image with the minimum possible error E for a given number of pixel clusters, as well as to formalize the concept of objects in the image, an accurate statement of the optimization problem and a system of minimizing methods E taking into account the features of the video data are necessary. The report describes the methods for minimizing of the error E, which are developed by obtaining and studying the sequences of optimal approximations for specific images basing on the concept of a hierarchically structured color image of N pixels.

An image is considered *structured* if it is defined a binary hierarchy of pixel clusters and the corresponding sequence of piecewise constant approximations, which are described by a convex sequence of E values depending on the number of clusters g. Any binary hierarchy of pixel clusters is allowed, which is the main feature of the development. The number of clusters $2N-1$ in the binary hierarchy is almost twice as large as the number of pixels in the image. Due to operations with these clusters, a real minimization of the error E is achieved and a hierarchy of color–distinguished objects from one, several, or many segments of arbitrary size and shape is detected.

The main method for obtaining a hierarchically structured image is Ward's clustering method, which is implemented by a comparable to N number of algorithms. For definiteness of the algorithm, Ward's method is parameterized by the *number of* g_0 *objects* in the image, which coincides with the number of clusters at which the value of E is as close as possible to the optimal one. To reduce the computational complexity, Ward's method is performed *in image parts*, i.e. within g_0 clusters of some initial image partition, which are processed as separate images. When recursively accelerating, the computational complexity of Ward's method in image parts decreases from N^2 to $N^{\frac{4}{3}}$, $N^{\frac{16}{15}}$, $N^{\frac{256}{255}}$,...,N. For the correct calculation of the set of $2N-1$ pixel clusters by Ward's method in image parts, which ends by iteratively merging of g_0 clusters into one cluster using the original Ward's method, it is enough that the initial partitioning of the image into g_0 clusters cannot be improved by reducing the error E by counter operations of dividing one cluster into two and merging of a pair of clusters into one.

This condition is provided by improving the quality of partitioning into g_0 clusters by CI(Clustering Improvement)–method. CI–method is reduced to iterative repetition of counter operations of splitting/merging of pixel clusters that correspond to the maximum reduction of E. CI–method does not change the partitions

constructed by Ward's method and is especially effective for improving of rough initial approximations of the image.

CI–method for improving of image partition into a given number of clusters is supplemented by K–meanless method (Dvoenko S. D., 2014). K–meanless method is designed to reduce the approximation error of arbitrary image approximation with fixed number of structured pixel clusters. The output of K–meanless method yields an approximation that cannot be improved by the approximation error E by means of reclassification of one or another part from preassigned number of parts of one pixel cluster into another pixel cluster. K–meanless method is an upgraded version of the conventional K–means methods, which when clustering pixels, as a rule, lead to false minimums E due to operations with only individual pixels, a coarsened criterion for their reclassification, and initial image approximation, which is oftenly far from the optimal. K–meanless method reclassifies the provided sets of pixels with the maximum reduction of E in the strategy from larger sets to smaller ones and supports the *minimizing* of the error E for image approximation with g_0 clusters, which is so close to the optimal approximation that stays invariant under the processing by CI–method. If the image approximation of g_0 clusters is selected from the number of approximations obtained by Ward's method, then due to the convexity property, it is guaranteed that the error E does not exceed a predetermined threshold.

As the measure of the heterogeneity for given pixel cluster it is considered the absolute value $\left|\frac{dE}{dg}\right|$ of the derivative of the error E with respect to the number of clusters g, which does not decrease upon the cluster enlargement.

Generation and storage of the hierarchy of image approximations, as well as high–speed operations with clusters of pixels are performed using the network, connecting the image pixels. The kernel network encodes a hierarchy of image partitions in a pair of arrays of N elements. In one array, an acyclic graph presenting Sleator–Tarjan dynamic tree is defined. In the second array the establishing order of arcs in the tree is specified by means of a cyclic graph. A dozens more graphs, systems of pointers, number arrays, and other components from N elements are generated for the image and kernel network, which allow working with pixel clusters as quickly as with individual pixels. This is achieved thanks to the Sleator–Tarjan trees, which, unlike the traditional trees(dendrograms), with minimal memory consumption support an arbitrary binary hierarchy of pixel clusters.

In the report, the computer network for object detecting is compared with a multilayer artificial neural network (ANN). It is useful for interpretation of visual perception.

The report is relevant for improving of commonly used cluster analysis methods, namely, in software tools like MatLab, as well as for creating of software for detection of objects in the image by means of a two–stage procedure for calculating of the hierarchy of pixel clusters and their transforming into an approximation of the im-

age with “objects of interest” according to the threshold value of the heterogeneity parameter H [1].

[1] *Kharinov M.* Localization of objects in a digital image by piecewise constant approximations // Izvestiya TulGU. Engineering, 2019. Vol. 6, Tula, 2019. — P. 160–169.

Алгоритмы подсчета нитей холстов картин по изображениям на основе максимизации взаимной информации

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

Работа посвящена решению задачи определения характеристик холстов картин, используемых для датировки, и является продолжением исследований, связанных с разработкой компьютерных методов анализа изображений для атрибуции произведений живописи. Одним из видов послойного исследования произведений живописи в атрибуции является определение производителя и датировка производства материала основы, в частности холста. Характеристики холстов зависят от уровня технологии производства. Поэтому специфика структуры ткани, использованной как основа для живописи, может служить источником информации о времени исполнения изучаемой картины. Для снижения трудоемкости и повышения точности анализа холстов необходимо разработать автоматизированные алгоритмы измерения параметров холстов по изображениям. В исследовании используются изображения, полученные фотосъемкой при направленном под острым углом по отношению к холсту освещении. Такой способ получения изображений позволил подчеркнуть текстуру холста в выбранном направлении.

Холсты картин имеют ряд особенностей, которые затрудняют импользование алгоритмов, созданных для контроля качества продукции текстильного производства, и алгоритмов анализа рентгеновских изображений холстов картин.

В данной работе для анализа изображений образцов применялись предложенные модификации известного подхода, основанного на фильтрации в Фурье-области и пороговой бинаризации. Для пороговой бинаризации фильтрованных изображений предложено использовать глобальный и локальный алгоритмы, максимизирующие взаимную информацию между полутоновым и бинаризованным изображениями (MIMax-based и Local MIMax-based). Разработаны процедуры обработки исходных, фильтрованных и бинаризованных изображений холстов. Предложенные алгоритмы сравнивались с известными алгоритмами на основе методов пороговой бинаризации Отсу (Otsu-based) и Ниблэка (Niblack-based), см. [1].

Для проверки эффективности предложенных алгоритмов проведен вычислительный эксперимент. Эксперимент включает два этапа. На первом этапе определяется диапазон пространственного разрешения изображений образцов, на котором достигается наибольшая точность подсчета нитей. На втором этапе указанные выше алгоритмы применяются к изображениям образцов холста,

полученных при выбранном на первом этапе диапазоне пространственного разрешения. Измерения количества нитей основы проводились в 33 изображениях, а количества нитей утка - на 42 избражениях. Для оценки точности алгоритмов полученные значения числа нитей сравнивались с результатами подсчета, выполненными экспертами. Построены гистограммы значений относительных ошибок, вычислены их средние значения и дисперсии (см. таблицу). Проведено сравнение результатов подсчета нитей алгоритмами на основе максимума взаимной информации и известными алгоритмами, разработанными для контроля качества ткани в текстильном производстве.На практике плотность холста измеряется экспертами в количестве нитей, приходящихся на единицу длины в направлении основы или утка. При оценивании плотности холстов в направлении нитей основы алгоритм на базе метода пороговой бинаризации Отсу в $92,8\%$ случаев продемонстрировал ошибку в пределах одной нити на сантиметр, алгоритм на базе метода Ниблэка – в 83%, алгоритм на основе глобального критерия максимума взаимной информации – в 100% случаев, а алгоритм на основе локального критерия максимума взаимной информации – в 97% случаев. При оценивании плотности холстов в направлении нитей утка алгоритм на базе метода Отсу в 100% случаев показал ошибку более одной нити на сантиметр, алгоритм на базе метода Ниблэка в 60% изображений. Алгоритм на основе глобального критерия максимума взаимной информации в 88% случаев обеспечил ошибку, не превосходящую одной нити на сантиметр, а на основе локального критерия максимума взаимной информации – в 95% случаев. Результаты проведенных исследований позволили сделать следующие выводы. Алгоритмы подсчета нитей на изображениях холстов картин, использующие глобальный и локальный методы пороговой бинаризации на основе критерия максимума взаимной информации и разработанные процедуры пред- и постобработки, являются более эффективным по сравнению с методами, созданными для контроля качества текстильного производства. Алгоритм подсчета нитей на основе критерия максимума взаимной информацип обеспечивает точность на уровне известных алгоритмов измерения плотности холста картин (например, методам подсчета нитей по рентгеновским изображениям холста). При этом предложенные алгоритмы не требуют специального оборудования, необходимого для получения рентгеновских снимков.

Дальнейшие исследования будут направлены на повышение точности подсчета нитей и разработку методов измерения других параметров холстов картин по изображениям.

Работа выполнена при частичной поддержке РФФИ, гранты №18-07-01385 и №18-07-01231.

[1] *Мурашов Д. М., Березин А. В., Иванова Е. Ю.* Определение количества нитей холстов картин по изображениям, полученным при направленном освещении // Машинное обучение и анализ данных, 2018.

Algorithms based on the mutual information maximization for measuring number of threads from painting canvas images

Dmitry Murashov[1]⋆ d_murashov@mail.ru
Aleksey Berezin[2] berezin_aleks@mail.ru
Ekaterina Ivanova[3] ivanova-e-u@yandex.ru
[1]Moscow, FRC CSC RAS
[2]Moscow, State Historical Museum
[3]Moscow, Glazunov Academy

This work deals with the problem of canvas threads counting in images of paintings. Counting of threads is necessary for measuring canvas density and a number of other parameters used by art historians for dating the artworks. To emphasize canvas texture here we use images acquired in raking light. We improve known techniques developed for inspecting fabrics in the textile industry. Two new threads counting algorithms based on filtering in the Fourier domain and mutual information maximization thresholding techniques (global and local MIMax) are proposed and tested. These algorithms are compared with two known algorithms based on the global Otsu thresholding technique (Otsu-based) and the local Niblack algorithm (Niblack-based) (see [1]). To evaluate the effectiveness of the described above algorithms, a computational experiment is carried out. The experiment includes two stages. At the first stage, the spatial resolution of the sample images, at which the highest accuracy of counting the threads can be achieved, is determined. At the second step, to count the warp and weft threads the described above algorithms are applied to the images of the canvas samples obtained at the selected optimal spatial resolution. To estimate the accuracy of the algorithms, the obtained values of the number of threads are compared with the results of counting performed by experts. We compute relative error values and obtain means and standard deviations (see the table below). In practice, the density of the canvas is measured by experts in the number of threads per unit length in the direction of the warp or weft.

When estimating the density of canvases in the direction of the warp threads, the algorithm based on the Otsu method in 92.8% of cases showed the error within one thread per centimeter, the algorithm based on the Niblack method - in 83% of cases, the global algorithm based on the mutual information maximization - in 100%, and the method using the local MIMax technique – in 97% of cases. When estimating the canvas density in the direction of weft threads, the algorithm based on the Otsu method always showed the error of more than one thread per centimeter, and the algorithm based on the Niblack method - for 60% of samples. The method based on the global MIMax criterion in 88% of cases provided the error not exceeding one thread per centimeter, and the method using the local MIMax technique – in 95% cases.

From the results of the experiment, one can conclude that the technique based on the global MIMax method is more efficient for counting warp threads, and the

technique based on the local MIMax method is more accurate for counting the weft threads. These algorithms for measuring the canvas density from images taken in raking light are efficient in cases when the analysis of canvas images acquired in X-rays and transmitted light is ineffective. The results of the experiment show that the accuracy of the proposed threads counting algorithms is comparable to the accuracy of known techniques.

The future research will be aimed at improving the accuracy of counting threads, developing methods for measuring other parameters of canvases of paintings.

This research is partially funded by RFBR, grants 18-07-01385 and 18-07-01231.

[1] *Murashov D., Berezin A., Ivanova E.* Painting canvas thread counting from images obtained in raking light // Machine Learning and Data Analysis, 2018.

Автоматическое совмещение изображений в задачах улучшенного и комбинированного видения с использованием генеративных состязательных сетей

Визильтер Юрий Валентинович[1] viz@gosniias.ru
Выголов Олег Вячеславович[1] o.vygolov@gosniias.ru
Доброходов Константин Викторович[1] konstantin.dobrokhodov@gmail.com
Лебедев Максим Алексеевич[1] mlebedev@gosniias.ru
Неклюдов Семен Александрович[1]⋆ neklyudov.semen97@gmail.com

[1]Москва, ФГУП «Государственный научно-исследовательский институт авиационных систем»

В современных системах непрерывного визуального представления закабинного пространства, повышающих ситуационную осведомленность экипажа воздушного судна, обрабатываются данные разной физической природы. Так, например, в системах улучшенного видения изображения от датчиков технического зрения (как правильно, телевизионных и инфракрасных камер) проходят фильтрацию, обработку и комплексирование специальными алгоритмами, а в системах комбинированного видения изображение закабинной обстановки формируется за счет интегрального представления реального от оптического датчика и синтезированного на основе пилотажно-навигационной информации (ПНИ) изображений.

В данной работе предлагается оригинальная архитектура генеративной состязательной нейронной сети, основанной на архитектуре LinkNET, позволяющей объединять изображения разной физической природы.

Работа выполнена при поддержке РФФИ, грант 18-07-01275А, и РНФ, грант № 19-11-11008.

[1] *Визильтер Ю. В., Выголов О. В., Доброходов К. В., Лебедев М. А., Неклюдов С. А.* Автоматическое совмещение изображений в задачах улучшенного и комбинированного видения с использованием генеративных состязательных сетей // Вестник компьютерных и информационных технологий, Москва: ООО «Издательский до «Спектр», 2019. — (принято в печать).

Automatic Images Fusion in Aviation Enhanced and Combined Vision Systems Using Generative Adversarial Networks

Yuriy Vizilter[1] viz@gosniias.ru
Oleg Vygolov[1] o.vygolov@gosniias.ru
Konstantin Dobrokhodov[1] konstantin.dobrokhodov@gmail.com
Maksim Lebedev[1] mlebedev@gosniias.ru
Semen Neklyudov [1]⋆ neklyudov.semen97@gmail.com

[1]Moscow, The Federal State Unitary Enterprise "State Research Institute of Aviation Systems"(FGUP "GosNIIAS")

In modern systems of continuous vision view of the cockpit space, which increase the situational awareness of the aircraft crew, data of different physical nature are processed. For example, special algorithms in enhanced vision systems are filtered, processed and integrated images from technical vision sensors (television or infrared cameras), and images outside the cockpit in combined vision systems are formed using integral representation of real image from optical sensor and image, which was synthesized based on navigation data.

In this paper, we propose the original architecture of a generative adversarial neural network based on the LinkNET architecture, which allows fusion images of different physical nature. The high quality of the fusion image is achieved by denoising, debluring and geometric mismatch of input images due to errors in the data of navigation data.

This work was performed with the support of RFBR, grant 18-07-01275A, and RSF, grant 19-11-11008.

[1] *Vizilter Yu., Vygolov O., Dobrokhodov K., Lebedev M., Neklyudov S.* Automatic Images Fusion in Aviation Enhanced and Combined Vision Systems Using Generative Adversarial Networks // Herald of computer and information technologies, Moscow: Publishing house "Spektr", 2019. — (in printing).

Алгоритм стабилизации видео с выбором ведущей группы движений с сохранением размерности кадра

Семенов Павел Владимирович[1]⋆ semenov.pv71@yandex.ru
Князев Денис Викторович[1] denis.denis-knyazev2018@yandex.ru
Копылов Андрей Валерьевич[1] av.kopylov@yandex.ru
[1]Тула, Тульский государственный университет

Задача стабилизации видео заключается в устранении нежелательного межкадрового смещения, которое появляется в результате непреднамеренного движения камеры. Широкое распространение промышленных систем, а также мобильных устройств с функцией видеонаблюдения и дальнейший анализ видеопоследовательности выдвигают повышенные требования к качеству получаемого изображения, в тоже время далеко не все устройства оснащены аппаратными системами стабилизации.

На данный момент существуют различные методы программной стабилизации видео, основанные на фильтрации Калмана и оценках движения камеры. В общем случае, все алгоритмы сводятся к поиску ключевых особенностей на соседних кадрах, их сопоставлению и сглаживанию движения. Эти методы дают хорошие результаты, но лишены важной особенности – они основаны на отделении глобального межкадрового движения от локального и не позволяют отделять локальные группы движений, принадлежащих различным объектам на кадре, друг от друга. В процессе стабилизации видео, вследствие применения межкадровых преобразований возникает проблема появления областей, для которых изображение на текущем преобразованном кадре отсутствует. Наличие этих областей ухудшает восприятие информации, содержащейся в видео. Одним из наиболее популярных методов является Motion Inpainting. Данный метод позволяет заполнять кадры как с динамическими, так и со статичными сценами. Также существуют методы, основанные на оценке глобального преобразования. Данные методы требуют проведения весьма громоздких вычислений и не подходят для стабилизации в реальном масштабе времени. Альтернативой является обрезка всех кадров под один размер, однако данный метод приводит к потере информации.

В данной работе мы предлагаем ввести кластеризацию в двумерном пространстве, сформированном векторами движения, а также метод позволяющий воссоздать кадр в полном размере без значительной потери информации [1]. Дополнительный шаг кластеризации, позволит нам выбирать ведущую группу векторов, движение которой мы хотим стабилизировать. В данной работе для достижения лучших результатов мы применили более точный способ кластеризации для поиска минимального покрывающего дерева. В основе данного метода лежит алгоритм Прима. В результате вычисляется матрица аффинных преобразований, которая применяется к текущему кадру для устранения неже-

лательного движения. После этого преобразования мы получаем области, для которых изображение на текущем кадре отсутствует.

Наш метод для воссоздания кадра, не обладает большой вычислительной сложностью. Для его работы сначала, посредством обратного аффинного преобразования мы представляем кадр с неполной информацией в системе координат заполненного кадра (как правило предыдущего). Затем посредством поиска точек пересечения крайних отрезков кадров, мы выделяем незаполненную область. После определения незаполненной области начинается второй этап алгоритма – совмещение сигналов интенсивностей. Для этого необходимо рассчитать интенсивность всех граничных пикселей заполненной области кадра с неполной информацией и соответствующих неопределенной области граничных пикселей с заполненного кадра. В результате будет получено два сигнала изменения интенсивности пикселей на границах соответствующих областей. Сигнал с заполненного кадра подвергается «эластичной» трансформации вдоль границы заполняемой области так, чтобы значения сигналов совпали как можно лучше. Для решения данной задачи использован быстрый алгоритм на основе динамического программирования. Таким образом, на стыке кадров будет правильно подобрана интенсивность каждого пикселя и динамические объекты сцены не будут деформироваться и искажаться. В результате на выходе получится полноценный кадр стабилизированного видео без потери информации и при минимальных вычислительных затратах, позволяющих осуществить видеостабилизацию в реальном масштабе времени.

Для работы метода стабилизации требуется относительно небольшое число ключевых особенностей на кадре и, следовательно, векторов движения.

Работа выполнена при поддержке РФФИ, гранты № 18-07-00942, 18-07-01087.

[1] *Семенов П. В., Князев Д. В., Копылов А. В.* Алгоритм стабилизации видео с выбором ведущей группы движений с сохранением размерности кадра // Известия ТулГУ, Технические науки Тула, Издательство ТулГУ, 2019. (в печати).

Video stabilization algorithm with selection of the leading group of movements with preservation of frame dimension

Semenov Pavel semenov.pv71@yandex.ru
Knyazev Denis denis.denis-knyazev2018@yandex.ru
Kopylov Andrey av.kopylov@yandex.ru
Tula, Russia, TulSU

The problem of video stabilization is to eliminate unwanted inter-frame bias that results from casual camera movement. The widespread use of mobile devices with video surveillance function and further analysis of video sequences put forward increased demands on the quality of the resulting image, while at the same time, not all devices are equipped with hardware stabilization systems.

At the moment, there are various methods of software video stabilization based on Kalman filtering and camera movement estimates. In the general case, all algorithms are reduced to searching for key features on adjacent frames, comparing them and smoothing the motion. These methods give good results, but devoid an important feature - they are based on the separation of global interframe movement from local and do not allow to separate local groups of movements belonging to different objects in the frame from each other. In the process of video stabilization, due to the use of inter-frame transformations, a problem arises of the appearance of missed areas. These areas affect the perception of the information contained in the video. One of the most popular methods is Motion Inpainting. This method allows you to fill frames with both dynamic and static scenes. There are also methods based on the estimation of global transformation. These methods require multiple calculations and are not suitable for stabilization in real time. An alternative is to crop all frames to the same size, but this method leads to loss of information.

In this paper, we propose introducing clustering in a two-dimensional space formed by motion vectors, as well as a method that allows us to recreate a frame in full size without significant loss of information [1]. This step of clustering will allow us to select a leading group of vectors whose motion we want to stabilize. In this work, for best results, we have used the accurate clustering method to find the minimum spanning tree. This method is based on the Prim algorithm. As a result, the matrix of affine transformations is calculated, which is applied to the current frame to eliminate unwanted movement. After this conversion, we get areas for which the image on the current frame is missing.

Our method for reconstructing a frame does not have much computational complexity. For the algorithm execution, firstly, through the inverse affine transformation, we present a frame with incomplete information in the coordinate system of the filled frame (usually the previous one). Then, by searching for the intersection points of the frames, we select a blank area. After determining the blank area, the second stage of the algorithm begins - the combination of intensity signals. For this, it is necessary to calculate the intensity of all boundary pixels of the filled region

of the frame with incomplete information and the corresponding indefinite region of boundary pixels from the filled frame. As a result, two signals of changes in pixel intensity at the boundaries of the respective regions will be obtained. The signal from the filled frame undergoes an "elastic" transformation along the boundary of the filled area so that the signal values match as best as possible. To solve this problem, a fast algorithm based on dynamic programming was used. Thus, at the junction of the frames, the intensity of each pixel will be correctly selected and the dynamic objects of the scene will not be deformed and distorted. As a result, the output will be a full-size frame of stabilized video without loss of information and with minimal computational costs, allowing real-time video stabilization.

The stabilization method requires a relatively small number of key features on the frame and, therefore, motion vectors.

This research is funded by RFBR, grants 18-07-00942, 18-07-01087.

[1] *Semenov P. V., Knyazev D. V., Kopylov A. V.* Video stabilization algorithm with selection of the leading group of movements with preservation of frame dimension // News of TulSU, Technical sciences of Tula, Publishing house of TulSU, 2019. (in press)

Вычислительно эффективный алгоритм распознавания изображения на основе последовательного анализа главных компонент нейросетевых признаков

Соколова Анастасия Дмитриевна[1]⋆ adsokolova96@mail.ru
Савченко Андрей Владимирович[1] avsavchenko@hse.ru

[1]Нижний Новгород, Национальный исследовательный университет Высшая школа экономики, Лаборатория алгоритмов и технологий анализа сетевых структур

В задаче классификации изображений необходимо поступающему на вход изображению поставить в соответствие один из $C > 1$ заранее точно неопределенных классов. Классы задаются с помощью обучающего множества из $R \geqslant C$ эталонных изображений с известной меткой класса $c(r) \in \{1, ..., C\}$. Для многих задач распознавания изображений создание большой ($R \gg C$) базы данных эталонов, необходимой для обучения глубоких сверточных нейронных сетей (СНС), является слишком дорогостоящей процедурой. В таком случае приходится ограничиваться малой обучающей выборкой ($R \approx C$), при этом предварительно обученная на других наборах данных (например, ImageNet-1000) СНС применяется для извлечения D-мерного вектора признаков входного изображения $\mathbf{x} = [x_1, ..., x_D]$. Окончательное принятие решения может осуществляться с помощью классификаторов, обученных с использованием векторов признаков эталонных изображения $\mathbf{x}_r = [x_{r;1}, ..., x_{r;D}], r \in \{1, ..., R\}$. Зачастую наибольшую точность классификации при очень малом числе изображений каждого класса достигается с помощью методов ближайших соседей (k-NN). Использование такого подхода порождает новые проблемы с вычислительной сложностью алгоритмов принятия решений, которая, в отличие от константной сложности прямого прохода (inference) нейронной сети, обычно линейно зависит от количества распознаваемых классов. Наиболее популярные способы повышения вычислительной эффективности (приближенный поиск ближайшего соседа и снижение размерности признаков) часто приводят к значимому снижению точности классификации.

Для преодоления отмеченной проблемы высокой вычислительной сложности распознавании изображений в работе [1] предложен новый алгоритм на основе последовательного анализа векторов признаков высокой размерности. На предварительном шаге выполнить анализ главных компонент векторов признаков эталонных изображений. Далее входной вектор признаков $\mathbf{x}$ преобразуется в вектор главных компонент $\tilde{\mathbf{x}} = [\tilde{x}_1, ..., \tilde{x}_D]$, после чего последовательно анализируется иерархическое представление признаков так, что на каждом l-м уровне иерархии ($l \in \{1, ..., L\}$) сопоставляются только d_l главных (первых) компонент. Количество компонент на каждом уровне иерархии d_l определяется, исходя из фиксированной доли объясненной дисперсии $\sigma_l^2 \in (0; 1]$ обучающего множества, при этом $\sigma_1^2 < \sigma_2^2 < ... < \sigma_L^2 = 1$. Если применяемая в методе k-NN мера близости является аддитивной, то можно использовать расстояния, вычисленные

на предыдущем уровне. Тогда может быть применен метод ближайшего соседа с использованием расстояния $\rho_c(\tilde{\mathbf{x}}^{(l)})$ между входным объектом и каждым классом c на l-м уровне из множества кандидатов C_l:

$$c_l^* = \arg\min_{c \in C_l} \rho_c(\tilde{\mathbf{x}}^{(l)}). \tag{1}$$

Множество C_l содержит вначале все классы: $C_1 = \{1, ..., C\}$. Векторы признаков рассматриваются как оценки вероятностных распределений случайных величин с D возможными значениями. Множество классов-кандидатов на каждом шаге уточняется следующим образом:

$$C_{l+1} = \left\{ c \in C_l \,\middle|\, \frac{\rho_c(\tilde{\mathbf{x}}^{(l)})}{\rho_{c_l^*}(\tilde{\mathbf{x}}^{(l)})} \leqslant \delta \right\}, \tag{2}$$

где $\delta \leqslant 1$ - фиксированный порог. Согласно этому выражению, на следующем иерархическом уровне проверяются только те классы, расстояния до которых не намного превышают минимальное расстояние до ближайшего соседа c_l^*. В противном случае анализируется большее число компонент на $(l+1)$-м уровне иерархии. Процесс повторяется до тех пор, пока не будут обработаны все компоненты вектора признаков.

Экспериментальное исследование было проведено для задачи распознавания лиц. Для извлечения признаков из изображения использовались 4 свободно доступных дескриптора лиц: VGGFace($D = 4096$), Lightened CNN версия C ($D = 256$), VGGFace2: ResNet50 модель ($D = 2048$), FaceNet ($D = 512$). Для применения косинусной меры близости осуществлялась нормализация вектора признаков в метрике Евклида (L_2), поэтому для определения использовался удовлетворяющий условию аддитивности квадрат расстояния Евклида для главных компонент. Значение порога для отношений рассогласований выбрано равным $\delta = 0.7$.

Эксперименты проводились для нескольких широко известных наборов данных. Результаты распознавания для Labeled Faces in the Wild (LFW) приведены в Таблице 1). Здесь время принятия решения предложенного алгоритма в 10 раз меньше, чем время традиционного подхода поиска ближайшего соседа.

Таблица 1. Результаты распознавания изображений лиц, набор данных LFW

	Классификатор	VGGFace	LCNN	VGGFace2	FaceNet
Точность (%)	k-NN, все признаки	96.31	97.48	98.66	98.15
	k-NN, 64 признаков	94.10	96.21	96.95	97.36
	Предложенный подход	95.80	96.81	97.98	97.94
Время (мс)	k-NN, все признаки	50.39	4.88	34.78	9.37
	k-NN, 64 признаков	2.53	2.50	2.53	2.52
	Предложенный подход	3.20	2.23	2.81	1.87

Статья подготовлена в результате проведения исследования (№ 19-04-004) в рамках Программы «Научный фонд Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ)» в 2019 г. и в рамках государственной поддержки ведущих университетов Российской Федерации "5-100".

[1] *Sokolova A. D., Savchenko A. V.* Fast Nearest-Neighbor Classifier based on Sequential Analysis of Principal Components // In International Conference on Analysis of Images, Social Networks and Texts, Springer, Cham, 2019.

Efficient image recognition with sequential analysis of principal components of off-the-shelf CNN features

Anastasiia Sokolova[1]⋆ adsokolova96@mail.ru
Andrey Savchenko[1] avsavchenko@hse.ru
[1]Nizhny Novgorod, National Research University Higher School of Economics, Laboratory of Algorithms and Technologies for Network Analysis

In the image recognition task it is required to assign an input image into one of $C > 1$ pre-undefined classes. The classes are specified by a training set of $R \geqslant C$ reference images with known class label $c(r) \in \{1, ..., C\}$. The creation of enormous ($R \gg C$) database which is required for deep convolutional neural networks (CNN) training is too expensive procedure for many recognition tasks. Therefore, one can use only a small training set ($R \approx c$) and CNN which was pre-trained on other datasets (such as ImageNet-1000) in order to extract D-dimensional feature vector of the input image $\mathbf{x} = [x_1, ..., x_D]$. The final decision can be made by a classifier trained using feature vectors of reference images $\mathbf{x}_r = [x_{r;1}, ..., x_{r;D}], r \in \{1, ..., R\}$. The highest classification accuracy for a very small number of images of each class is often achieved by the nearest neighbors methods (k-NN). The use of such approach raises new problems with the computational complexity of decision-making algorithms which linearly depends on the number of recognized classes in contrast to the constant complexity of direct inference in a neural network. Unfortunately, the most popular ways of increasing computational efficiency, namely, an approximate nearest neighbor search and reducing the dimension of features, often lead to a significant decrease in the classification accuracy.

In order to overcome the noted problem of high computational complexity of image recognition the novel approach was suggested in the paper [1] based on a sequential analysis of high-dimensional feature vectors. On the first stage the principal component analysis of reference feature vectors is implemented. Then input feature vector $\mathbf{x}$ is transformed into the vector of principal components $\tilde{\mathbf{x}} = [\tilde{x}_1, ..., \tilde{x}_D]$. After that we proposed to sequentially process a hierarchy of features, so that at each l-th level of hierarchy ($l \in \{1, ..., L\}$) only d_l first principal components are matched. The number of components at each hierarchy level d_l is chosen in order to explain the variance rate $\sigma_l^2 \in (0; 1]$ for the training set so that $\sigma_1^2 < \sigma_2^2 < ... < \sigma_L^2 = 1$. If the dissimilarity measure is additive, we can utilize the distances from the previous level to speed-up the matching. Then the k-NN method is applied using the distance $\rho_c(\tilde{\mathbf{x}}^{(l)})$ between input object and every class c at l-th level from the set of candidates C_l:

$$c_l^* = \arg\min_{c \in C_l} \rho_c(\tilde{\mathbf{x}}^{(l)}). \quad (1)$$

The set of candidates C_l initially contains all subjects: $C_1 = \{1, ..., C\}$. We treat the feature vectors as the estimates of probability distributions of random variables with D possible values. We propose to refine the set of candidates as follows:

$$C_{l+1} = \left\{ c \in C_l \,\middle|\, \frac{\rho_c(\tilde{\mathbf{x}}^{(l)})}{\rho_{c_l^*}(\tilde{\mathbf{x}}^{(l)})} \leqslant \delta \right\}, \tag{2}$$

where $\delta \leqslant 1$ is a fixed threshold. According to this expression, at the next hierarchical level only those classes are checked, distances to which are not much higher than the minimal distance to the nearest neighbor c_l^*. Otherwise, we increase the representation level $l + 1$ and the number of components d_{l+1} in order to compute the distances between new features. It can continue as long as we process the full feature vector.

The experimental study is devoted to unconstrained face recognition. Four publicly available CNNs are used for feature extraction: VGGFace($D = 4096$), Lightened CNN version C ($D = 256$), VGGFace2: ResNet50 model ($D = 2048$), FaceNet ($D = = 512$). To apply the cosine similarity measure the feature vectors are normalized in the Euclidean metric (L_2). Therefore, the Euclidean distance square which satisfies the additivity condition is used for principal component definition. The threshold value for the mismatch relations is chosen equal to $\delta = 0.7$.

The experiments were conducted for several popular datasets. The recognition results for the Labeled Faces in the Wild (LFW) dataset are demonstrated in (Table 1). The decision-making time in the proposed approach is 10-times lower than the conventional implementation of the k-NN.

Table 1. Face recognition results for the k-NN classifier, LFW dataset

	Classifier	VGGFace	LCNN	VGGFace2	FaceNet
Accuracy (%)	k-NN, all features	96.31	97.48	98.66	98.15
	k-NN, 64 features	94.10	96.21	96.95	97.36
	Proposed approach	95.80	96.81	97.98	97.94
Time (ms)	k-NN, all features	50.39	4.88	34.78	9.37
	k-NN, 64 features	2.53	2.50	2.53	2.52
	Proposed approach	3.20	2.23	2.81	1.87

This research was prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economics (HSE University) in 2019 (grant No. 19-04-004) and within the framework of the Russian Academic Excellence Project "5-100".

[1] *Sokolova A. D., Savchenko A. V.* Fast Nearest-Neighbor Classifier based on Sequential Analysis of Principal Components // In International Conference on Analysis of Images, Social Networks and Texts, Springer, Cham, 2019.

Метод встраивания криптографического ключа в биометрический эталон радужной оболочки глаза

Зайнулина Эльвира Талиповна[1,2]⋆ zaynulina.et@phystech.edu
Матвеев Иван Алексеевич[2] matveev@ccas.ru
[1]Москва, Московский физико-технический институт (НИУ)
[2]Москва, Федеральный исследовательский центр «Информатика и управление» РАН

Представлен метод встраивания криптографического ключа в биометрические данные, дающий возможность передавать полученный код по открытым каналам и впоследствии извлекать ключ при предъявлении биометрии владельца. В качестве биометрических данных используются эталоны радужной оболочки глаза. Метод встраивания (кодер) представляет собой четырёхступенчатую схему, состоящую из последовательных кодирования Рида-Соломона, кодирования Адамара, кодирования дублированием битов и псевдослучайного перемешивания битов, полученный код побитово суммируется с биометрическим эталоном по модулю 2. Декодер выполняет операции в обратной последовательности. Подбор параметров схемы осуществляется с помощью решения задачи оптимизации, состоящей в том, чтобы при некотором фиксированном пороге для коэффициента ложного допуска (FAR) минимизировать значение коэффициента ложного отказа в допуске (FRR); при этом существуют ограничения на минимальную длину криптографического ключа и максимальный размер итогового кода.

Проведены эксперименты на базе данных ICE (2954 изображения). В результате при $\mathrm{FAR} = 2\times10^{-2}\%$ удалось добиться $\mathrm{FRR} = 2.7\%$. Планируется провести тесты на других базах данных.

Работа поддержана грантом РФФИ №19-07-01231.

[1] *Зайнулина Э. Т., Матвеев И. А.* Метод встраивания криптографического ключа в биометрический эталон радужной оболочки глаза // Машинное обучение и анализ данных. 2019.

Method of embedding a cryptographic key in the biometric iris template

Elvira Zainulina[1,2]⋆ zaynulina.et@phystech.edu
Ivan Matveev [2] matveev@ccas.ru
[1]Moscow, Moscow Institute of Physics and Technology (National Research University)
[2]Moscow, Federal Research Center "Computer Science and Control" of RAS

A method of embedding a cryptographic key in biometric data is presented, which makes it possible to transmit the obtained code through open channels and subsequently retrieve the key upon presentation of the owner's biometrics. The templates of the iris are used as biometric data. The embedding method (encoder) is a four-stage scheme consisting of sequential Reed-Solomon coding, Hadamard coding, bit duplication coding and pseudo-random bit mixing, the resulting code is bitwise summed with the biometric standard modulo 2. The decoder performs the operations in the reverse order. The selection of the parameters of the scheme is carried out using the solution of the optimization problem, which consists in minimizing the value of the coefficient of false denial of tolerance (FRR) for a certain threshold for the false tolerance coefficient (FAR); at that there are restrictions on the minimum length of the cryptographic key and the maximum size of the resulting code.

Experiments were conducted on the ICE database (2954 images). As a result, at FAR $= 2 \times 10^{-2}\%$, FRR $= 2.7\%$ was achieved. It is planned to conduct tests on other databases.

This research is funded by RFBR, grant 19-07-01231.

[1] *Zainulina E. T., Matveev I. A.* Method of embedding a cryptographic key in the biometric iris template // Machine Learning and Data Analysis, 2019.

Развитие обобщенной схемы классификации алгоритмов сегментации изображений

Ханыков Игорь Георгиевич[1]⋆ igk@iias.spb.su

[1]Санкт-Петербург, Институт информатики и автоматизации Российской академии наук

Схемы классификации алгоритмов сегментации изображений (АСИ) можно разделить на категории: смешанные, специализированные, классификации по единственному признаку и обобщенные классификации. Последние представляют особый интерес, поскольку они позволяют однозначно классифицировать существующие АСИ; предсказывают появление и формируют требования к новым АСИ.

Классификационный признак принимает одно из двух доступных значений. Пара признаков формирует уровень. Несколько уровней формируют обобщенную классификационную схему. По способу обработки изображений АСИ классифицируются либо на группы нахождения области по свойствам схожести, либо нахождения границ по свойствам различия. По стратегии исполнения АСИ классифицируются на группу последовательного и группу параллельного исполнения вычислительных операции. По типу изображения АСИ делится на группы, обрабатывающие либо цветные, либо полутоновые изображения. Четвертый признак – наличие критерия качества – разделяет АСИ на группы с критерием качества и без такового.

В работе вводится дополнительный признак – число разбиений исходного изображения на выходе алгоритма – для разделения групп алгоритмов, генерирующих единственное разбиение и множество. У первой группы АСИ в ходе вычислительного процесса число однородных по некоторой характеристике множеств (сегментов, кластеров) фиксировано. У второй группы число однородных множеств варьируется в некотором диапазоне.

Исследование выполнено при поддержке НИР № 0073-2018-0001 „Состояние и перспективы развития информационного общества в России“ с 2014 г. по 2021 г.

[1] *Ханыков И. Г.* Классификация алгоритмов сегментации изображений // Санкт-Петербург: Изв. вузов. Приборостроение, 2018. С. 978–987.

The Development of Generalized Classification Scheme for Image Segmentation Algorithms

Igor Khanykov[1]⋆ igk@iias.spb.su

[1]Saint Petersburg, Institute for Informatics and Automation of the Russian Academy of Sciences

The classification schemes for image segmentation algorithms (ISA) can be divided into categories: mixed, specialized, classifications on a single basis, and generalized classifications. The latter are of particular interest, since they make it possible to unambiguously classify existing ISA; predict the appearance and formulate the requirements for the new ISAs.

The classification attribute takes one of two available values. A pair of attributes forms a level. Several levels form a generalized classification scheme. According to the method of image processing, ISAs are classified either into groups for finding a region according to similarity properties, or for finding boundaries according to difference properties. According to the execution strategy, ISAs are classified into a group of sequential and a group of parallel execution of computational operations. By type of image, ISAs are divided into groups that process either color or grayscale images . The fourth attribute - the presence of a quality criterion - divides ISAs into groups with and without a quality criterion.

An additional attribute is introduced in the paper — the number of partitions of the original image at the output of the algorithm — to separate algorithms into the groups with a single partition and with multiple partitions at output. The number of homogeneous sets by some characteristic (segments, clusters) of the first ISA group is fixed during the computational process. The number of homogeneous sets of the second group varies in a certain range.

The study was supported by the research work No 0073-2018-0001 “The State and Prospects of the Development of the Information Society in Russia” from 2014 to 2021.

[1] *Khanykov I. Klassifikaciya algoritmov segmentacii izobrazheniy* [The Classification of Image Segmentation Algorithms] // Journal of Instrument Engineering, Saint Petersburg. 2018. Vol.61, N.11. P. 978–987. (In Russian).

Деревья и леса решений, основанные на сходстве, в задачах анализа КТ изображений

Бериков Владимир Борисович[1,2]⋆ berikov@math.nsc.ru
Пестунов Игорь Алексеевич[3] pestunov@ict.sbras.ru
Козинец Роман Максимович[2] romanec1954@gmail.com
Рылов Сергей Александрович[3] rylovs@mail.ru
[1]Новосибирск, Институт математики им. С. Л. Соболева СО РАН
[2]Новосибирск, Новосибирский государственный университет
[3]Новосибирск, Институт вычислительных технологий СО РАН

Предложен метод распознавания образов с применением модификации класса логических решающих функций, представленных в виде дерева решений. Вместо стандартных высказываний, соответствующих вершинам дерева, в которых проверяется принадлежность некоторой переменной тем или иным множествам ее значений, используется более общий тип высказываний относительно близости рассматриваемой точки к различным подмножествам наблюдений. При этом для определения степени похожести могут выбираться различные метрики и подпространства признаков. Этот тип дерева решений позволяет получить более сложные границы принятия решений, которые в то же время имеют понятную пользователю логическую интерпретацию. Рассмотрено несколько стратегий построения решения: на основе преобразования данных с использованием опорных точек и с использованием набора деревьев. Метод применен для анализа томографических изображений. Эксперименты показали, что предложенный алгоритм, в условиях малой обучающей выборки, дает более точные прогнозы, чем ряд других известных алгоритмов, в том числе с использованием глубокой сверточной нейронной сети.

Работа поддержана грантами РФФИ 18-07-00600а, 19-29-01175.

[1] *Berikov V, Pestunov I., Kozinets R., Rylov S.* Similarity-based decision tree induction method and its application to cancer recognition on tomographic images // Journal of Physics: Conference Series. 2019. (in press)

Similarity-based decision trees and forests in CT images analysis

Vladimir Berikov[1,2]⋆ berikov@math.nsc.ru
Igor Pestunov[3] pestunov@ict.sbras.ru
Roman Kozinets[2] romanec1954@gmail.com
Sergey Rylov[3] rylovs@mail.ru
[1]Novosibirsk, Sobolev Institute of mathematics SB RAS
[2]Novosibirsk, Novosibirsk State University
[3]Novosibirsk, Institute of Computational Technologies SB RAS

The paper proposes a pattern recognition method using a modification of the class of logical decision functions presented in the form of decision tree. Instead of standard statements corresponding to the tree nodes, in which a variable is tested for a certain set of its values, a more general type of statements is used regarding the similarity of the point in question to different subsets of the observations. At the same time, to determine the degree of similarity, various metrics and subspaces of features can be used. This type of decision tree allows one to obtain more complex decision boundaries, which at the same time have a clear logical interpretation for the user. Several tree induction strategies are considered based on data transformation using support points and with a collection of trees. The method is experimentally investigated on the problem of tomographic images analysis. Experiments have shown that the proposed method gives more accurate predictions in the condition of small training sample size than a number of other known classifiers and deep convolutional neural network.

The work is supported by RFBR projects 18-07-00600a, 19-29-01175.

[1] *Berikov V, Pestunov I., Kozinets R., Rylov S.* Similarity-based decision tree induction method and its application to cancer recognition on tomographic images // Journal of Physics: Conference Series. 2019. (in press)

Построение двухступенчатого линейно-нелинейного фильтра для восстановления и коррекции изображений

Фурсов Владимир Алексеевич[1,2]⋆ fursov@ssau.ru
Гошин Егор Вячеславович[1,2] goshine@yandex.ru
Медведева Ксения Сергеевна[1] aksiniyame@gmail.com

[1]Самара, Самарский национальный исследовательский университет имени академика С.П. Королева
[2]Самара, Институт систем обработки изображений РАН – филиал ФНИЦ «Кристаллография и фотоника» РАН

В настоящей работе мы рассматриваем двухэтапную технологию повышения четкости изображений. На первом этапе обработка изображения осуществляется с использованием линейного квадратично-экспоненциального (SE) фильтра с частотной характеристикой, обладающей центральной симметрией. На втором этапе проводится нелинейная коррекция. Идея фильтра заключается в том, чтобы увеличить значение центрального отсчета опорной области, если он находится на границе существенно различающихся уровней интенсивности.

Технология двухэтапного линейно-нелинейного фильтра описана в нашей статье [1]. Эта технология направлена на повышение качества изображений в мобильных устройствах. Доклад по этой теме представлен и принят на 12-й Международной конференции по машинному видению (ICMV-2019) (Амстердам, Нидерланды. 16-18 ноября 2019 г.). Результаты первых экспериментов, которые приведены в вышеуказанных работах, показали высокое качество восстановления и коррекции изображений. Однако в этих работах параметры фильтра на стадии нелинейной коррекции подбирались экспериментально.

В настоящей работе предлагается методика предварительного анализа исходных изображений для оптимизации параметров нелинейного фильтра. Приведены соотношения для оценки параметров фильтра и результаты экспериментов, подтверждающие эффективность методики.

Рассматриваемая технология повышения качества изображений реализуется при относительно низких вычислительных затратах, что открывает перспективы ее использования для обработки изображений, получаемых с помощью дифракционных оптических элементов в мобильных устройствах.

Работа была поддержана Российским фондом фундаментальных исследований (проекты № 17-29-03112 и № 18-07-01390), а также Министерством науки и высшего образования в рамках государственного задания.

[1] *Фурсов В. А. , Гошин Е. В. , Медведева К. С.* Технология повышения детализации изображений с нелинейной коррекцией высокоградиентных фрагментов // Компьютерная оптика, Самара: Т. 43, No 3, 2019. -— С. 484–491.

The build a two-stage linear-nonlinear filter to restore and correct of images

Vladimir Fursov[1,2]⋆ fursov@ssau.ru
Yelena Goshin[1,2] goshine@yandex.ru
Ksenia Medvedeva[2] aksiniyame@gmail.com
[1]Samara, Image Processing Systems Institute of RAS - Branch of the FRSC "Crystallography and Photopic" RAS
[2]Samara, Samara National Research University

In this study, we developed a two-stage technology for improving the sharpness of images. In the first stage, the correction was performed using a linear square exponential (SE) filter with a centrally symmetric frequency response in the form of quadratic and exponential functions. In the second stage, non-linear correction was carried out. The idea of the filter was to increase the impact of the central value, if it was at the edge of different intensity levels.

The technology of two-stage linear-nonlinear filter is described in our article [1]. The proposed technology was aimed at improving the quality of images in mobile devices. A report on this theme has been submitted to the 12th International Conference on Machine Vision (ICMV-2019) (Amsterdam, Netherlands. November, 16-18, 2019.). The results of the first experiments showed high quality of images restoration and correction. However, in previous works, the filter parameters in the nonlinear correction step were selected experimentally.

This paper proposes a technique for preliminary analysis of initial images in order to optimize parameters. The expressions for estimation of filter parameters and results of experiments confirm effectiveness of the proposed method are given. The ability to obtain more details in low-resolution images at low computational costs opens up the prospects for the use of mobile devices based on the diffraction optical elements.

This work was supported by the Russian Foundation for Basic Research under grant # 17-29-03112 and # 18-07-01390, and by the Ministry of Science and Higher Education within the State assignment.

[1] *Fursov VA, Goshin YeV, Medvedeva KS.* Technology of enhancing image detalization with nonlinear correction of highly gradient fragments. // Computer Optics, 2019.43(3) -— p.: 484 - 491.

Глобальный анализ изображений и детектирование и распознавание дорожной разметки в реальном времени

Досаев Роман Владимирович[1] romandosaev@gmail.com
Кий Константин Иванович[1]⋆ konst.kiy@site.ru
[1]Москва, Институт прикладной математики им. Келдыша РАН

В докладе предлагается новый подход к детектированию и распознаванию дорожной разметки. Данная тема является очень актуальной и ее приложения лежат в области создания советующих систем для помощи водителям (так называемым ADAS), внедряемым ведущими компаниями производителями автомобилей. Также данная тема очень важна для разработки систем управления беспилотных транспортных средств. Наличие нерешенных проблем в данной области подтверждается известными аварийными ситуациями с беспилотными автомобилями (часто со смертельными исходами) ведущих производителей. Особенно актуально решение данных задач для стран с состоянием дорог и климатическими условиями такими как в Российской Федерации. Наиболее современный обзор работ и полученных результатов в области детектирования и распознавания разметки опубликован в [1]. Первая публикация авторов на эту тему может быть найдена в [3]. В данной работе развиваются методы предложенные в [3] и рассматриваются такие задачи как обнаружение белой постоянной дорожной и временной окрашенной (в цвета от желтого до оранжевого-красного) и предлагается метод выделения временной дорожной разметки при наличии постоянной белой разметки. На этот счет нами не было найдено публикаций. Это связано с проблемами при работе с цветными изображениями в реальном времени [3]. Рассматриваются также вопросы, связанные с выделением стоп линий и разметки пешеходных переходов.

Предлагаемые методы опираются на глобальный метод анализа изображений в реальном времени, предложенный вторым автором [2]. Данный метод позволяет производить выделение различных объектов на цветных изображениях и анализировать их совместное поведение в реальном времени. В отличие от основных существующих методов сегментации и анализа изображений удается одновременно находить как большие объекты, так и малые контрастные объекты (сигнальные зоны автомобилей, летательных аппаратов, строительные конуса, знаки аварийной остановки и т.д.), и анализировать их совместное поведение в реальном времени.

Основным понятием является структура (граф) цветовых сгустков [4]. Структуры цветовых сгустков сжато и эффективно описывают изображения. Каждая разметка дает некоторый непрерывный объект в графе цветовых сгустков [3]. В работе [3] описаны эффективные алгоритмы построения непрерывных цепочек на графе цветовых сгустков, которые могут соответствовать дорожной разметке. Заметим, что при данном подходе удается выделять разметку на сильно искривленных ее частях, что выгодно отличает наш метод от методов,

основанных на поиске прямых линий с помощью преобразования Хафа. В настоящей работе предлагаются методы выделения кусков реальной разметки из множества построенных кандидатов на нее.

Написаны комплексы программ, реализующие разработанные методы. Произведено их тестирование на базе изображений и видео последовательностей из стандартных сайтов и съемок, проведенных на различных российских дорогах. Примеры обработки будут продемонстрированы.

Работа поддержана грантами РФФИ № 18-07-00127 и 19-08-01159.

[1] *Norote S. P., Bhujbal P. N., et al.* A review of recent advances in lane detection and departuture warning system // Pattern Recognition, Elsevier, 2018. — P. 216–34.

[2] *Dosaev R. V., Kiy K. I.* A new real-time method for finding temporary and permanent road marking and its applications // Proceedings CEUR, CEUR.org, 2019. vol. 2391 — P. 86–96

[3] *Kiy K. I.* A new method of global image analysis and its application in understanding road scenes // Pattern Recognit. Image Anal. , Pleades, 2018. —P. 483–94.

[4] *Kiy K. I.*. Segmentation and detection of contrast objects and their application in robot navigation // Pattern Recognit. Image Anal., Pleades,, 2015. — C. 338–46.

Global image analysis and detection and recognition of road marking in real time

Roman Dosaev[1] romandosaev@gmail.com
Konstantin Kiy[1]⋆ konst.kiy@site.ru
[1]Moscow,Keldysh institute of applied mathematics of RAS

In this paper, a new approach to detecting and recognizing road marking is proposed. This subject is very topical and its applications lie in the field of developing advising driver assistance systems (the so-called ADAS), introduced by leading automobile corporations. This subject is very important for developing control systems of pilotless vehicles. The presence of unsolved problems in this field is supported by known accidents with driverless vehicles (frequently with fatal accidents) of leading manufacturers. To solve these problems is especially important for countries with the state of roads and climatic conditions similar to those of the Russian Federation.

The most modern review of papers and obtained results in the field of detecting and recognizing road marking can be found in [1].

The first publication of the authors on this topic can be found in [2]. In this paper, methods proposed in [2] are developed and the problems of finding white permanent marking and temporary colored marking (painted in colors from yellow to orange-red) and a method for selecting temporary road marking under the presence of white permanent marking is proposed. We were not able to find any publications on this topic. It is connected with problems of processing large color images in real time when dealing [2]. The problems connected with detecting stop lines and pedestrian crossings are also considered.

The proposed methods lean on the method of global image analysis proposed by the second author [**?**]. This method allows one to detect various objects on color images and to analyze their joint behavior in real time. In contrast to the main existing methods for segmenting and analyzing images, it is possible to find simultaneously both big objects and small contrast objects (signal zones of vehicles and flying vehicles, construction cones, emergency triangles etc.) and to analyze their joint behavior in real time.

The basic notion is the structure of color bunches [2]. Structures of color bunches describe the image concisely and efficiently. Each road marking gives a certain continuous object in the graph of color bunches [2]. In [2] efficient algorithms for finding continuous chains on the graph of color bunches that may correspond to road markings are described. Note that in this approach it is possible to detect road marking in its parts with big curvature, which distinct our method from methods based on Hough transforms. In this paper, methods for selecting parts of real road marking from the set of candidates constructed are proposed.

In this paper, methods for selecting parts of real road marking from the set of candidates constructed.

Complexes of programs implementing the developed methods are written. They have been tested on an image base from images and video sequences from standard cites and those taken on various Russian roads. The results will be demonstrated.

This research is funded by RFBR, grants 18-07-00127 and 19-08-01159.

[1] *Norote S. P., Bhujbal P. N., et al.* A review of recent advances in lane detection and departuture warning system // Pattern Recognition, Elsevier, 2018. — P. 216–34.

[2] *Dosaev R. V., Kiy K. I.* A new real-time method for finding temporary and permanent road marking and its applications // Proceedings CEUR, CEUR.org, 2019. vol. 2391 — P. 86–96

[3] *Kiy K. I.* A new method of global image analysis and its application in understanding road scenes // Pattern Recognit. Image Anal. , Pleades, 2018. —P. 483–94.

[4] *Kiy K. I..* Segmentation and detection of contrast objects and their application in robot navigation // Pattern Recognit. Image Anal., Pleades,, 2015. — P. 338–46.

Использование вейвлет-нейронных сетей для решения обратных задач спектроскопии многокомпонентных растворов

Доленко Сергей Анатольевич[1]⋆ dolenko@srd.sinp.msu.ru
Ефиторов Александр Олегович[1] a.efitorov@sinp.msu.ru
Доленко Татьяна Альдефонсовна[1,2] tdolenko@mail.ru
Лаптинский Кирилл Андреевич[1,2] onelumen@gmail.com
Буриков Сергей Алексеевич[1,2] sergey.burikov@gmail.com

[1]Москва, НИИ ядерной физики имени Д.В.Скобельцына МГУ имени М.В.Ломоносова
[2]Москва, Физический факультет МГУ имени М.В.Ломоносова

Вейвлет-нейронные сети (ВНС) представляют собой семейство аппроксимационных алгоритмов, использующих для разложения вейвлет-функции. Являясь более гибкими, чем обыкновенные многослойные персептроны (МСП), они имеют более высокую вычислительную стоимость и требуют более значительных усилий по поиску оптимальных параметров.

В данной работе решаются обратные задачи по определению концентраций компонентов в многокомпонентных растворах по их спектрам комбинационного рассеяния (КР) света.

Так как ВНС весьма чувствительны к количеству входных признаков, решению рассматриваемых задач предшествовало извлечение оптимальных признаков. Среди методов извлечения признаков наилучшие результаты показало непрерывное вейвлет-преобразование.

Результаты, продемонстрированные ВНС, сравниваются с результатами, полученными с помощью МСП и с помощью линейного метода проекций на латентные структуры (ПЛС). Продемонстрированы некоторые проблемы в обеспечении эффективного обучения ВНС. Намечены пути дальнейшего улучшения алгоритма обучения.

Работа поддержана грантами РФФИ № 17-07-01479 и № 19-01-00738.

[1] *Efitorov, A. et al.* Use of Wavelet Neural Networks to Solve Inverse Problems in Spectroscopy of Multi-component Solutions // Studies in Computational Intelligence, V.856. Springer Nature, 2020. — p. 285–294. https://doi.org/10.1007/978-3-030-30425-6_33.

Use of Wavelet Neural Networks to Solve Inverse Problems in Spectroscopy of Multi-Component Solutions

Sergey Dolenko[1]⋆ dolenko@srd.sinp.msu.ru
Alexander Efitorov[1] a.efitorov@sinp.msu.ru
Tatiana Dolenko[1,2] tdolenko@mail.ru
Kirill Laptinskiy[1,2] onelumen@gmail.com
Sergey Burikov[1,2] sergey.burikov@gmail.com
[1]D.V. Skobeltsyn Institute of Nuclear Physics, M.V.Lomonosov Moscow State University
[2]Physical Department, M.V.Lomonosov Moscow State University

Wavelet neural networks (WNN) are a family of approximation algorithms that use wavelet functions to decompose the approximated function. They are more flexible than conventional multi-layer perceptrons (MLP), but they are more computationally expensive, and require more effort to find optimal parameters.

In this study, we solve the inverse problems of determination of concentrations of components in multi-component solutions by their Raman spectra.

As WNN is very sensitive to the number of input features, the solution of the studied problems was preceded with feature extraction. The best result among the feature extraction methods was demonstrated by continuous wavelet transformation.

The results demonstrated by WNN are compared to those obtained by MLP and by the linear partial least squares (PLS) method. Several problems in performing efficient WNN training have been demonstrated. Directions of possible improvement of the WNN training algorithm have been formulated.

This research was supported by RFBR, grants no.17-07-01479 and no.19-01-00738.

[1] *Efitorov, A. et al.* Use of Wavelet Neural Networks to Solve Inverse Problems in Spectroscopy of Multi-component Solutions // Studies in Computational Intelligence, V.856. Springer Nature, 2020. — p. 285–294. https://doi.org/10.1007/978-3-030-30425-6_33.

Выделение предварительно записанных голосовых сообщений в аудиозаписях телефонных разговоров

Копылов Андрей Валериевич[1] And.Kopylov@gmail.com
Середин Олег Сергеевич[1] oseredin@yandex.ru
Тышкевич Борис Владимирович[2] bvt@itoolabs.com
Филин Андрей Игоревич[1,2]⋆ afilin@itoolabs.com
[1]Тула, Тульский государственный университет
[2]Тула, ITooLabs

Call-центры генерируют огромные объемы аудиоинформации, что делает необходимым применение интеллектуальных алгоритмов для их анализа. Среди множества задач интерес представляет выделение речевых фрагментов в записях телефонных разговоров с целью дальнейшей обработки. В то же время, большинство современных call-центров используют предварительно записанные сообщения (Interactive Voice Responce, IVR) с целью автоматизации взаимодействия с клиентом (например, маршрутизация звонков, создание интерактивной очереди, так называемый "холодный"обзвон и др.). Отделение IVR от речевых фрагментов, возникающих непосредственно во время телефонного диалога играет значительную роль в задачах построения систем распознавания эмоций, фильтрации нежелательных звонков, автоматического определения наличия автоответчиков, сокращение длины хранимой записи, фильтрации спама в голосовой почте и т.д. Сложность задачи заключается в том, что в отличии от традиционного выделения речевых фрагментов (Voice Activity Detection, VAD) в потоке сообщений, в которой требуется отделить тишину, музыку и шумы, IVR, как правило, содержит речь.

В большинстве случаев человек способен определить на слух, является ли фраза предварительно записанной или произнесена в процессе живого диалога. Это дает надежду на решение такой задачи при помощи современных методов машинного обучения.

В настоящее время, на сколько нам известно, данный вопрос слабо исследован в литературе по анализу данных. Большинство описанных исследований решают близкие задачи: например, детектирование живого пользователя, и используют технологии активного взаимодействия и/или определение резких изменений характеристик канала связи.

Одним из препятствий к применению алгоритмов машинного обучения для решения данной задачи является отсутствие размеченных баз телефонных разговоров на русском языке, включающих IVR. В ходе построения системы оценки эмоционального фона диалога с оператором центра обработки вызовов [1], такая база была создана.

В данной работе проведено сравнительное исследование трех алгоритмов: на основе машины опорных векторов (SVM), градиентного бустинга (XGBoost) а также на основе сверточной нейронной сети (CNN) для решения задачи детекти-

рования IVR-фрагментов в аудиозаписи телефонного разговора. Для XGBoost и SVM использовалось признаковое представление The Geneva Minimalistic Acoustic Parameter Set (GeMAPS), для CNN - log-спектрограмма. Эксперименты показывают сопоставимые результаты работы алгоритмов с незначительным преимуществом CNN.

[1] *Копылов А. В., Середин О. С., Найденов А. В., Зенин Д. Г.* Формирование базы данных для системы оценки эмоционального фона диалога с оператором центра обработки вызовов // Математические методы распознавания образов: Тезисы докладов 18-й Всероссийской конференции с международным участием, г. Таганрог, 2017 г. М.: Торус Пресс, 2017. — С. 132–133.

Detection of Interactive Voice Responce (IVR) in audio records of phone conversations

Andrei Kopylov[1] And.Kopylov@gmail.com
Oleg Seredin[1] oseredin@yandex.ru
Boris Tyshkevich[2] bvt@itoolabs.com
Andrey Filin[1,2]⋆ afilin@itoolabs.com
[1]Tula, Tula State University
[2]Tula, ITooLabs

Call centers generate huge amounts of audio data, which makes it necessary to use intelligent data analysis algorithms. Among the many tasks interest is the allocation of speech fragments in the records of telephone conversations for postprocessing. At the same time, most modern call centers use pre-recorded messages (Interactive Voice Responce, IVR) to automate interaction with a client (for example, routing calls, creating an interactive queue, the so-called "cold" calls, etc.). Separating IVR from speech fragments that occur directly during a telephone conversation plays a significant role for developing of emotion recognition systems, filtering unwanted calls, automatic responding machines detection, reducing size of stored records, filtering spam in voice mail, etc. The complexity of the task is consist in fact that, in opposite to the traditional Voice Activity Detection (VAD), where goal is to separate silence, music and noise from speech, IVR, most commonly, contains speech.

In most cases, a person can to determine by ear whether the phrase is pre-recorded or uttered during a live dialogue. This gives hope to solve this problem using modern machine learning methods.

At present, as far as we know, this issue is poorly studied in the literature on data analysis. Most of the studies solve similar, but not exactly, problems: for example, liveness detection, and use active interaction technologies and/or determining sharp changes in the communication channel characteristics.

One of the obstacles to the use of machine learning algorithms to solve this problem is the lack of labeled database of telephone conversations, which include IVR, in Russian. Such database was created during the construction of a system for assessment the emotional state of a dialogue with a call center operator [**?**].

In this paper, we performed a comparative study of three algorithms: based on a support vector machine (SVM), gradient boosting (XGBoost) and also based on a convolutional neural network (CNN) for solving the problem of detecting IVR fragments in audio records of telephone conversations. For XGBoost and SVM was used The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) as features, for CNN was used a log spectrogram. Experiments show comparable results of the algorithms with a slight advantage of CNN.

[1] *Kopylov A. V., Seredin O. S., Naidyonov A. V., Zenin D. G.* The creation of a corpus of emotional data for the system of emotion-related states assessment of a dialogue with the call center operator // Mathematical methods of pattern recognition: Lecture notes

of 18-th Russian National Conference MMPR-2017. Russia, Taganrog, 2017 M.: Torus Press, 2017. — p. 132–133.

Метод моделирования параметров ионосферы и обнаружения ионосферных возмущений

Мандрикова Оксана Викторовна[1] oksanam1@mail.ru
Фетисова Надежда Владимировна[1]⋆ nv.glushkova@yandex.ru
Полозов Юрий Александрович[1] up_agent@mail.ru

[1]Петропавловск-Камчатский, Институт космофизических исследований и распространения радиоволн ДВО РАН

Работа направлена на создание методов моделирования и анализа параметров ионосферы. Ионосфера Земли чутко реагирует на изменения в околоземном космическом пространстве (корональные выбросы и вспышки на Солнце, изменения параметров солнечного ветра, магнитные бури и суббури). В возмущенные периоды в ионосфере возникают аномальные процессы, характеризующие возникновение ионосферных бурь. Ионосферные бури оказывают негативное влияние на работу радиоканалов и вызывают нарушения в функционировании технических средств наземного и космического базирования. Хотя механизмы возникновения ионосферных возмущений известны, их оперативное прогнозирование в настоящее время не реализовано.

В работе предложена многокомпонентная модель временного ряда параметров ионосферы, позволяющая адекватно описать регулярные вариации параметров и аномальные изменения разной интенсивности. Модель имеет вид:

$$f(t) = A^{REG}(t) + U(t) + e(t) = \sum_{\mu=1,\dots,T} \sum_{k=1,\dots,N^{\mu}_{j^{\text{reg}}}} s^{\mu}_{j^{\text{reg}},k} b^{\mu}_{j^{\text{reg}},k}(t) + \sum_{\eta,n} P^{ad}_{1,\eta}(d_{\eta,n})\Psi_{\eta,n}(t) + \sum_{\eta,n} P^{ad}_{2,\eta}(d_{\eta,n})\Psi_{\eta,n}(t) + e(t), \quad (1)$$

где $A^{REG}(t) = \sum_{\mu=1,\dots,T} \sum_{k=1,\dots,N^{\mu}_{j^{\text{reg}}}} s^{\mu}_{j^{\text{reg}},k} b^{\mu}_{j^{\text{reg}},k}(t) + e(t)$ является регулярной компонентой, которая описывает характерные изменения параметров ионосферы, $s^{\mu}_{j^{\text{reg}},k} = \sum\limits_{l=1}^{p^{\mu}_{j^{\text{reg}}}} \gamma^{\mu}_{j^{\text{reg}},l}\omega^{\mu}_{j^{\text{reg}},k-l} - \sum\limits_{n=1}^{h^{\mu}_{j^{\text{reg}}}} \theta^{\mu}_{j^{\text{reg}},n} a^{\mu}_{j^{\text{reg}},k-n}$, $p^{\mu}_{j^{\text{reg}}}$, $\theta^{\mu}_{j^{\text{reg}},n}$ – параметры μ-ой составляющей, $\omega^{\mu}_{j^{\text{reg}},k} = \nabla^{\nu^{\mu}}\delta^{\mu}_{j^{\text{reg}},k}$, $\delta^{1}_{-m^{\text{reg}},k} = c_{-m^{\text{reg}},k}$, $\delta^{\mu}_{j^{\text{reg}},k} = d_{j^{\text{reg}},k}$, $\mu = 2,\dots,T$, $a^{\mu}_{j^{\text{reg}},k}$ – остаточные ошибки модели μ-ой составляющей, $b^{1}_{-m^{\text{reg}},k} = \varphi_{-m^{\text{reg}},k}$ – масштабирующая функция, $b^{\mu}_{j^{\text{reg}},k} = \Psi_{j^{\text{reg}},k}$, $\mu = 2,\dots,T$ – вейвлет-базис.

$U(t) = \sum\limits_{\eta,n} P^{ad}_{1,\eta}(d_{\eta,n})\Psi_{\eta,n}(t) + \sum\limits_{\eta,n} P^{ad}_{2,\eta}(d_{\eta,n})\Psi_{\eta,n}(t)$ – аномальная компонента, описывающая нестационарные короткопериодные изменения параметров в возмущенные периоды, $P^{ad}_{i,\eta} = V_i * St_{\eta}$ – адаптивный порог, $d_{\eta,n} = \langle f, \Psi_{\eta,n}\rangle$, $\{\Psi_{\eta,n}\}_{\eta,n\in Z}$ – вейвлет-базис, $e(t)$ – шум.

Идентификация модели основана на комплексном подходе, объединяющем разные схемы вейвлет-преобразования с методами авторегрессии – проинтегрированного скользящего среднего. Предлагаемый подход показал свою эффективность в задачах обнаружения ионосферных возмущений. Разработанные на основе модели и предложенные в данной работе вычислительные алгоритмы, в отличие от аналогов, позволяют обнаружить внезапные аномальные изменения в ионосфере и оценить их параметры. Алгоритмы реализованы в системе оперативного анализа данных критической частоты ионосферы foF2 района Камчатки (http://lsaoperanalysis.ikir.ru/lsaoperanalysis.html).

[1] *Мандрикова О. В., Фетисова Н. В., Полозов Ю. А.* Моделирование параметров ионосферы и выделение ионосферных аномалий в режиме оперативного анализа данных // Машинное обучение и анализ данных, 2019. (в процессе)

A method for modeling of ionospheric parameters and detection of ionospheric disturbances

Oksana Mandrikova[1] oksanam1@mail.ru
Nadezhda Fetisova[1]⋆ nv.glushkova@ya.ru
Yuryi Polozov[1] up_agent@mail.ru

[1]Petropavlovsk-Kamchatskiy, Institute of Cosmophysical Research and Radio Wave Propagation FEB RAS

The paper is aimed at creating methods for modeling and analysis of the ionospheric parameters. The Earth's ionosphere is sensitive to changes in near-Earth space (coronal emissions and flashes on the Sun, changes in the solar wind parameters, magnetic storms and substorms). During disturbed periods anomalous processes occur in the ionosphere. They characterize the occurrence of ionospheric storms. Ionospheric storms have a significant impact on the radio channels and cause disturbances in the ground-based and space-based technical equipment. Although the mechanisms of the occurrence of ionospheric disturbances are known, their real-time forecasting is not currently implemented.

A multicomponent model of the ionospheric parameter time series is proposed in the paper. The model allows us to adequately describe regular variations of the parameters and anomalous changes of different intensities. The model is represented in the form:

$$f(t) = A^{REG}(t) + U(t) + e(t) = \\ = \sum_{\mu=1,\dots,T} \sum_{k=1,\dots,N^{\mu}_{j^{\mathrm{reg}}}} s^{\mu}_{j^{\mathrm{reg}},k} b^{\mu}_{j^{\mathrm{reg}},k}(t) + \sum_{\eta,n} P^{ad}_{1,\eta}(d_{\eta,n})\Psi_{\eta,n}(t) + \\ + \sum_{\eta,n} P^{ad}_{2,\eta}(d_{\eta,n})\Psi_{\eta,n}(t) + e(t), \quad (1)$$

where $A^{REG}(t) = \sum_{\mu=1,\dots,T} \sum_{k=1,\dots,N^{\mu}_{j^{\mathrm{reg}}}} s^{\mu}_{j^{\mathrm{reg}},k} b^{\mu}_{j^{\mathrm{reg}},k}(t) + e(t)$ is a regular component describing characteristic variations of the ionospheric parameters, $s^{\mu}_{j^{\mathrm{reg}},k} = \sum_{l=1}^{p^{\mu}_{j^{\mathrm{reg}}}} \gamma^{\mu}_{j^{\mathrm{reg}},l}\omega^{\mu}_{j^{\mathrm{reg}},k-l} - \sum_{n=1}^{h^{\mu}_{j^{\mathrm{reg}}}} \theta^{\mu}_{j^{\mathrm{reg}},n} a^{\mu}_{j^{\mathrm{reg}},k-n}$, $p^{\mu}_{j^{\mathrm{reg}}}$, $\theta^{\mu}_{j^{\mathrm{reg}},n}$ are parameters of μ-th component, $\omega^{\mu}_{j^{\mathrm{reg}},k} = \nabla^{\nu^{\mu}}\delta^{\mu}_{j^{\mathrm{reg}},k}$, $\delta^{1}_{-m^{\mathrm{reg}},k} = c_{-m^{\mathrm{reg}},k}$, $\delta^{\mu}_{j^{\mathrm{reg}},k} = d_{j^{\mathrm{reg}},k}$, $\mu = 2,\dots,T$, $a^{\mu}_{j^{\mathrm{reg}},k}$ are residual errors of μ-th component model, $b^{1}_{-m^{\mathrm{reg}},k} = \varphi_{-m^{\mathrm{reg}},k}$ is a scaling function, $b^{\mu}_{j^{\mathrm{reg}},k} = \Psi_{j^{\mathrm{reg}},k}$, $\mu = 2,\dots,T$ is a wavelet-basis.

$U(t) = \sum_{\eta,n} P^{ad}_{1,\eta}(d_{\eta,n})\Psi_{\eta,n}(t) + \sum_{\eta,n} P^{ad}_{2,\eta}(d_{\eta,n})\Psi_{\eta,n}(t)$ is an anomalous component describing non-stationary short-period changes of parameters during disturbed periods, $P^{ad}_{i,\eta} = V_i * St_{\eta}$ is an adaptive threshold, $d_{\eta,n} = \langle f, \Psi_{\eta,n}\rangle$, $\{\Psi_{\eta,n}\}_{\eta,n\in Z}$ is a wavelet-basis, $e(t)$ is a noise component.

Model identification is based on an integrated approach that combines different wavelet transform schemes with the ARIMA methods (autoregressive - integrated moving average). The proposed approach has shown efficiency in tasks of detecting ionospheric disturbances. Computational algorithms developed on the basis of the model are proposed in this paper. The algorithms in comparison to analogs make it possible to detect sudden anomalous changes in the ionosphere and estimate their parameters. The algorithms are implemented in a system for the online analysis of the ionospheric critical frequency data (foF2) of the Kamchatka region (http://lsaoperanalysis.ikir.ru/lsaoperanalysis.html).

[1] *Mandrikova O, Fetisova N., Polozov Yu.* A method for modeling of ionospheric parameters and detection of ionospheric disturbances // Machine Learning and Data Analysis, 2019. (in proccess)

Исследование рядов динамики метеорологических показателей

Зюзина Нина Александровна[1]⋆ zjuzina.na15@physics.msu.ru
Газарян Варвара Арамовна[1,2] varvaragazaryan@yandex.ru
Курбатова Юлия Александровна[4] kurbatova.j@gmail.com
Шапкина Наталья Евгеньевна[1,5] neshapkina@mail.ru
Чуличков Алексей Иванович[1,4] achulichkov@gmail.com

[1]Москва, Московский государственный университет им. М. В. Ломоносова
[2]Москва, Финансовый университет при правительстве РФ
[3]Москва, Институт проблем экологии и эволюции им. А. Н. Северцова РАН
[4]Москва, ВИНИТИ РАН
[5]Москва, ИТПЭ РАН

Введение. Анализ рядов динамики метеорологических параметров является важным аспектом в изучении современных изменений глобального климата. Данные метеорологической станции, упорядоченная совокупность значений переменных, измеряемых через постоянный временной промежуток, представляются в виде ряда динамики метеорологического параметра.

Временные ряды динамики метеорологических параметров, как правило, нестационарны, поэтому для их исследования удобно вейвлет-преобразование, которое позволяет устранить недостатки преобразования Фурье, которое локализовано по частоте, но не имеет временного разрешения.

Вейвлеты — это семейство функций, которые получаются из одной функции посредством ее сдвигов и растяжений по оси времени. С помощью вейвлет-преобразования функция рассматривается в виде разложения на колебания, локализованные по времени и частоте, что удобно для нестационарного временного ряда.

Предмет исследования. В качестве метеорологических параметров в данной работе используются показатели среднесуточной температуры на протяжении 45 лет (1971-2016гг) на юго-западе Валдайской возвышенности (метеорологическая станция «Заповедник»), и показатели посекундной концентрации углекислого газа за 6 лет (конец 2011-2018гг), во Вьетнаме (метеорологическая станция «AsiaFlux»). Статистические данные предоставлены ИПЭЭ им. А.Н. Северцова РАН.

В качестве исследуемых рядов динамики были выбраны: (1) показатели среднесуточной температуры, усредненные по неделям; (2) показатели среднесуточной температуры, относящиеся к метеорологической весне; (3)показатели (максимумы и минимумы) дневной концентрации CO_2 на разных высотах (0.3 м. и 46 м.); (4)показатели концентрации CO_2 за полтора года, усредненные с периодом в 30 минут.

Исследуемые временные ряды динамики были проверены на стационарность с помощью расширенного теста Дикки–Фуллера. Результат оказался предска-

зуем — ряды 1, 2, 4 нестационарны, тем самым обосновано применение вейвлет-анализа. Временные ряды 3, напротив, оказались стационарными, что, вообще говоря, не исключает возможность использования вейвлет-преобразования в качестве одной из методик исследования.

Непрерывное вейвлет-преобразование. Сущность непрерывного вейвлет-преобразования заключается в следующем. С помощью подходящего материнского вейвлета $\psi(t)$, вычисляются вейвлет-функции $\psi_{a,b}(t)$:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right).$$

Параметр a называется параметром масштаба вейвлет-преобразования, он принимает строго положительные значения и отвечает за ширину вейвлета; величина b есть параметр сдвига, который определяет положение вейвлета на оси t.

Далее определяются вейвлет-коэффициенты $W(a,b)$:

$$W(a,b) = \frac{1}{\sqrt{a}}\int_{-\infty}^{+\infty} s(t)\psi^*_{a,b}dt,$$

где $s\,(t)$ – анализируемый ряд динамики, * обозначает комплексное сопряжение.

После этого проводится качественный анализ картины вейвлет-коэффициентов и построение интегрального спектра:

$$S\,(a) = \int_{a_1}^{a_2} |W\,(a,b)|^2 db,$$

который показывает наличие циклов в исходном временном ряде. Масштаб $\boldsymbol{a}$ связан с координатами оси времени $\boldsymbol{t}$ как: $t = \frac{a}{F_c}$, где F_c — центральная частота вейвлета.

После обработки исследуемых рядов в качестве материнских вейвлетов были выбраны вейвлет Морле и вейвлет Гаусса 7, дающие наиболее информативные и наглядные результаты.

Заключение. В рядах динамики показателей температуры (ряды 1,2) были определены периоды всех цикличностей. Было произведено сравнение с результатами, полученными ранее с использованием других методов или полученных с помощью данных в других регионах. Результаты хорошо согласуются между собой. Были получены качественные картины временных положений цикличностей, которые, как показывают исследования, могут являться суперпозициями более сложных явлений.

В ходе анализа данных показателей концентрации CO_2 (ряды 3,4) была выявлена сложная структура спектров, для всех исследуемых рядов сложно выделить высокочастотные периодичности, так как спектр содержит множество

непродолжительных по времени слабых цикличностей. Интегральные спектры размыты, были выделены в основном только слабые пики, произведено сравнение результатов с методом линейной регрессионной модели и Фурье-анализом. Сложная структура высокочастотной периодичности может быть вызвана такими природными явлениями, как особенности климата Вьетнама, эмиссия CO_2 из почвы, перемешивание слоев воздуха и др.

Работа поддержана грантами РФФИ № 17-07 -00832 А, 18-07-00424 А, РФФИ-РГО-а № 17-05-41127.

[1] *Газарян В. А. , Курбатова Ю. А. , Овсянников Т. А. , Шапкина Н. Е.* Статистический анализ циклических изменений в рядах динамики метеорологических показателей на юго-западе Валдайской возвышенности.// ВМУ. Серия 3. ФИЗИКА. АСТРОНОМИЯ. 2018. № 1.

[2] *Астафьева Н.М.* Вейвлет-анализ: основы теории и примеры применения. // Успехи физических наук, 1996, т. 166, №11, с. 1145-1170.

[3] *Тимохина А.В., Прокрушин А.С., Онучкин А.А., Панов А.В., Кофман Г.Б., Хаймани М.* Динамика приземной концентрации CO2 в среднетаежной подзоне Приенисейской Сибири. // Экология. 2015. Т. 46. № 2. С. 143-151.

Study of series of dynamics of meteorological parameters by wavelet analysis

Nina Ziuzina[1]⋆ zjuzina.na15@physics.msu.ru
Varvara Gazaryan[1,2] varvaragazaryan@yandex.ru
Yulia Kurbatova[4] kurbatova.j@gmail.com
Natalya Shapkina[1,5] neshapkina@mail.ru
Alexey Chulichkov[1,4] achulichkov@gmail.com

[1]Moscow, Lomonosov Moscow State University
[2]Moscow, Financial University under the Government of the Russian Federation
[3]Moscow, A.N. Severtsov Institute of Ecology and Evolution
[4]Moscow, VINITI RAS
[5]Moscow, ITAE RAS

Introduction. The analysis of the series of dynamics (time series) of meteorological parameters is important to study modern changes in the global climate. The data from the meteorological station (an ordered set of values of variables measured over a constant time period) are presented in the form of a series of dynamics of the meteorological parameter.

The time series of meteorological parameters are usually non-stationary, therefore, to study them, the wavelet transform is convenient, eliminating the disadvantages of the Fourier transform, because having a good localization in frequency, a conversion has no time resolution.

Wavelets are a family of functions that are obtained from one function through its shifts and stretches along the time axis. Using a wavelet transform, a function is considered as a decomposition of oscillations localized both in time and frequency. It is the reason for using the wavelet transform instead of the Fourier transform, in case of studying an unsteady time series.

Subject of study. As the meteorological parameters in this work, we use the average daily temperature for 45 years (1971-2016) in the south-west of the Valdai Upland (meteorological station "Zapovednik"), and the parameters of the second-minute carbon dioxide concentration for 6 years (end of 2011-2018), in Vietnam (meteorological station "AsiaFlux"). Statistic data was provided by A.N. Severtsov Institute of Ecology and Evolution.

As the studied series of dynamics the following parameters were selected: (1)The average daily temperature, averaged by week; (2) The average daily temperature related to the meteorological spring; (3)Data (maxima and minima) of daily CO_2 concentration at different altitudes (0.3 m and 46 m); (4) CO_2 concentration data for a year and a half averaged over a period of 30 minutes.

The studied time series were checked for stationarity using the Augmented Dickey-Fuller test. The result turned out to be predictable - series 1, 2, 4 are non-stationary, thereby the use of wavelet analysis is justified. Time series 3, on the

contrary, turned out to be stationary, which, generally speaking, does not exclude the possibility of using the wavelet transform as one of the research methods.

Continuous Wavelet Transform. The essence of continuous wavelet transform is as follows. Using a suitable mother wavelet $\psi(t)$, wavelet functions $\psi_{a,b}(t)$ are computed:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right).$$

The parameter a is called the scale parameter of the wavelet transform; it takes strictly positive values and is responsible for the width of the wavelet. The parameter b is the shift parameter that determines the position of the wavelet on the t axis.

Next, the wavelet coefficients $W(a,b)$ are determined::

$$W(a,b) = \frac{1}{\sqrt{a}}\int_{-\infty}^{+\infty} s(t)\psi^*_{a,b}dt,$$

where $s\,(t)$ – analyzed time series, * means complex pairing.

After that, a qualitative analysis of the picture of wavelet coefficients and the construction of the integrated spectrum are carried out:

$$S\,(a) = \int_{a_1}^{a_2} |W\,(a,b)|^2 db,$$

that shows the existence of cycles in the original time series. Scale a is related to the coordinates of the time axis $\boldsymbol{t}$ as:

$$t = \frac{a}{F_c},$$

where F_c — center wavelet frequency.

After processing the time series under study, the Morlet wavelet and the Gaussian wavelet 7 were selected as the mother wavelets, giving the most informative and visual results.

Conclusion. In the series of dynamics of temperature indicators (series 1, 2), periods of all cycles were determined. A comparison was made with the results obtained previously using other methods or using data in other regions. The results are in good agreement with each other. Qualitative pictures were obtained of the temporal positions of cyclicities, which, as studies show, can be superpositions of more complex phenomena.

A compound structure of the spectra was revealed during the analysis of these parameters of CO_2 concentration (series 3, 4); for all the studied series of carbon dioxide dynamics it is difficult to distinguish high-frequency periodicities, since the spectrum contains many short-lived weak cycles. The integrated spectra of all obtained time series are blurred, mainly only weak peaks were identified, and the results were compared with the linear regression model and Fourier analysis. The

compound structure of high-frequency periodicity can be caused by such natural phenomena as the climate of Vietnam, CO_2 emissions from the soil, mixing of layers of air, etc.

Work supported by grants RFBR No 17-07 -00832 A, 18-07-00424 A, RFBR -RGS-a No17-05-41127.

[1] *V. A. Gazaryan, J. A. Kurbatova, T. A. Ovsyannikov, N. E. Shapkina.* 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. 73. No 1.

[2] *Astafyeva N. M.* 'Wavelet analysis: basic theory and some applications" Phys. Usp. 1996, b. 166, No 11, pp. 1145-1170.

[3] *A.V. Timokhina, A.S. Prokushkin, A.A. Onuchin, A.V. Panov, G.B. Kofman, M. Heimann* 2015, published in Ekologiya, 2015, No. 2, pp. 110–119.

Анализ состава инвестиционного портфеля по данным о доходностях ценных бумаг в условиях нестационарного фондового рынка

Красоткина Ольга Вячеславовна[1] okrasotkina@markovprocesses.com
Марков Михаил[1] mmarkov@markovprocesses.com
Моттль Вадим Вячеславович[2]⋆ vmottl@yandex.ru
Пугач Илья Александрович[3] iliapugach@gmail.com
[1]Summit, NJ, USA, Markov Processes International
[2]Москва, Вычислительный центр РАН
[3]Московский физико-технический институт

Задача восстановления скрытого состава инвестиционного портфеля по известным временным рядам стоимостной доходности самого исследуемого портфеля и доходностей ценных бумаг (биржевых активов), в которые предположительно мог быть вложен капитал, сформулирована лауреатом нобелевской премии по экономике 1990 года Уильямом Шарпом под названием Returns Based Style Analysis – RBSA [1, 2]. Предполагается, что портфель построен по принципу Buy and Hold, т.е. количество ценных бумаг каждого вида не изменяется в течение периода наблюдения. Математическая модель Шарпа связывает вектор доходностей ценных бумаг с доходностью портфеля регрессионной зависимостью, в которой роль векторного коэффициента играет постоянное неизвестное долевое стоимостное распределение капитала, которое надо найти как решение задачи линейной регрессии.

Однако даже при постоянном количественном составе портфеля Buy and Hold его стоимостной долевой состав зависит от теущих цен биржевых активов и не будет постоянным. В силу этого обстоятельства модель Шарпа является неточной, тем не менее, она приближенно характеризует скрытый количественный состав портфеля в условиях стационарного рынка ценных бумаг, когда цены активов колеблются вокруг некоторых постоянных значений.

Последнее условие сответствует предположению, что доходности активов, выражающие "скорости" роста цен, имеют в среднем нулевые значения. Если на очень коротких интервалах наблюдения это грубое предположение еще можно приближенно принять, то при более длительном наблюдении оно совершенно неадекватно реальности. Более того, именно факт наличия положительной составляющей волатильных доходностей активов лежит в основе самой идеи биржевых инвестиций как способа сохранения и умножения капитала в условиях нормальной экономики.

В данной работе мы рассматриваем усложнение исходной регрессионной модели Шарпа, учитывающее зависимость долевого стоимостного распределения капитала в портфеле Buy and Hold от изменяющихся доходностей составляющих его биржевых активов и, следовательно, от времени. Хотя такая модель неизбежно является нелинейной, она относится к классу обобщенных линейных

моделей [3]. Этот термин означает, что в такой модели нелинейность полностью сосредоточена в выборе функции связи, соединяющей целевую переменную, имеющую смысл предсказываемой доходности портфеля, с обычной скалярной линейной функцией от вектора доходностей биржевых активов. Неизвестный векторный коэффициент этой функции и есть искомое долевое стоимостное распределение капитала.

Обобщенная регрессионая задача RBSA сформулирована как задача минимизации остаточной суммы квадратов разности между известными и предсказанными значениями доходности портфеля в интервале наблюдения, однако предсказанная доходность портфеля теперь уже не является линейной функцией доходностей активов, в отличие от модели Шарпа. Как следствие, критерий наименьших не обязательно будет выпуклым, но в множестве экспериментов мы никогда не наблюдали его невыпуклости.

Построен алгоритм решения обобщенной регрессионой задачи RBSA в двойственной форме, имеющий полиномиальную вычислительную сложность относительно длины временного ряда наблюдения доходностей и линейную вычислительную сложность по числу биржевых активов.

Работа поддержана грантом РФФИ №17-07-00993.

[1] *W.F. Sharpe*. Determining a fund's effective asset mix. Investment Management Review, November/December 1988, pp. 59-69.

[2] *W.F. Sharpe* Asset allocation: Management style and performance measurement. The Journal of Portfolio Management, Winter 1992, 18 (2), pp. 7-19.

[3] *P. McCullagh, J. Nelder.* Generalized Linear Models, Second Edition. Chapman and Hall, 1989, 511 p.

Returns-based analysis of investment portfolios accounting for time-varying asset prices

Olga Krasotkina[1] okrasotkina@markovprocesses.com
Michael Markov[1] mmarkov@markovprocesses.com
Vadim Mottl[2]⋆ vmottl@yandex.ru
Ilya Pugach[3] iliapugach@gmail.com
[1]Summit, NJ, USA, Markov Processes International
[2]Moscow, Computing Center of the Russian Academy of Sciences
[2]Moscow Institute of Physics and Technology

The problem of estimating the asset composition of an investment portfolio from readily available return time series of both the portfolio and relevant investment assets was formulated by 1990 Nobel prize winner William Sharpe and became known under the name of RBSA – Returns Based Style Analysis [1]. The model became quite popular and widely used in Finance, as it allows an investment professional to explain the behavior of a very large and complex portfolio, consisting of both large and small number of securities, by a handful of asset indices. In its original formulation RBSA assumes that asset weights are constant within the estimation time window, whether such window represents the entire date range or a "rolling" window of a smaller size [2]. Sharpe's mathematical model ties the portfolio return to the returns of the assets via a constrained linear regression model, in which vector of regression coefficients represents asset weights that have to be estimated. In practical terms, RBSA is finding a portfolio with constant allocations to asset indices or factors (regressors) that closely approximates the portfolio's returns.

However, very few investment strategies are designed to continuously "rebalance" portfolios, e.g., maintain constant asset weights. First, this could prove to be quite expensive as such rebalancing requires significant trading, but it also creates a significant tax burden for a taxable portfolio among other issues. In many cases of so-called "passive" strategies portfolio managers allow their investments appreciate/depreciate with little or no trading, especially on the asset class or sector level. Such portfolio is typically called Buy-and-Hold and its allocations are changing over time but are driven primarily by market prices.

Sharpe's RBSA model with its constant allocations is inaccurate for analysis of a buy-and-hold portfolio, unless asset volatilities and drifts are either identical or very similar. Such assumption is atypical, as investments behave quite differently between countries and regions, economic sectors and asset classes.

In this work, we consider a significant enhancement of Sharpe's original regression model by way of allowing time-varying asset portfolio weights as in a Buy-and-Hold portfolio. Though such a model is inevitably nonlinear, it belongs to the class of Generalized Linear Models [3]. This term means that the nonlinearity is fully concentrated in the choice of the link function, which ties the goal variable, namely, the portfolio return to be predicted, to the usual scalar linear function of the asset

returns vector, whose regression coefficient vector is the sought-for cost composition of the prtfolio.

The generalized RBSA regression problem is formulated as that of minimizing the residual sum of squared differences between the reported and the predicted values of the portfolio returns in the observation interval, however the predicted returns are no longer linear functions of the asset returns, as it was the case in Sharpe's model. As a result the criterion is not obligatory convex, but we never observed its non-convexity in a variety of experiments.

The algorithm of solving the generalized RBSA regression problem in its dual form has the polynomial computational complexity with respect to the length of the observed returns time series and the linear computational complexity in the number of stock-market assets.

This research is funded by RFBR, grant No17-07-00993.

[1] *W.F. Sharpe.* Determining a fund's effective asset mix. Investment Management Review, November/December 1988, pp. 59-69.

[2] *W.F. Sharpe* Asset allocation: Management style and performance measurement. The Journal of Portfolio Management, Winter 1992, 18 (2), pp. 7-19.

[3] *P. McCullagh, J. Nelder.* Generalized Linear Models, Second Edition. Chapman and Hall, 1989, 511 p.

Анализ временных рядов в задаче распознавания видов физической активности человека

Мотренко Анастасия Петровна[1]⋆ motrenko@forecsys.ru
Симчук Егор Александрович[1] simchuk_egor@forecsys.ru
Стрижов Вадим Викторович[1] strijov@gmail.com
Каширин Даниил Олегович[1] kashirin@forecsys.ru
Инякин Андрей Сергеевич[1] andrey.inyakin@frccsc.ru
Хайруллин Ринат Ильдарович[1] khayrullin@forecsys.ru

[1]Москва, ООО «Форексис»

Решается задача анализа временных рядов для распознавания видов физической активности человека. Под активностью понимается продолжительная деятельность, для определения которой требуется понимание контекста. Различные виды активности могут включать одни и те же элементарные действия. Разработана методика решения задачи распознавания вида активности, включающая рекомендации по сбору данных и проведению экспериментов, разметке данных для конструирования и обучения математических моделей распознавания.

Задача распознавания вида физической активности решается в иерархической постановке: на нижнем уровне иерархии решается задача классификации микродействий или элементарных действий, результаты распознавания используются для генерации расширенного признакового описания в задаче верхнего уровня классификации.

Экспертный выбор элементарных действий для постановки задачи нижнего уровня иерархии обладает рядом ограничений. Это трудоемкий процесс, который не всегда позволяет задать достаточное количество классов для решения задачи распознавания видов физической активности. Кроме того, выбранный набор действий, как правило не обобщается на другие типы деятельности. В связи с этим предложен метод генерации признаков на базе алгоритма автоматического обнаружения эталонов из работы [1]. Эталонами называются сегменты временного ряда, характерные для данной активности, и не характерные для других распознаваемых видов активности. При поиске эталонов для проекций временного ряда трех-осевого датчика на внутреннюю систему координат датчика возникает проблема, связанная с ориентацией датчика в пространстве – одно и то же действие в разных положениях имеет различные проекции. Предложена модификация метода поиска эталонов с локальным проектированием временных рядов на базис главных компонент.

Ошибки в разметке временных рядов, выполненной асессорами и связанные с человеческим фактором или недостаточной формализацией классов разметки, приводят, во-первых, к ухудшению качества модели, во-вторых, к невозможности объективной оценку качества моделей распознавания и классификации. Предложены методы обучения на «зашумленной» и/или неточной разметке,

учитывающие наличие шума и позволяющие повысить качество распознавания по сравнению с «доверчивым» обучением на результатах асессорской разметки. Возможность обрабатывать неточную разметку позволяет сократить время и стоимость разметки, и, как следствие, увеличить количество размеченных данных для обучения.

[1] *B. Hu, Y. Chen, E. Keogh* Time Series Classification under More Realistic Assumptions // Proceedings of the 2013 SIAM International Conference on Data Mining, 2013, 578 – 586 *Motrenko A., Strijov V.* Extracting fundamental periods to segment biomedical signals // Journal of Biomedical and Health Informatics, 2016, Vol. 20, No 6, 1466 – 1476.

Time series analysis for continuous human physical activity recognition

Anastasia Motrenko[1]⋆ motrenko@forecsys.ru
Egor Simchuk[1] simchuk_egor@forecsys.ru
Vadim Strijov[1] strijov@gmail.com
Danil Kashirin[1] kashirin@forecsys.ru
Andrey Inyakin[1] andrey.inyakin@frccsc.ru
Rinat Khayrulin[1] khayrullin@forecsys.ru
[1]Moscow, Forecsys

The talk addressees a problem of time series analysis for continuous human activity recognition. We define "activity" as a long-term complex of movements, unified by a common purpose. Different activities might include the same movements. We consider movements as elementary units of recognition. We describe a methodology for activity recognition, which includes recommendations for data acquisition and experimental design, data labeling and learning classification models.

We adopt hierarchical approach to the problem of activity recognition. At the bottom level of the hierarchy the problem movement classification is solved. Results of movement classification are further used to generate features for top-level problem: activity recognition.

Expert-based selection of movements to classify suffers from a number of limitations. While the process is time- and labor-consuming, it does not always results in selecting enough classes for adequate representation of the activity at question. Moreover, the results rarely generalize to other sets of activities. To overcome these issues we propose a method of feature generation for activity recognition, based on automatic pattern detection [1]. This method detects patterns which are shape-specific to a given activity and are rarely found within other activities. Results of pattern recognition for separate projections of a tri-axial sensor time series are sensitive to sensor orientation. The same movement, performed in different positions, results in different signal shapes for a given projection. We propose a modification of the pattern detection procedure, which improves position invariance by projecting time series onto local principal components basis.

Errors in human data labeling deteriorate model performance and impede model evaluation. We propose noise-aware learning methods. Compared to "naive" learning, the proposed method demonstrates better classification quality. Softer requirements to labeling precision allow to reduce labeling cost and time and increase the number of labeled samples.

[1] *B. Hu, Y. Chen, E. Keogh* Time Series Classification under More Realistic Assumptions // Proceedings of the 2013 SIAM International Conference on Data Mining, 2013, 578 – 586 *Motrenko A., Strijov V.* Extracting fundamental periods to segment biomedical signals // Journal of Biomedical and Health Informatics, 2016, Vol. 20, No 6, 1466 – 1476.

Локально-аппроскимирующие модели в задаче декодирования сигналов головного мозга

Маркин Валерий Олегович[1]⋆ markin.vo@phystech.edu
Исаченко Роман Владимирович[1]⋆ isa-ro@yandex.ru
Стрижов Вадим Викторович[1] strijov@ccas.ru
[1]Москва, Московский физико-технический институт

В работе рассматривается задача построения оптимального признакового описания в задаче декодирования сигналов электрокортикографии (ECoG). Сигнал представляет собой многомерный временной ряд, значения которого – напряжение на каждом из 32 электродов в различные моменты времени. По исходным данным необходимо предсказать траекторию движения руки испытуемого в пространстве. Для борьбы с избыточностью пространства признаков и неустойчивостью модели предлагается построить локальную модель аппроксимации сигнала. Это позволяет существенно снизить размерность признакового пространства и учесть пространственную структуру сигнала.

В качестве основной локальной модели в работе используется аппроксимация сигнала двумерным нормальным распределением. В качестве признаков вместо сигнала с каждого электрода используется информация о центре и разбросе совокупного многомерного сигнала всех электродов. Использование данного метода позволило учесть пространственные зависимости в данных, решить проблему сильной корреляции сигнала на близко расположенных электродах и, помимо этого, привело к снижению размерности признакового описания в три раза.

В работе проведены эксперименты на реальных данных, полученных в экспериментах на обезьянах. Проведено сравнение предложенного метода с методом частных наименьших квадратов (PLS). В ходе эксперимента было установлено, что предложенный метод дает лучший результат, чем метод PLS, а также менее склонен к переобучению.

Работа выполнена при поддержке РФФИ (проекты 19-07-1155, 19-07-0875).

[1] *Isachenko R.V., Strijov V.V.* Quadratic Programming Optimization with Feature Selection for Non-linear Models // Lobachevskii Journal of Mathematics, 2018, 39(9) : 1179-1187.

Local-approximating models in brain signals decoding

Markin Valerii[1]⋆ markin.vo@phystech.edu
Isachenko Roman[1] isa-ro@yandex.ru
Strijov Vadim[1] strijov@ccas.ru
[1] Moscow, MIPT

The paper studies the problem of building an optimal feature description in the task of signal decoding. Electrical signals in the cerebral cortex recorded by electrocortical imaging (ECoG) are considered. The signal is a multivariate time series, the values of which are the voltage at each of the 32 electrodes at different times. Given the initial data, it is necessary to predict the trajectory of the patient's arm movement in space. The initial feature space is excessive and the forecasting model is unstable. To solve this problem it is proposed to build a local model of signal approximation. This allows to significantly reduce the dimensionality of the feature space and take into account the spatial structure of the signal.

In the current work approximation of a signal by two-dimensional normal distribution is used as a basic local model. Instead of the signal from each electrode, it uses information about the mean and the variance of the combined multivariate signal of all electrodes. The use of this method allowed to take into account the spatial dependencies in the data, to solve the problem of strong correlation between the signals from close electrodes, and, in addition, led to a three-fold decrease in the dimension of the feature space.

Experiments were carried out on real data obtained in experiments on monkeys. The proposed method is compared with the method of partial least squares (PLS). In the course of the experiment it was established that the proposed method gives a better result than the PLS method, and is less inclined to retraining.

The research is founded by RFBR (projects 19-07-1155, 19-07-0875)

[1] *Isachenko R. V., Strijov V. V.* Quadratic Programming Optimization with Feature Selection for Non-linear Models // Lobachevskii Journal of Mathematics, 2018, 39(9) : 1179-1187.

Улучшение визуального качества изображений в авиационных системах улучшенного видения с использованием генеративных состязательных сетей

Визильтер Юрий Валентинович[1] viz@gosniias.ru
Выголов Олег Вячеславович[1] o.vygolov@gosniias.ru
Доброходов Константин Викторович[1]⋆ konstantin.dobrokhodov@gmail.com
Комаров Денис Валерьевич[1] dkomarov@gosniias.ru
Лебедев Максим Алексеевич[1] mlebedev@gosniias.ru

[1]Москва, ФГУП «Государственный научно-исследовательский институт авиационных систем»

Информационная поддержка экипажа воздушного судна (ВС) в плохих условиях видимости в режимах захода на посадку, посадки и руления является важной задачей обеспечения безопасности полетов. Одной из актуальных систем, решающей задачу повышения ситуационной информированности летчика, является система улучшенного видения (СУВ). В таких системах данные от датчиков технического зрения (как правильно, телевизионных (ТВ) и инфракрасных (ИК) камер) проходят обработку специальными алгоритмами, повышая информативность и улучшая восприятие летчиками видеоинформации.

В данной работе предлагается оригинальная архитектура генеративной состязательной нейронной сети, основанной на архитектуре pix2pix, позволяющей повышать визуальное качество изображений за счет устранения шума и размытости, а также повышения резкости изображения.

Работа выполнена при поддержке РФФИ, грант 18-07-01275А, и РНФ, грант № 19-11-11008.

[1] *Визильтер Ю. В., Выголов О. В., Доброходов К. В., Комаров Д. В., Лебедев М. А.* Улучшение визуального качества изображений в авиационных системах улучшенного видения с использованием генеративных состязательных сетей // Вестник компьютерных и информационных технологий, Москва: ООО «Издательский до «Спектр», 2019. — (принято в печать).

Image Enhancement in Aviation Enhanced Vision Systems Using Generative Adversarial Networks

Vizilter Yuriy[1] viz@gosniias.ru
Vygolov Oleg[1] o.vygolov@gosniias.ru
Dobrokhodov Konstantin[1]⋆ konstantin.dobrokhodov@gmail.com
Komarov Denis[1] dkomarov@gosniias.ru
Lebedev Maksim[1] mlebedev@gosniias.ru

[1]Moscow, The Federal State Unitary Enterprise "State Research Institute of Aviation Systems"(FGUP "GosNIIAS")

Informational support for the aircraft crew during approach, landing and taxiing in poor visibility conditions is an important task in order to ensure flight safety. One of the relevant systems that solve the problem of increasing situational awareness of the pilot is the enhanced vision system (EVS). In such systems, data from vision sensors (generally, television (TV) and infrared (IR) cameras) are processed by special algorithms, increasing information content and improving the perception of video information by pilots.

This paper proposed the original architecture of a generative adversarial neural network based on the pix2pix architecture, which allows improving the visual quality of images by eliminating noise and blur, as well as increasing image sharpness.

This work was performed with the support of RFBR, grant 18-07-01275A, and RSF, grant 19-11-11008.

[1] *Vizilter Yu., Vygolov O., Dobrokhodov K., Komarov D., Lebedev M.* Image Enhancement in Aviation Enhanced Vision Systems Using Generative Adversarial Networks // Herald of computer and information technologies, Moscow: Publishing house "Spektr", 2019. — (in printing).

Алгоритм получения топологических признаков цифровых изображений на основе компьютерной топологии

Еремеев Сергей Владимирович[1] sv-eremeev@yandex.ru
Романов Семен Алексеевич[1,2]⋆ cwwc@bk.ru
[1]Владимир, Владимирский Государственный Университет
[2]Ульяновск, SimbirSoft

Высокий технологический уровень изделий устанавливает перед изготовителями исходных материалов соответствующие требования по качеству выпускаемой продукции. Таким образом, цена человеческой ошибки при контроле качества материала значительно возрастает за счет потерянных средств при изготовлении технически сложной продукции. Поэтому человеческий контроль должен быть либо полностью заменен автоматическим, либо дополнен системами компьютерного зрения. В свою очередь, перед системами компьютерного зрения ставится непростая задача по обнаружению и распознаванию дефектов. Основная сложность обусловлена особенностями области производства, будь то высокие температуры, загрязненность продукции или неидеальные условия освещенности. Для того чтобы максимально нивелировать влияние большинства факторов предлагается сегментировать изображения анализируемого продукта методами компьютерной топологии [1, 2]. Такой подход будет объединять приближенные по яркости пиксели в единую одноцветную группу, что позволит фильтровать шум, обусловленный тенями и особенностями структуры материала.

(а) (б)

Рис. 1. Изображение металла с инородными вкраплениями (а) до обработки, (б) после обработки.

Методы вычислительной топологии подразумевают представление всей структуры изображения в виде устойчивых образований, называемых дырами. Под дырами подразумевается объединения групп прилегающих пикселей по признаку приближенности их цветового значения. Предполагается, что по мере увеличения порога, при котором пиксели будут объединяться в группы, будут разрастаться и дыры, тем самым упрощая структуру всего изображения. Это позволит представить ранее зашумленную тенями и маловыраженными свой-

ствами материала поверхность единой группой с усредненным цветом. На более поздних этапах отдельные группы пикселей начинают объединяться, образуя тем самым древовидную структуру, представляющую собой иерархию наследования дыр. Корнем этого дерева будет являться представление всей поверхности в виде одного единого цвета, а наследники второго и последующих порядков — наиболее выраженные отличия материала. Данные отличия и предлагается характеризовать в качестве выявленных дефектов и в дальнейшем использовать сегментированное изображение для последующего анализа. Применив алгоритм сегментации к изображению металла с дефектом рис. 1а мы получим его сегментированную версию, где четко прослеживаются границы дефекта и границы тени рис. 1б.

Работа поддержана грантом 14281GU/2019.

[1] *Eremeev S.*, *Kuptsov K.*, *Romanov S.* An Approach to Establishing the Correspondence of Spatial Objects on Heterogeneous Maps Based on Methods of Computational Topology // In: van der Aalst W. et al. (eds) Analysis of Images, Social Networks and Texts. AIST 2017. Lecture Notes in Computer Science, vol 10716. Springer, pp. 172–182.

[2] *Edelsbrunner H.* Computational Topology: An Introduction. American Mathematical Society, 2009.

Algorithm for obtaining features of digital images based on computer topology

Eremeev Sergey[1] sv-eremeev@yandex.ru
Romanov Semyon[1,2]⋆ cwwc@bk.ru
[1]Vladimir, Vladimir State University
[2]Ulianovsk, SimbirSoft

A high technological level of production sets the appropriate requirements for the quality of products for manufacturers of raw materials. This means that the cost of human error increases significantly in material quality control due to the loss of money in the manufacture of technically complex products. Therefore, human control must be either fully replaced by automatic control or complemented by computer vision systems. Nowadays there is a complex problem for detecting and recognizing defects with the help of computer vision systems. The main difficulty is to take into account the peculiarities of the area of production such as high temperatures, product contamination or not ideal lighting conditions.

In order to minimize the influence of most factors, it is proposed to segment the images of the analyzed product by methods of computer topology [1, 2]. Such an approach will combine the pixels approximate in brightness into a single color group which will allow us to filter the noise formed by shadows and peculiarities of the material structure.

(a) (b)

Fig. 1. Image of metal with foreign inclusions (a) before processing, (b) after processing.

Computational topology methods represent the entire structure of the image in the form of stable formations called holes. Holes are the unions of groups of adjacent pixels based on the approximation of their color value. Pixels will be merged into groups with the increase of some threshold. Holes will grow and the structure of the entire image will become easier. It will allow us to present a surface with noisy shadows and low material properties as a single group with an average color. At later stages, individual groups of pixels begin to unite, thus forming a tree-like structure that represents a hierarchy of hole inheritance. The root of this tree will be the representation of the entire surface as a single color. The heirs of the second and subsequent orders are the most pronounced differences in material. It is proposed to

characterize these differences as defects and to use the segmented image for further analysis. After applying the segmentation algorithm to the image of metal with a defect in Fig. 11a, we obtain its segmented version. Fig. 1b shows the boundaries of the defect and the borders of the shadow.

The reported study was funded by Foundation for Assistance to Innovations according to the research project 14281GU/2019.

[1] *Eremeev S.*, *Kuptsov K.*, *Romanov S.* An Approach to Establishing the Correspondence of Spatial Objects on Heterogeneous Maps Based on Methods of Computational Topology // In: van der Aalst W. et al. (eds) Analysis of Images, Social Networks and Texts. AIST 2017. Lecture Notes in Computer Science, vol 10716. Springer, pp. 172–182.

[2] *Edelsbrunner H.* Computational Topology: An Introduction. American Mathematical Society, 2009.

Исследование сокращения скелетного описания для задачи детектирования падений

Середин Олег Сергеевич[1]⋆ oseredin@yandex.ru
Копылов Андрей Валериевич[1] And.Kopylov@gmail.com
Сурков Егор Эдуардович[1] eg-su@mail.ru
[1]Тула, Тульский государственный университет

В задачах обработки полученных данных с камер и сенсоров возникает вопрос безопасности и как следствие – анализа этих данных непосредственно на облачных ресурсах. Периферийные вычисления позволяют выполнить основную часть работы на отдалённом сервере. Для этого нужно подготовить обезличенные и в тоже время, учитывая ограниченную пропускную способность между датчиками и облачными ресурсами, минимально возможные по размеру данные для передачи их на удалённый сервер. Такой вопрос возникает в том случае, если обработать данные полностью на стороне потребителя становится невозможно, или информация, получаемая в процессе обработки, требует более серьёзных мер защиты. В предложенном ранее методе детектирования падений человека [1], для его представления используется скелет, построенный по семнадцати точкам, предоставленных Microsoft Kinect v2, в качестве описательных признаков используется эвклидова матрица расстояний между точками, скорость изменения этих расстояний между соседними кадрами, а также изменение межкадровых скоростных характеристик. Кроме того, используются высоты выбранных точек скелета. При решении задачи используется одноклассовый классификатор, SVM и CUSUM-метод. Эксперимент проводился на базе данных TST Fall Detection Dataset v2. Такая методика обеспечивает конфиденциальность информации о лицах, использующих RGB-D сенсор ввиду отсутствия данных, которые смогли бы отследить скелетное описание человека. В рамках данной работы выполнено исследование способа сокращения скелетного описания фигуры, что не только позволяет уменьшить объём передаваемой информации на облачные ресурсы, но и повышает точность распознавания факта падения человека. Качество распознавания зависит от тщательного отбора признаков, которые максимально исключают избыточную информацию, приводящую к использованию сложных моделей и переобучению. Прийти к более простому описанию скелета можно сокращая исходную матрицу расстояний, исключая из неё те расстояния, которые при любых движениях человека не меняются, в виду строения его скелета. Также такая процедура ещё больше обезличивает данные и восстановить какое-либо скелетное описание без дополнительной информации становится невозможно. Проведён эксперимент по сокращению скелетного описания фигуры человека в ранее предложенном методе детектирования падений. Произведено исключение восьми расстояний, а именно расстояний между следующими суставами: плечевым и локтевым, локтевым и лучезапястным, тазобедренным и коленным и между коленным и голеностоп-

ным, что повлекло за собой исключение восьми атрибутов скорости изменения расстояний на соседних кадрах и восьми атрибутов изменения скоростных характеристик между соседними кадрами. В итоге удаляется двадцать четыре избыточных признака. Результатом работы стало увеличение точности распознавания факта падения до 0.924 и точности совпадения момента падения на тестовой выборке до 0.881. Общая скорость вычислений с новыми моделями для распознавания была увеличена на 15-20%. Данные результаты были получены процедурой оценки качества, опирающейся на метод последовательного исключения персоны из обучающей выборки (Leave-One-Person-Out).

Работа выполнена при поддержке РФФИ, гранты № 18-07-00942, 18-07-01087.

[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 // International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. – 2019. – Vol. 42. – No. 2/W12, pp. 189-195.

The Study of Skeleton Description Reduction in the Human Fall-Detection Task

Seredin Oleg[1]⋆ oseredin@yandex.ru
Kopylov Andrei[1] And.Kopylov@gmail.com
Surkov Egor[1] eg-su@mail.ru
[1]Tula, Tula State University

In the tasks of processing the data, obtained from cameras and sensors, there is a question of security and, as a consequence, the analysis of these data directly on cloud resources. Edge computing and fog computing allow to execute the main part of calculations on a remote server. Therefore, the data must be impersonal and have a minimum size. This is necessary because bandwidth between sensors and cloud resources is limited. Such new computing paradigm should be apply if it becomes impossible to process the data completely on the consumer side, or the information obtained in the process of analysis requires more serious security measures. In the previously proposed human fall detection method [1], a skeleton constructed from seventeen points is used to represent it. These data were provided by the Microsoft Kinect v2. As descriptive features used the Euclidian Distance Matrix (EDM) of skeleton points is used, as well as the change of distances in neighboring frames and the change of interframe speed features between neighboring frames. In addition, heights of points are taken into account. When solving the problem, a one-class classifier, SVM and CUSUM procedure are used. The experiment was carried out using the data of The TST Fall Detection Dataset v2. This technique ensures the confidentiality of information of people using RGB-D sensor because of the absence of data that be possible to trace the skeleton description of the person. Since the data on which it would be possible to trace the skeletal description of the person are available. Within the frames of this work the study of the method of reducing the skeletal description of the figure was conducted. Was assumed it this would not only reduce the amount of information transmitted to the cloud resources, but also would increase the accuracy of recognition of the fact of a person falling. The accurate recognition depends on a careful selection of features that maximally eliminate redundant information leading to the use of complex models, which leads to over-training. A simpler description of the skeleton should be achieved by reducing the original distance matrix, excluding those distances that, with any movements of a person, do not change because of the structure of his skeleton. Furthermore, such a procedure depersonalizes the data and even more, it becomes impossible to restore any skeletal description without additional information. An experiment to reduce the skeletal description of a human figure in the previously proposed method for detecting falls was conducted. The exclusion of eight distances, namely the distances between the following joints: shoulder and elbow, elbow and wrist, hip and knee and between knee and ankle is made. These distances were resulted in eight attributes of the rate of change of distances on neighboring frames and eight attributes of change

of speed characteristics between neighboring frames. Finally, twenty-four excessive features were removed. The result of this study was an increase in the accuracy of classification to 0.924 and the accuracy of coincidence of the moment of falling on the test sample to 0.881. In addition, the speed of all calculations with new models for recognition was increased by 15-20%. These results were obtained by the quality assessment procedure, which is based on the method of sequential exclusion of a person from the training sample (Leave-One-Person-Out).

This work is supported by RFBR, grants 18-07-00942, 18-07-01087.

[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 // International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. – 2019. – Vol. 42. – No. 2/W12, pp. 189-195.

Исследование метода визуальной навигации по векторной карте в задаче автоматической посадки на Луну

Бобков Александр Валентинович[1,2] Alexander.Bobkov@bmstu.ru
Сюй Ян[2]⋆ Xuyang785506380@gmail.com
[1]Москва, МГТУ им. Н.Э. Баумана

Данная работа посвящена решению задачи определения собственного положения спускаемого аппарата путем сравнения наблюдаемого изображения с электронной бортовой картой. Рассматриваются алгоритмы, использующие модификацию обобщенного преобразования Хафа, исследуются их характеристики, оценивается их пригодность для задач лунной навигации.

Современный этап освоения космического пространства напрямую связан с освоением ближайших небесных тел, таких как Луна и Марс, и развертывании на их поверхности исследовательских станций. Одной из ключевых проблем здесь является выполнение высокоточной мягкой посадки. Предыдущие миссии характеризовались относительно невысокой точностью посадки, которой недостаточно ни для современных исследовательских миссий, ни тем более для задач развертывания и долговременного обслуживания станций.

Радикально повысить точность посадки можно, с одной стороны, путем совершенствования алгоритмов управления посадкой, а с другой – совершенствованием и интеллектуализацией измерительных средств, их комплексированием.

Подсистема визуальной навигации служит для высокоточного определения собственного положения спускаемого аппарата путем сравнения наблюдаемого изображения поверхности Луны с бортовой электронной картой местности. Существующие методы не обеспечивают требованиям решаемой задачи по производительности, устойчивости к изменениям ракурса, освещенности, воздействию шума. Это заставляет искать альтернативные подходы, лишенные подобных недостатков.

В основе предлагаемого подхода лежит использование преобразования Хафа, модифицированного для случая, когда карта задана в векторном виде - в виде списка окружностей-кратеров с известными положением и радиусом.

Предлагаемый алгоритм можно описать следующим образом. Каждой точке границы кадра A гребня кратера можно поставить в соответствие точки B_i каждой i-той окружности карты. Пара точек AB_i образуют гипотезу (x,y) о предполагаемом совпадении точки A кадра и точки B_i карты. Введём пространство гипотез (ПГ) H[x,y] и будем размечать в нём каждую гипотезу, увеличивая соответствующий счётчик H[x,y] на единицу. Верные гипотезы при этом будут совпадать и будут увеличивать один и тот же счётчик, порождая один большой отклик в ПГ. Неправильные гипотезы будут, наоборот, разбросаны хаотически беспорядочно, и существенных откликов порождать не будут. Теперь нахождение объекта на изображении будет сводиться к поиску положения максимального отклика $h^*=H[x^*,y^*]$ в ПГ.

Данный алгоритм производит по одной гипотезе для каждой пары "точка контура – окружность". Количество гипотез для кадра разумных размеров не превышает миллиона, и их учёт вполне может быть выполнен в режиме реального времени даже на относительно маломощных вычислителях.

Исследования показали, что предложенный алгоритм удовлетворяет требованиям точности и производительности, устойчив к изменениям условий видимости и позволяет безошибочно определять положение кадра при наличии в поле зрения хотя бы трех кратеров, обозначенных на карте.

Однако метод оказался чувствительным к изменениям ракурса съемки, которые могут вызываться как неточностями измерения высоты и ориентации аппарата, так и собственными движениями камеры. При этом отклики отдельных кратеров более не будут собираться в одну точку, а будут образовывать область (кластер верных гипотез) с набором характерных пиков. Были опробованы несколько подходов по повышению надежности и оценена эффективность каждого из них.

1) Учет ошибки градиента, возникающей из-за погрешности алгоритма, отклонения формы гребня кратера от окружности и т.д. Из-за этой ошибки вектор градиента более не будет направлен в центр кратера, и его отклик в ПГ будет размыт. Вариант коррекции - каждая точка окружности должна голосовать сразу за несколько положений, с учетом возможной ошибки.

2) Учет изменения ракурса, вызывающих несоответствие ориентации и масштаба кадра и карты. Это приведет к размытию отклика в ПГ и к его распаду на отдельные локальные максимумы. Для восстановления положения и величины максимума использовалось суммирование по окну.

3) Фильтрация шума и мелких деталей. Наличие шума приводит к возникновению ложных точек границ и порождению ложных гипотез в ПГ. Мелкие детали, такие как небольшие сколы и провалы на гребне кратера, будут отклонять форму кратера от окружности, и также будут голосовать за неверные положения. Задача фильтрации – максимально подавить точки границ, не принадлежащих гребням кратеров. В работе использовалась медианная фильтрация с различным размером окна.

4) Нормирование откликов. Области карты, имеющие большую плотность кратеров, при наложении на кадр могут порождать множество совпадающих точек границ, принадлежащих самым разным объектам, например – неровностям рельефа, теням, шуму и т.д. Поэтому отклики в ПГ необходимо нормировать, чтобы учесть различие в плотности расположения кратеров.

Анализ этих дополнений показал, что можно обеспечить достаточную надежность работы алгоритма к небольшим изменениям ракурса, при этом существенно не снижая производительности. Алгоритм оказался работоспособен даже для достаточно сложных условий освещенности, с низкорасположенным Солнцем, что характерно для полярных областей. Это позволяет надеяться, что

предлагаемый подход может быть положен в основу разработки визуальной навигационной системы.

[1] *Бобков А. В., Сюй Ян* Исследование метода визуальной навигации по векторной карте в задаче автоматической посадки на Луну // Машинное обучения и анализ данных, 2019.

Research of the method of visual navigation by a vector map in the task of automatic landing on the Moon

Alexander Bobkov[1] Alexander.Bobkov@bmstu.ru
Yang Xu[1]⋆ Xuyang785506380@gmail.com
[1]Moscow, Bauman Moscow State Technical University

This work is devoted to solving the problem of determining the own position of the descent vehicle by comparing the observed image with the electronic on-board map. Algorithms using a modification of the generalized Hough transform are considered, their characteristics are investigated, their suitability for lunar navigation problems is evaluated.

The modern stage of space exploration is directly related to the development of nearby celestial bodies, such as the Moon and Mars, and the deployment of research stations on their surface. One of the key issues here is the performance of precision precision landing. Previous missions were characterized by relatively low landing accuracy, which is neither sufficient for modern research missions, much less for deployment and long-term station maintenance tasks. Improving the accuracy of landing is possible, on the one hand, by improving landing control algorithms, and, on the other, by improving and intellectualizing measuring instruments and their integration.

The visual navigation subsystem is used for high-precision determination of the own position of the descent vehicle by comparing the observed image of the lunar surface with the on-board electronic map of the area. Existing methods do not allow to satisfy all the requirements of the problem being solved in terms of performance, resistance to changes in angle, lighting, and noise. This leads to the search for alternative approaches, devoid of such shortcomings.

The proposed approach is based on the use of the Hough transform modified for the case when the map is specified in vector form - as a list of crater circles with known position and radius.

The proposed algorithm can be described as follows. Each point of the border of frame A of the crater ridge can be associated with a point B_i of each i-th circle of the map. A pair of points AB_i form the hypothesis (x,y) about the assumed coincidence of point A of the frame and point B_i of the map. We introduce the hypothesis space (HS) H [x, y] and mark each hypothesis in it, increasing the corresponding counter H[x,y] by one. The correct hypotheses will coincide and will increase the same counter, generating one large response in the HS. Wrong hypotheses, on the contrary, will be scattered randomly randomly, and they will not generate significant responses. Now finding the object in the image will come down to finding the position of the maximum response $h^* = H[x^*,y^*]$ in the HS.

This algorithm produces one hypothesis for each pair "contour point - circle". The number of hypotheses for a frame of reasonable size does not exceed a million,

and their accounting can very well be performed in real time even on relatively low-power computers.

Studies have shown that the proposed algorithm satisfies the requirements of accuracy and performance, is resistant to changes in visibility conditions, and allows you to accurately determine the position of the frame when there are at least three craters in the field of view indicated on the map.

The considered method turned out to be sensitive to changes in the shooting angle, which can be caused both by inaccuracies in measuring the height and orientation of the apparatus, and by the camera's self-movement. In this case, the responses of individual craters will no longer be collected at one point, but will form a region (a cluster of correct hypotheses) with a set of peaks. Several approaches were tested and the effectiveness of each of them was evaluated.

1) Taking into account the gradient error arising due to the error of the algorithm, deviation of the shape of the crest of the crater from the circle etc. Due to this error, the gradient vector will no longer be directed to the center of the crater, and its response in the HS will be blurred. Correction option - each point of the circle must vote for several positions at once, taking into account a possible error.

2) Accounting for changes in angle, causing a mismatch in the orientation and scale of the frame and map. This will lead to a smearing of the response in the HS and to its decay into individual local maxima. To restore the position and maximum value, summation over the window was used.

3) Filtering noise and small parts. The presence of noise leads to the appearance of false points of boundaries and the generation of false hypotheses in the HS. Small details, such as small chips and dips on the crest of the crater, will deflect the shape of the crater from the circle, and will also vote for incorrect positions. The task of filtering is to suppress as much as possible the points of boundaries that do not belong to the crests of craters. We used median filtering with different spot sizes.

4) Rationing of responses. When overlaid on a frame, areas of a map that have a high density of craters can generate many coincident boundary points that belong to a wide variety of objects, for example, uneven terrain, shadows, noise, etc. Therefore, the responses in the HS need to be normalized to take into account the difference in the density of the craters.

An analysis of these additions showed that it is possible to ensure sufficient reliability of the algorithm to small changes in angle, while not significantly reducing performance. The algorithm turned out to be workable even for rather difficult lighting conditions, with a low-lying Sun, which is typical for polar regions. This allows us to hope that the proposed approach can be the basis for the development of a visual navigation system.

[1] *Bobkov A.V., Xu Yang* Research of the method of visual navigation by a vector map in the task of automatic landing on the Moon // Machine Learning and Data Analysis, 2019.

Диагностика водно-этанольных растворов по спектрам комбинационного рассеяния с помощью искусственных нейронных сетей: методы повышения устойчивости решения к искажениям спектров

Исаев Игорь Викторович[1,2]⋆ isaev_igor@mail.ru
Буриков Сергей Алексеевич[1,2] sergey.burikov@gmail.com
Доленко Татьяна Альдефонсовна[1,2] tdolenko@mail.ru
Лаптинский Кирилл Андреевич[1,2] onelumen@gmail.com
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, Физический факультет МГУ имени М.В.Ломоносова

В данной работе рассматривается добавление шума при обучении нейронной сети как метод повышения устойчивости решения к шумам в данных. Метод тестируется при решении обратной задачи спектроскопии комбинационного рассеяния света водно-этанольных растворов, для специального типа искажений, вызываемого изменениями мощности лазерной накачки, которые ведут к сжатию или растяжению спектра.

Подтверждено, что чем выше уровень шума в тренировочном наборе данных, тем медленнее качество решения деградирует с увеличением уровня шума в тестовом наборе.

Далее метод был протестирован на спектрах реальных алкогольных напитков. Обнаружено, что их спектры существенно отличаются от спектров использованного для обучения набора растворов, эмулирующих алкогольные напитки. Поэтому приемлемые результаты показывают только сети, обученные с добавлением шума. При добавлении 20% гауссова шума среднее отклонение при определении концентрации этанола составило 2.07%.

Исследование выполнено за счёт гранта Российского Научного фонда, проект № 19-11-00333.

[1] *Isaev, I. et al.* Diagnostics of Water-Ethanol Solutions by Raman Spectra with Artificial Neural Networks: Methods to Improve Resilience of the Solution to Distortions of Spectra // Studies in Computational Intelligence, V.856. Springer Nature, 2020. — p. 319–325. https://doi.org/10.1007/978-3-030-30425-6_37.

Diagnostics of Water-Ethanol Solutions by Raman Spectra with Artificial Neural Networks: Methods to Improve Resilience of the Solution to Distortions of Spectra

Igor Isaev[1,2]⋆ isaev_igor@mail.ru
Sergey Burikov[1,2] sergey.burikov@gmail.com
Tatiana Dolenko[1,2] tdolenko@mail.ru
Kirill Laptinskiy[1,2] onelumen@gmail.com
Sergey Dolenko[1] dolenko@srd.sinp.msu.ru
[1]D.V.Skobeltsyn Institute of Nuclear Physics, M.V.Lomonosov Moscow State University
[2]Physical Department, M.V.Lomonosov Moscow State University

In this study, we consider adding noise during training of a neural network as a method of improving the stability of its solution to noise in the data. We tested this method in solving the inverse problem of Raman spectroscopy of aqueous ethanol solutions, for a special type of distortion caused by changes in the power of laser pump leading to compression or stretching of the spectrum.

It has been confirmed that the higher is the noise level in the training set, the slower the solution quality decreases with increase of the noise level in the test set.

The method was also tested on the spectra of real alcoholic beverages. It has been found that their spectra differ significantly from those in the dataset simulating alcoholic beverages, used for training. Therefore, acceptable results were demonstrated only by networks trained with adding noise. Networks trained with the addition of 20% Gaussian noise showed an average deviation of 2.07 % vol in determination of ethanol concentration.

This study has been performed at the expense of Russian Science Foundation, grant no.19-11-00333.

[1] *Isaev, I. et al.* Diagnostics of Water-Ethanol Solutions by Raman Spectra with Artificial Neural Networks: Methods to Improve Resilience of the Solution to Distortions of Spectra // Studies in Computational Intelligence, V.856. Springer Nature, 2020. — p. 319–325. https://doi.org/10.1007/978-3-030-30425-6_37.

Определение вида и параметров искажений изображения по Фурье-спектру сигнала

Чочиа Павел Антонович[1]⋆ chochia@iitp.ru
[1]Москва, ИППИ РАН

Исследуется диагностика искажений изображения по Фурье-спектру получаемого сигнала. Рассматриваются вопросы определения типа и параметров линейных однородных сглаживающих операторов следующих видов: круговой формы прямоугольного профиля, круговой формы Гауссова профиля и линейной формы прямоугольного профиля. Спектральная плотность изображения анализируется с применением *среднего радиального профиля* и *среднего профиля по направлению*. Предложены и исследованы алгоритмы диагностики для каждого из рассматриваемых типов искажений. Показано влияние точности данных и шума на достоверность результатов анализа, а также возможности диагностики при суперпозиции нескольких искажений.

Предполагаем искажающий оператор пространственно-инвариантным $h(x-u, y-v)$, шум $\xi(x,y)$ некоррелированным и аддитивным, а процесс искажения — линейным. Тогда получаемое изображение $f(x,y)$ описывается интегралом свертки:

$$f(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(u,v)h(x-u,y-v)dudv + \xi(x,y).$$

Данное уравнение эквивалентно произведению Фурье-спектров сигнала $G(\omega_x,\omega_y)$ и искажающего оператора $H(\omega_x,\omega_y)$ в сумме со спектром шума $\Xi(\omega_x,\omega_y)$: $F(\omega_x,\omega_y) = H(\omega_x,\omega_y)G(\omega_x,\omega_y) + \Xi(\omega_x,\omega_y)$. Таким образом, с точностью, определяемой уровнем шума $\xi(x,y)$, модуль спектральной плотности (МСП) получаемого изображения $f(x,y)$ равен произведению МСП исходного сигнала $g(x,y)$ и ядра искажающего оператора $h(x,y)$.

Для диагностики параметров искажающих операторов, имеющих круговую симметрию, введем понятие *среднего радиального профиля* модуля спектральной плотности $P(\omega)$ как функции частоты. Пусть $M(\omega,\varphi)$ — амплитуда спектра $F(\omega_x,\omega_y)$ в полярных координатах. Усреднив $M(\omega,\varphi)$ по φ, получим среднее радиальное значение (профиля) МСП для частоты ω:

$$P(\omega) = \int_{\varphi} M(\omega,\varphi)d\varphi/2\pi\omega.$$

Эксперименты над реальными изображениями показывают, что значения $P(\omega)$ убывают быстро с ростом ω, а форма профиля $P(\omega)$ слабо зависит от сюжета изображения.

Круговое рассеяние. Оператор кругового рассеяния прямоугольного профиля радиуса r имеет вид: $h_{\mathrm{C}}(x,y) = 1/(\pi r^2)$ для $x^2+y^2 \leqslant r^2$, и $h_{\mathrm{C}}(x,y) = 0$ в остальных точках. Спектр $h_{\mathrm{C}}(x,y)$ выглядит как набор концентрических колец

с нулями на радиусах $\omega = \rho n$, где $\rho = \pi/r$, а $n = 1, 2, \ldots$. Нули также будут присутствовать в двумерном спектре изображения и могут быть обнаружены в профиле $P(\omega)$. Период ρ минимумов/максимумов $P(\omega)$ отыскивается как точка максимума функционала

$$\rho_{\mathrm{C}} = \underset{\rho_{\min} \leqslant \rho \leqslant \rho_{\max}}{\arg\max} \left(\frac{1}{[W/2\rho] - 1} \sum_{n=1}^{[W/2\rho]-1} (P(n\rho + \rho/2) - P(n\rho)) \right),$$

где W — размер изображения, а $[\cdot]$ — взятие целой части. Радиус рассеяния r_{C} находится как $r_{\mathrm{C}} = W/(2\rho_{\mathrm{C}})$.

Гауссово рассеяние. Оператор Гауссова рассеяния с круговой симметрией $h_{\mathrm{G}}(x, y) = C \exp\{-(x^2 + y^2)/2\sigma^2\}$ имеет профиль нормального распределения; тот же вид и у профиля МСП Гауссова оператора. Средний радиальный профиль МСП будет:

$$P(\omega) = C \exp\{-\omega^2/2\sigma^2\} G_{\mathrm{M}}(\omega) + \Xi_{\mathrm{M}}(\omega).$$

Здесь $G_{\mathrm{M}}(\omega)$ и $\Xi_{\mathrm{M}}(\omega)$ — средние радиальные профили МСП неискаженного сигнала и шума. Первое слагаемое спадает быстро с ростом ω, а второе можно считать постоянным.

При больших ω, когда $C \exp\{-\omega^2/2\sigma^2\} G_{\mathrm{M}}(\omega) < \Xi_{\mathrm{M}}(\omega)/2$, значения $\log P(\omega)$ близки к константе, определяемой уровнем шума Ξ_{M}. При малых ω, когда $C \exp\{-\omega^2/2\sigma^2\} G_{\mathrm{M}}(\omega) > 2\Xi_{\mathrm{M}}(\omega)$, получим: $\log P(\omega) \approx -\omega^2/2\sigma^2 + \log G_{\mathrm{M}}(\omega) + C_2$. Здесь $A(\omega) = -\omega^2/2\sigma^2$ определяется Гауссовым рассеянием, $B(\omega) = \log G_{\mathrm{M}}(\omega)$ зависит от конкретного изображения, а C_2 — константа.

Составляющая $A(\omega)$ задается формулой $y = -ax^2 + b$ и всюду выпукла вверх; оценка Гауссова рассеяния сводится к нахождению значения a. На интересующем участке ω кривая $B(\omega)$ выпукла вниз, причем амплитуда ее выпуклости ниже, чем у $A(\omega)$. Значение a находится с помощью преобразования Хафа. Зная a, дисперсия Гауссова сглаживания вычисляется как $\sigma_{\mathrm{G}} = 1/\sqrt{2a}$.

Линейный смаз вызывается равномерным прямолинейным смещением наблюдаемой сцены. Оператор смаза эквивалентен свертке изображения с отрезком длиной r, повернутым на угол α $(0 \leqslant \alpha < \pi)$; его формула: $h_{\mathrm{S}}(x, y) = 1/r$ для $y = x \tan\alpha$ при $x^2 + y^2 \leqslant r^2/4$, и $h_{\mathrm{S}}(x, y) = 0$ в остальных точках.

Амплитуда спектра оператора смаза имеет вид полос, перпендикулярных направлению смаза α, с периодом W/r, где W — размер изображения, а r — величина смаза. Для диагностики подобных искажений воспользуемся вариантом кепстрального преобразования: $C(q_1, q_2) = |\mathcal{F}^{-1}\{\log|S(\omega_x, \omega_y)|^2\}|$. Здесь $|\cdot|$ — операция модуля, а $\mathcal{F}^{-1}\{\cdot\}$ — обратное преобразование Фурье. В кепстре искаженного смазом изображения появится проходящая через начало координат линия повышенных значений, совпадающая с направлением смаза, наклон которой α легко определяется.

Период полос ρ_L находится через максимум функционала

$$\rho_L = \underset{\rho_{\min}\leqslant\rho\leqslant\rho_{\max}}{\arg\max} \left(\frac{1}{[W/\rho]-1} \sum_{n=1}^{[W/\rho]-1} (P_\alpha(n\rho+\rho/2) - P_\alpha(n\rho)) \right).$$

Здесь $P_\alpha(q)$ — *средний профиль* МСП изображения *по направлению*, перпендикулярному смазу. Получив значение ρ_L и зная размер изображения W, вычисляется величина смаза: $r_{\mathrm{S}} = W/\rho_L$.

Эксперименты показали, что точность диагностики в меньшей степени зависит от размеров искажающих операторов, но в большей — от наличия шума и характеристик сигнала. Если изображение подверглось нескольким искажениям одновременно, то описанные алгоритмы уверенно диагностируют суперпозицию искажений в случае точного представления сигнала. Снижение точности и наличии шума заметно ухудшает диагностику.

[1] *Чочиа П. А.* Диагностика линейного однородного искажающего оператора по спектру наблюдаемого изображения // Информационные процессы, Москва, 2019, т. 19, № 3, С. 313–326.

Estimation of Image Distortion Type and Parameters from Fourier Spectrum of Signal

Chochia Pavel[1]⋆ chochia@iitp.ru

[1]Moscow, IITP RAS

We consider the problem of assessing the type and parameters of distortions from Fourier spectrum of observed image (the *blind estimation* task). We explore the questions of diagnostics of linear homogeneous distorting operators of three types: circular form with rectangular profile, circular form with Gaussian profile, and motion blur ones. The *average radial profile* and *average directional profile* of spectrum density are used for distortion identification. The diagnostics and parameters estimation algorithms for every considered distortion types are proposed and researched. The influence of noise and data accuracy to the reliability of the results is demonstrated.

Let an image $g(u, v)$ is distorted by spatially invariant operator $h(x - u, y - v)$ and non-correlated additive noise $\xi(x, y)$, and the distortion process is linear. Then obtained image $f(x, y)$ may be described by the following convolution integral:

$$f(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(u, v)h(x - u, y - v)dudv + \xi(x, y).$$

It is equivalent to the product of Fourier spectra of source signal $G(\omega_x, \omega_y)$ and distorting operator $H(\omega_x, \omega_y)$ in addition with noise spectrum $\Xi(\omega_x, \omega_y)$: $F(\omega_x, \omega_y) = = H(\omega_x, \omega_y)G(\omega_x, \omega_y) + \Xi(\omega_x, \omega_y)$. That is, with an accuracy of noise level $\xi(x, y)$, the modulo of spectral density (MSD) of obtained image $f(x, y)$ equals to the product of MSD's of source signal $g(x, y)$ and distorting operator kernel $h(x, y)$.

To find the parameters of distorting operators with circular symmetry, we introduce the concept of *average radial profile* of MSD $P(\omega)$ as the function of frequency. Let $M(\omega, \varphi)$ is the spectrum $F(\omega_x, \omega_y)$ amplitude in polar coordinates. Averaging $M(\omega, \varphi)$ over φ, we obtain the average radial value of MSD (profile) for the frequency ω:

$$P(\omega) = \int_{\varphi} M(\omega, \varphi)d\varphi/2\pi\omega.$$

The experiments on real images demonstrate that $P(\omega)$ values quickly descend when ω growth, and the shape of profile $P(\omega)$ weakly depends on an image subject.

Circular dissipation. The distorting operator of circular dissipation with rectangular profile and scattering radius r, centered at zero, is: $h_{\mathrm{C}}(x, y) = 1/(\pi r^2)$ for $x^2 + y^2 \leqslant r^2$, and $h_{\mathrm{C}}(x, y) = 0$ in other points. The spectrum of $h_{\mathrm{C}}(x, y)$ looks like a set of concentric rings with zeroes at radii $\omega = \rho n$, where $\rho = \pi/r$, and $n = 1, 2, \ldots$. These zeroes will also be present in two-dimensional image spectrum and may be detected in the profile $P(\omega)$. The period ρ of minima/maxima in $P(\omega)$ is found as

the point of maximum of the functional

$$\rho_{\mathrm{C}} = \underset{\rho_{\min}\leqslant\rho\leqslant\rho_{\max}}{\arg\max}\left(\frac{1}{[W/2\rho]-1}\sum_{n=1}^{[W/2\rho]-1}\left(P(n\rho+\rho/2)-P(n\rho)\right)\right),$$

where W is the image size, and $[\cdot]$ is the integer part operation. The scattering radius r_{C} is calculated as $r_{\mathrm{C}} = W/(2\rho_{\mathrm{C}})$.

Gaussian dissipation operator of distortion with circular symmetry $h_{\mathrm{G}}(x,y) = C\exp\{-(x^2+y^2)/2\sigma^2\}$ has the profile of normal distribution; MSD of Gaussian operator has the same shape. Average radial profile of MSD of Gaussian dissipation will be:

$$P(\omega) = C\exp\{-\omega^2/2\sigma^2\}G_{\mathrm{M}}(\omega) + \Xi_{\mathrm{M}}(\omega).$$

Here $G_{\mathrm{M}}(\omega)$ and $\Xi_{\mathrm{M}}(\omega)$ are the average radial profiles of MSD's for no distorted image and for noise. The first item drops down rapidly when ω growth, while the second one is almost constant.

Under large ω, when $C\exp\{-\omega^2/2\sigma^2\}G_{\mathrm{M}}(\omega) < \Xi_{\mathrm{M}}(\omega)/2$, the values $\log P(\omega)$ are close to a constant definable by the noise level Ξ_{M}. Under small ω, when $C\exp\{-\omega^2/2\sigma^2\}G_{\mathrm{M}}(\omega) > 2\Xi_{\mathrm{M}}(\omega)$, we obtain: $\log P(\omega) \approx -\omega^2/2\sigma^2 + \log G_{\mathrm{M}}(\omega) + C_2$. Here $A(\omega) = -\omega^2/2\sigma^2$ is defined by Gaussian dissipation, $B(\omega) = \log G_{\mathrm{M}}(\omega)$ depends on specific image, and C_2 is a constant.

The component $A(\omega)$ is specified by $y = -ax^2 + b$ law and everywhere is convex upwards; the estimate of Gaussian dissipation consists in finding the value a. The component $B(\omega)$ is convex downwards in the interesting section of ω, and the amplitude of its curvature is less than $A(\omega)$ curvature. The value a is evaluated with the help of Hough transform. When a is known, the dissipation of Gaussian smoothing is calculated as $\sigma_{\mathrm{G}} = 1/\sqrt{2a}$.

Linear blur appears due to uniform rectilinear displacement of observable scene. This blur is equivalent to convolution of an image with the segment of length r, rotated at angle α ($0 \leqslant \alpha < \pi$), and is described by the formula: $h_{\mathrm{S}}(x,y) = 1/r$ for $y = x\tan\alpha$ under $x^2 + y^2 \leqslant r^2/4$, and $h_{\mathrm{S}}(x,y) = 0$ in other points.

The amplitude of linear blur operator spectrum looks like a set of strips that are perpendicular to blur direction α. The period of strips equals W/r, where W is the image size and r is the blur value. For diagnostic of these distortions let use the variant of cepstrum transformation: $C(q_1,q_2) = |\mathcal{F}^{-1}\{\log|S(\omega_x,\omega_y)|^2\}|$. Here $|\cdot|$ is the modulo operation and $\mathcal{F}^{-1}\{\cdot\}$ is the inverse Fourier transform. In the blurred image cepstrum, the line of increased values will appear, which passes through the origin. The skewing angle of the line coincides with the blur direction α, and may be simply determined.

The strips period ρ_L is evaluated using the maximum of functional

$$\rho_L = \underset{\rho_{\min}\leqslant\rho\leqslant\rho_{\max}}{\arg\max}\left(\frac{1}{[W/\rho]-1}\sum_{n=1}^{[W/\rho]-1}\left(P_\alpha(n\rho+\rho/2)-P_\alpha(n\rho)\right)\right).$$

Here $P_\alpha(\omega)$ is the *average directional profile* of image's MSD in direction perpendicular to blurring one. When ρ_L value and the image size W are known, the length of the blur is calculated as $r_S = W/\rho_L$.

The experiments demonstrate that the accuracy of diagnostic less depends on the size of distorting operators, but more on the signal characteristics and the noise level. If an image was subjected to several distortions simultaneously, the described algorithms enough confidently detect the superposition of distortions when the signal is in high accuracy form. The reduction of accuracy and the presence of noise appreciably impair the diagnostic quality.

[1] *Chochia P.* Estimation of Linear Homogeneous Distorting Operator from Observed Image Spectrum // Information processes, Moscow, 2019, vol. 19, no. 3, pp. 313–326.

Мера TF–IDF и оценка близости смысловому эталону заголовков и аннотаций научных статей

Михайлов Дмитрий Владимирович[1]⋆ mdv74@list.ru
Емельянов Геннадий Мартинович[1] Gennady.Emelyanov@novsu.ru
[1]Великий Новгород, Россия, НовГУ

При подготовке учебного материала для реализации электронного обучения преподаватель должен иметь доступ к некоторому срезу информационного пространства, элементами которого являются публикации либо Internet-страницы, релевантные учебному курсу. Актуальная здесь задача — сортировка источников информации по степени отражения наиболее существенных понятий изучаемой предметной области (ПО) при максимальной компактности и безызбыточности изложения. Первостепенную роль при этом играет поиск набора единиц текста и их связей, необходимого и достаточного для представления единицы знаний и отвечающего смысловому эталону. Предлагаемый метод оценки близости эталону [1] не требует перефразирования текста и основан на разбиении слов каждой его фразы на классы по значению меры TF-IDF относительно текстов корпуса, предварительно формируемого экспертом. В роли анализируемых текстов выступают аннотации научных статей вместе с их заголовками. При этом смысловые образы наиболее близких эталону текстов определяют слова с наибольшими значениями TF-IDF, которые при расположении по соседству во фразе с наибольшей вероятностью связаны по смыслу и образуют ключевые сочетания [2] вместе со словами, близкими среднему значению данной меры. Для отнесения сочетаний слов к ключевым в работе вводится интерпретация меры TF-IDF, оценивающая число одновременных вхождений всех слов анализируемого сочетания во фразы отдельного документа. При подсчете общего числа слов документа здесь раздельно учитываются случаи совместной встречаемости слов сочетания и встречаемость без одновременного вхождения во фразу. Само значение TF-IDF ключевого сочетания должно быть не ниже минимума указанной меры по его отдельным словам.

Оценка близости отдельной фразы эталону, не требующая поиска перифраз для языкового описания соответствующей единицы знаний, строится из следующих эмпирических соображений.

Во-первых, разделение на общую лексику и термины должно быть выражено как можно в большей степени, а слова в кластерах, формируемых по TF-IDF — распределены более или менее равномерно. Кроме того, число получившихся кластеров должно стремиться к трём при максимуме TF-IDF для слов кластера наибольших значений указанной меры. Данное требование означает максимальную релевантность терминов в составе фраз отбираемого документа сформированному корпусу.

Для группы фраз, первая из которых — заголовок статьи, а остальные представляют аннотацию, вводятся два варианта оценки близости эталону, в равной

мере предусматривающие минимум среднеквадратического отклонения значения близости эталону по всем фразам группы. Первый подразумевает максимальную близость эталону для заголовка и отвечает общепринятому в научной периодике требованию отражения в заголовке содержания статьи. Второй вариант предполагает максимизацию близости эталону по всем фразам анализируемого текста. Максимальный итоговый рейтинг по коллекции, из которой производится отбор, получает статья с наибольшим значением первого варианта, попадающим в один кластер со значением второго варианта оценки для той же статьи. При этом значения первого варианта оценки для статьи, получившей максимальный итоговый рейтинг, и максимального значения первого варианта оценки по коллекции должны относиться к одному кластеру. В случае отсутствия в коллекции статьи, отвечающей данному требованию, максимальный итоговый рейтинг получает статья с наибольшим значением первого варианта оценки по анализируемой коллекции.

Предложенный метод даёт минимум трёхкратное сокращение числа документов, с которыми следует ознакомиться в первую очередь при изучении заданной ПО, например, студентами.

Работа поддержана грантом РФФИ № 19-01-00006.

[1] *Mikhaylov D., Emelyanov G.* Estimation by phrases for the closeness of a topical text to the semantic pattern without paraphrasing // Interactive Systems: Problems of Human-Computer Interaction. Collection of scientific papers, Ульяновск: УлГТУ, 2019 (в печати).

[2] *Михайлов Д. В., Емельянов Г. М.* Оценка близости тематического текста смысловому эталону без конструирования перифраз // Pattern Recognition and Image Analysis, 2019. Т. 29, №4 (в печати).

TF"=IDF metrics and estimation of affinity to a sense standard for titles and abstracts of scientific articles

Dmitry Mikhaylov[1]⋆ mdv74@list.ru
Gennady Emelyanov[1] Gennady.Emelyanov@novsu.ru
[1]Russia, Veliky Novgorod, Yaroslav-the-Wise Novgorod State University

The preparing of teaching material in e"=learning requires to have access for a teacher to a certain section of informational space, the elements of which are publications or Internet pages relevant to the course for study. The actual problem here is the sorting of information sources by degree of reflection of the most significant concepts of the studied subject area at a maximal compactness and non"=redundancy of narration. A primary role here plays a revelation of a set of text units and their relations necessary and enough to represent a knowledge unit and satisfies the sense standard. The offered method for estimation of the closeness to the sense standard [1] does not require paraphrasing the text and is based on the splitting of words of each its phrase into classes by the value of TF"=IDF metric relative to the texts of corpus pre"=formed by expert. Herewith as the analyzed texts the abstracts of scientific articles together with their titles are considered. These parts of articles reflect the main content of each paper and the most important results without unnecessary methodological details. Here the sense images of texts closest to the sense standard are defined by words with the greatest values of TF"=IDF which being neighbors in a phrase be related by sense most probably and form the keyword combinations [2] together with the words close to average value of this measure. To select keyword combinations from defining the semantic image of phrase the interpretation of TF"=IDF metrics which respects a number of simultaneous occurrences of all words of analyzed combination in the phrases of separate document of corpus is entered into consideration in current paper. When calculating the total number of document words we'll separately take into account the cases of co"=occurrence of combination words and occurrence without simultaneous presence in a phrase. Herewith the value of TF"=IDF metrics for key word combination should not be less than the minimum of values of mentioned measure for its separate words.

The estimation of the closeness of a separate phrase to the sense standard without paraphrasing the natural"=language description of a corresponding knowledge unit is based on the following empirical consideration. First, the division of words into general vocabulary and terms here should be expressed as greatly as possible. Another moment is that the words in clusters formed by TF"=IDF should be distributed more or less evenly. In addition, the number of resulted clusters must be close to three as much as possible at maximum of TF"=IDF for words related to the cluster of greatest values of mentioned measure. This requirement means the maximal relevance of term words in phrases of selected documents to the formed corpus.

For a group of phrases, first of which is the title of article and others represent its abstract, two variants for estimation of the affinity to the sense standard are introduced. Both variants are equally assumed the minimum of root"=mean"=square deviation for value of affinity to the standard for all phrases of group. The first variant assumes the maximal closeness to the standard for the article title and corresponds to the requirement general accepted in scientific periodicals to reflect in the title the paper content. The second variant assumes maximizing the affinity to the standard for all phrases of analyzed text. The maximal final rank in the collection for paper selection will be designated to the article with a greatest value of the first variant of estimation related to the same cluster with the value of the second variant for the same paper. Herewith the value of the first estimation variant for article with a maximal final rank, and a maximal value of the first estimation variant in the collection must be in the same cluster. In a case of absence of article meets this requirement, the maximal final rank will be designated to the article with a greatest value of the first variant of estimation in analyzed collection.

The proposed method gives at least a threefold reduction in the number of documents (i. e. scientific articles) that should be read first when studying a given subject area, for example, by students.

This research is funded by RFBR, grant 19-01-00006.

[1] *Mikhaylov D., Emelyanov G.* 2019. Estimation by phrases for the closeness of a topical text to the semantic pattern without paraphrasing. *Interactive Systems: Problems of Human"=Computer Interaction. Collection of scientific papers.* Ulyanovsk: ULSTU (in press).

[2] *Mikhaylov D., Emelyanov G.* 2019. Estimation of the closeness to a semantic pattern of a topical text without construction of periphrases. *Pattern Recognition and Image Analysis.* 29(4) (in press).

Применение многомерных формальных контекстов в анализе текстов естественного языка

Богатырев Михаил Юрьевич[1]⋆ okkambo@mail.ru
Самодуров Кирилл Викторович[1] zmeymc@gmail.com
[1]Тула, Тульский государственный университет

В системах информационного поиска и анализа текстов большое значение имеет разработка моделей, отражающих семантику обрабатываемых данных. Особенно это актуально для текстов естественного языка. Одним из направлений, претендующих на создание семантических моделей данных, является концептуальное моделирование. Концептуальная модель представляет собой граф, вершины которого определяются как понятия (концепции, концепты), а с помощью рёбер графа задаются отношения на множестве понятий.

К концептуальному моделированию относится Анализ формальных понятий (АФП) – раздел прикладной теории решёток, в котором исследуются иерархические связи на данных, заданных объектно-признаковыми описаниями, и образующими формальный контекст. Концептуальной моделью здесь является решётка понятий.

В работе [1] предложен метод извлечения фактов из текстов естественного языка, основанный на применении двух концептуальных моделей: концептуальных графов и решёток понятий. *Концептуальный граф* - это двудольный направленный граф, состоящий из двух типов вершин: концептов и концептуальных отношений. Концепты – это слова из обрабатываемого текста, а отношения формируются алгоритмом построения концептуальных графов.

Метод извлечения фактов [1] включает следующие этапы.

1. На предложениях обрабатываемых текстов строится множество концептуальных графов. Для этого используется разработанный нами генератор концептуальных графов для англо- и русскоязычных текстов.

2. На множестве концептуальных графов решается задача их агрегирования. Агрегирование необходимо для исключения избыточной размерности концептуальных моделей, не связанной с полезной информацией. Для агрегирования применяются эвристики отбора значимых данных, присущие конкретной задаче извлечения фактов, а также кластеризация концептуальных графов.

3. На агрегированном множестве концептуальных графов строится формальный контекст. Формальный контекст $\mathbf{K} = (G,M,I)$ представляет собой отношение $I \subseteq G \times M$ на множествах объектов G и их атрибутов M и задаётся матрицей, реализующей данное отношение. Построение формального контекста на текстах является самой сложной задачей метода. Для её решения применяется алгоритм отбора элементов концептуальных графов в формальный контекст, который должен максимально сохранять «семантическую выразительность» концептуальных графов, используемую в дальнейшем. Такой алгоритм работает

с данными из внешних ресурсов – словарей, тезаурусов, онтологий, которые выбираются на основании тематики обрабатываемых текстов.

4. На формальном контексте выделяются формальные понятия и строится другая концептуальная графовая модель – решётка понятий. Пара подмножеств $(A, B), A \subseteq G, B \subseteq M$ таких, что $A^{'} = B, B^{'} = A$, называется формальным понятием контекста **K**. Здесь штрихом обозначается оператор, реализующий связь между объектами и атрибутами в силу отношения I. В матрице контекста понятия (A,B) задаются максимальными по вложению подматрицами со всеми ненулевыми элементами. Частично упорядоченное по вложению объёмов множество формальных понятий контекста **K** образует математический объект - решётку, которая называется *решётка понятий*. Имея решётку понятий, можно выявлять связи между понятиями по принципу «общее – частное». Понятия – узлы решётки – интерпретируются как множество потенциальных фактов определённого уровня (тематики), которое связано с другими фактами.

Доклад содержит результаты исследований, развивающих метод концептуального моделирования [1] в направлении повышения семантической выразительности формальных контекстов и как следствие – повышения качества извлечения фактов из текстовых данных.

Термин «семантическая выразительность» уточняется с помощью следующей гипотезы. Если для моделирования смысла текста применяется концептуальная модель, в которой имеет место понятие размерности, то чем выше размерность такой модели, тем глубже она позволяет моделировать смысл текста. Соответственно, концептуальная модель более высокой размерности обладает и большей семантической выразительностью. Для проверки данной гипотезы необходимы не только двумерные, но также трёх- и в общем случае, n – мерные формальные контексты.

В докладе представлены результаты экспериментов по построению трёх- и четырёхмерных формальных контекстов на данных, формируемых для стандартных задач извлечения фактов из текстов биомедицинской тематики BioNLP Shared Tasks. Данные представляют собой тексты аннотаций статей из системы PubMed, слабо связанных друг с другом, но в целом посвящённых исследованию мутаций генов.

Для построения формальных контекстов применялись концептуальные графы, агрегированные посредством выделения в них AMR-схем. Модель Абстрактного представления смысла (Abstract Meaning Representation – AMR) известна как макро-модель, применяемая для моделирования семантики текстов. Трёхмерные контексты строятся по трёхэлементной схеме с парами семантических ролей типа «агенс» и «пациенс». Для построения четырёхмерных контекстов применялась четырёхэлементная AMR-схема «что – как влияет(глагол) – на что- атрибут влияния»

Результаты экспериментов демонстрируют справедливость гипотезы семантической выразительности. Четырёхмерные формальные контексты позволяют

посредством понятий и кластеров устанавливать значительно больше связей между текстами, чем трёхмерные контексты. При этом вычислительная сложность использованного нами алгоритма построения трёх- и четырехмерных контекстов не превышает известных оценок.

Работа поддержана грантом РФФИ № 19-07-01178.

[1] *Bogatyrev Mikhail* Fact Extraction from Natural Language Texts with Conceptual Modeling // Communications in Computer and Information Science, Vol. 706, Springer-Verlag, 2017. — Pp. 89–102.

Application of Multidimensional Formal Contexts in Natural Language Texts Analysis

Mikhail Bogatyrev[1]⋆ okkambo@mail.ru
Kirill Samodurov[1] zmeymc@gmail.com
[1]Tula, Tula Sate University

The development of models reflecting the semantics of the processed data is of great importance in the systems of information retrieval and text analysis. This is especially true for natural language texts. One of the areas that claim to create semantic data models is conceptual modeling. A conceptual model is a graph whose vertices are defined as concepts and whose edges are used to define relations on the set of concepts.

Conceptual modeling includes Formal Concept Analysis (FCA), a branch of applied lattice theory that explores hierarchical relationships on data defined by object-attribute descriptions that form a formal context. The conceptual model here is a conceptual lattice.

In [1] the method of fact extraction from natural language texts is proposed, it is based on application of two conceptual models : conceptual graphs and of conceptual lattices. *Conceptual graph* is a bipartite directed graph consisting of two types of vertices: concepts and conceptual relations. Concepts are words from the processed text, and relations are formed by the algorithm of constructing conceptual graphs.

The fact extraction method [1] includes the following steps.

1. A set of conceptual graphs are acquired from the sentences of the processed texts. To do this, we use our generator of conceptual graphs for English and Russian texts.

2. On the set of conceptual graphs, the problem of their aggregation is solved. Aggregation is necessary to eliminate the excessive dimension of conceptual models that are not associated with useful information. For aggregation, heuristics of the selection of significant data inherent in the particular fact extraction task of are used, as well as clustering of conceptual graphs.

3. A formal context is constructed on the aggregated set of conceptual graphs. The formal context $\mathbf{K} = (G,M,I)$ is a relation $I \subseteq G \times M$ on sets of objects G and their attributes M and is given by a matrix implementing this relation. Building a formal context on texts is the most difficult task of the method. To solve this problem, we use an algorithm for selecting elements of conceptual graphs into a formal context, which should preserve as much as possible the "semantic expressiveness" of conceptual graphs, which is used in the future. This algorithm works with data from external resources-dictionaries, thesauruses, ontologies, which selection is based on the subject of the processed texts.

4. Formal concepts are distinguished on the formal context and another conceptual graph model-the conceptual lattice-is constructed. A pair of subsets (A, B), $A \subseteq G, B \subseteq M$ such that $A^{'} = B, B^{'} = A$ is called formal concept of a context $\mathbf{K}$. Here,

a stroke denotes an operator that implements the relationship between objects and attributes by virtue of the I relationship. In the context matrix, the concepts (A,B) are given by embedding-maximal submatrices with all nonzero elements. The set of formal concepts in a context $\mathbf{K}$ is partially ordered by volume embedding, it forms a mathematical object of lattice called *conceptual lattice*. Having conceptual lattice, it is possible to identify the relationship between the concepts on the principle of "general-particular". Conceptual lattice nodes are interpreted as a set of potential facts of a certain level (subject), which is related to other facts.

The report contains the results of studies developing the method of conceptual modeling [1] in the direction of increasing the semantic expressiveness of formal contexts and, as a consequence, improving the quality of extracting facts from textual data.

The term "semantic expressiveness" is specified by the following hypothesis. If a conceptual model is used for modeling the meaning of a text, in which the notion of dimension takes place, the higher the dimension of such a model, the deeper it allows to model the meaning of the text. Accordingly, the conceptual model of a higher dimension has a greater semantic expressiveness. To test this hypothesis, not only two - dimensional but also three – and in general, n - dimensional formal contexts are needed.

The report presents the results of experiments on building three- and four-dimensional formal contexts on data generated for standard problems of fact extraction from texts of biomedical subjects of BioNLP Shared Tasks. The data are texts of abstracts of articles from the PubMed system, weakly related to each other, but generally devoted to the study of gene mutations.

Conceptual graphs aggregated by highlighting AMR schemes in them were used to construct formal contexts. The Abstract Meaning Representation (AMR) model is known as the macro model for modeling semantics of texts. Three-dimensional contexts are built with the three-element scheme with pairs of semantic roles such as "agent"and " patient". To build a four-dimensional contexts, the four -element AMR-scheme of "what – what is the impact(verb) – to which- the attribute of an effect» was used.

The experimental results demonstrate the validity of the hypothesis of semantic expressiveness. Four-dimensional formal contexts, through concepts and clusters, allow establishing significantly more links between texts than three-dimensional contexts. Moreover, the computational complexity of the algorithm used to construct three- and four-dimensional contexts does not exceed the known estimates.

This research is funded by RFBR, grant 19-07-01178.

[1] *Bogatyrev Mikhail* Fact Extraction from Natural Language Texts with Conceptual Modeling // Communications in Computer and Information Science, Vol. 706, Springer-Verlag, 2017. — Pp. 89–102.

Автоматическое выделение библиографии в научных текстах

Огальцов Александр Владимирович[1,3]⋆ avogaltsov@edu.hse.ru
Сафин Камиль Фанисович[1,2] kamil.safin@phystech.edu
[1]Россия, Москва, АО «Антиплагиат»
[2]Россия, Москва, МФТИ
[3]Россия, Москва, ВШЭ

Предлагается алгоритм автоматического выделения списка литературы из текстов научных статей. Использование только текстовой информации без форматирования, делает алгоритм более универсальным. Также в методе используются интерпретируемые правила, что делает его более понятным и простым для отладки.

Вводятся предположения относительно рассматриваемых документов. Во-первых, каждая строка рассматриваемого документа должна строго относиться либо к библиографической записи, либо к обычному тексту. Во-вторых, допускается, что тексты могут иметь более одного списка литературы. Некоторые документы, например, сборники тезисов, могут иметь несколько списков литературы по всему тексту.

Предлагаемый алгоритм является композицией двух модулей, каждый из которых выделяет список литературы независимо. Так как некоторые списки литературы имеют определенную структуру, их легко извлечь с помощью набора правил. Для списков литературы, которые не имеют четкой структуры, мы используем признаковую модель. Результатом работы алгоритма является объединение этих двух подходов.

Для анализа качества проводится тестирование на выборке научных документов. Предложенный алгоритм показывает высокое качество на данной тестовой выборке.

Работа выполнена при поддержке Российского фонда фундаментальных исследований (проект № 18-07-01441) и Фонда содействия развитию малых форм предприятий в научно-технической сфере (проект № 44116).

[1] *Огальцов А. В., Бахтеев О. Ю.* Автоматическое извлечение метаданных из научных PDF-документов // Информ. и её примен., Москва, 2018. — С. 75–82.

Automatic bibliography extraction from scientific papers

Aleksandr Ogalstsov[1,3]⋆ avogaltsov@edu.hse.ru
Kamil Safin[1,2] kamil.safin@phystech.edu
[1]Moscow, Russia, Antiplagiat Company
[2]Moscow, Russia, MIPT
[3]Moscow, Russia, HSE

We propose an automatic approach of extraction references from texts of scientific papers. This method uses only text data from paper. It makes this method more universal. Also, it uses simple and understandable heuristics, which makes algorithm more interpretable without affecting the quality of the proposed method.

We make some assumptions about documents considered. Firstly, we assume that each line of document strongly belongs to one of two classes: text or reference line. It means, that there is no line in document, which contains non-reference text and reference item at the same time. This assumption is true for scientific papers. Secondly, we allow documents to have more than one references blocks of literature. The most scientific papers, such as articles, diplomas, reports etc., have only one references section. But some works, for example proceedings, may have multiple bibliographic sections all around the document.

Proposed model consists of two components, each of which extracts bibliography independently. Model architecture is built so, because some reference blocks have determined structure, that can be easily extracted by rule-based methods. For cases with no certain structure we use feature-based method. The final result is a union of these components.

This work was supported by RFBR project 18-07-01441 and FASIE project 44116.

[1] *Ogaltsov A., Bakhteev O.* Automatic metadata extraction from scientific PDF documents // Inform. Primen., Moscow, 2018. — p. 75–82.

К вопросу о математической и программной поддержке в решении задачи атрибуции текстов

Кулаков Кирилл Александрович[1] kulakov@cs.karelia.ru
Рогов Александр Александрович[1] rogov@petrsu.ru
Москин Николай Дмитриевич[1]⋆ moskin@petrsu.ru
[1]Петрозаводск, Петрозаводский государственный университет

Эффективное решение актуальных задач анализа текстовых произведений, включая поиск заимствований и определение авторства, требует совмещения как работы эксперта-лингвиста, так и применения различных математических методов над большими объемами данных.

В центре внимания нашей работы находятся лингвостатистические методы, которые были применены для анализа публицистических статей XIX века, а именно около 500 неатрибутированных текстов из журналов "Время" (1861-1863), "Эпоха" (1864-1865) и еженедельника "Гражданин" (1873-1874). Известно, что Ф. М. Достоевский (вместе со своим братом М. М. Достоевским) редактировал и возглавлял эти журналы, поэтому уже давно ведутся исследования на предмет принадлежности его перу данных произведений.

Рассматривается описание модифицированного программного комплекса "СМАЛТ" для реализации инструментария задачи атрибуции текстов и модульной структуры информационной системы "Фольклор" для автоматизированного построения и анализа теоретико-графовых моделей текстов.

Обнаружена перспективность использования методов "дерево решений" (Decision Tree) и "лес решений" (Random Forest) для задачи атрибуции текстов с использованием ряда лингвостатистических параметров (в том числе параметров Г.Хетсо). Правила в форме "если ..., то ..." могут быть сформулированы на естественном языке и понятны для филологов.

Работа поддержана грантом РФФИ № 18-012-90026.

[1] *Кулаков К. А., Рогов А. А., Москин Н. Д.* Программная поддержка в решении задачи атрибуции текстов // Программная инженерия, Москва: Издательство "Новые технологии", 2019. — Т. 10. № 5. — С. 234–240. http://www.novtex.ru/prin/rus/10.17587/prin.10.234-240.html.

On the question of mathematical and software support in solving the problem of text attribution

Kulakov Kirill[1] kulakov@cs.karelia.ru
Rogov Alexander[1] rogov@petrsu.ru
Moskin Nikolai[1]⋆ moskin@petrsu.ru
[1]Petrozavodsk, Petrozavodsk State University

Effective solution of the actual problems of analysis of textual works, including the search for borrowings and the determination of authorship, requires combining the work of an expert linguist and the use of various mathematical methods on large amounts of data.

The focus of our work is on linguostatistical methods that were used to analyze journalistic articles of the XIX century, namely about 500 unattributed texts from the magazines "Time" (1861-1863), "Epoch" (1864-1865) and the weekly "Citizen" (1873-1874). It is known that F. M. Dostoevsky (together with his brother M. M. Dostoevsky) edited and headed these magazines, so research has long been conducted on the subject of belonging to his pen of these works.

The description of the modified software complex "SMALT" for the implementation of tools for the task of text attribution and the modular structure of the information system "Folklore" for the automated construction and analysis of graph-theoretical models of texts is considered.

The prospects of using "Decision Tree" and "Random Forest" methods for the task of text attribution using a number linguostatistical parameters (including parameters G. Kjetsaa) were discovered. Rules in the form of "if ..., that. .." can be formulated in natural language and understood by philologists.

This research is funded by RFBR, grant 18-012-90026.

[1] *Kulakov K., Rogov A., Moskin N.* Software support in solving the problem of text attribution // Software engineering, Moscow: New technologies Publ., 2019. — Vol. 10. No. 5. — p. 234–240. http://www.novtex.ru/prin/rus/10.17587/prin.10.234-240.html.

Регуляризованные мультимодальные иерархические тематические модели для разведочного поиска документов по документам

Янина Анастасия Олеговна[1]⋆ yanina@phystech.edu
Воронцов Константин Вячеславович[1] vokov@forecsys.ru
[1]Москва, Московский физико-технический институт

Разведочный поиск (exploratory search) — это разновидность информационного поиска, нацеленного на самообразование и приобретение новых знаний пользователем. Это многошаговый процесс, в ходе которого пользователь постоянно корректирует цели и стратегию поиска. Мы рассматриваем элементарный шаг этого поискового сценария, в котором цель поиска выражается запросом в виде длинного текста.

Нами разработан прототип разведочного поиска на основе вероятностного тематического моделирования. Тематическая модель дает векторное представление документа d в виде разреженного дискретного распределения по темам t. Поисковая система ранжирует тематические векторы документов $p(t|d)$ по их сходству с тематическим вектором запроса $p(t|q)$ и представляет пользователю k документов, наиболее близких тексту запроса.

В данной работе используются иерархические тематические модели, рекурсивно разделяющие темы на подтемы. Каждый уровень иерархии представляется плоской тематической моделью, каждая тема которой связывается с одной или (реже) с несколькими темами родительского уровня. Число тем на дочернем уровне, как правило, в несколько раз больше, чем на родительском. Иерархия тем моделирует стратегию поиска, когда человек постепенно фокусирует внимание на подтемах, отбрасывая заведомо нерелевантные темы. Иерархический тематический поиск сначала ранжирует документы по сходству тематических векторов верхнего уровня, затем оставшиеся документы ранжируются по сходству тематических векторов второго уровня, и так далее.

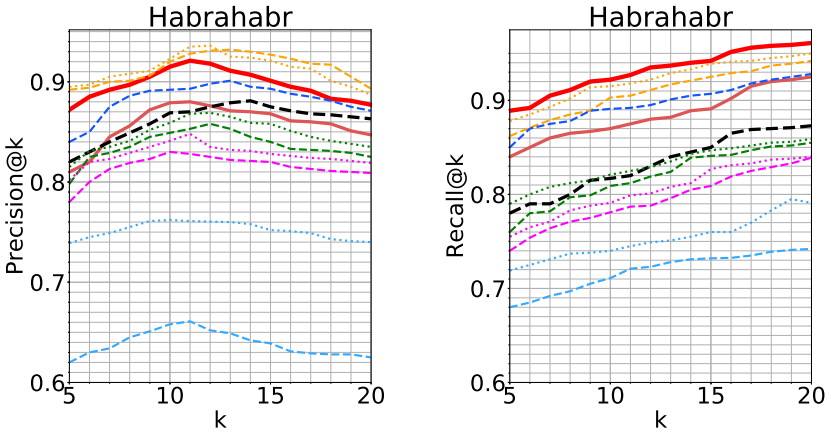
Для построения тематических моделей использовался подход аддитивной регуляризации (ARTM) и библиотека BigARTM с открытым исходным кодом. ARTM максимизирует взвешенную сумму логарифма правдоподобия модели и нескольких критериев регуляризации. В данной работе использовались три регуляризатора: (1) декорреляция распределений терминов в темах обеспечивает попарную различность тем, (2) разреживание распределений тем в документах повышает контрастность сравнения тематических векторов, (3) сглаживание распределений терминов в темах предотвращает вырождение тем. Кроме того, ARTM позволяет описывать тексты, содержащие не только слова, но и термы других модальностей. В данном исследовании в качестве модальностей использовались терминологические словосочетания, теги, категории и авторы документов. Для подбора коэффициентов регуляризации и весов модальностей использовалась покоординатная оптимизация по грубым дискретным сеткам.

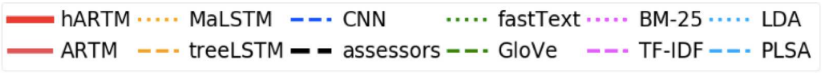
Эксперименты производились на двух коллекциях технических новостей: Habr.com на русском языке и TechCrunch.com на английском языке. Для каждой коллекции мы составили по 100 запросов путем копирования абзацев текста из внешних источников — stackoverflow.com, ixbt.com, и других IT-блогов.

Для измерения качества разведочного поиска использовалась двухэтапная методика оценивания с помощью асессоров. На первом этапе асессору предлагалось найти в заданной коллекции как можно больше документов, релевантных запросу. При этом ему разрешалось пользоваться любыми доступными инструментами поиска. На втором этапе асессор размечал поисковую выдачу, полученную системой тематического поиска по тому же запросу. Каждый запрос обрабатывался тремя асессорами для уменьшения дисперсии результата. Для каждого запроса измерялись стандартные метрики качества поиска Precision@k и Recall@k.

Результаты тематического поиска с плоской и иерархической моделью ARTM сравнивались как с результатами асессоров, так и с результатами других моделей векторного поиска. Для сравнения были взяты простые тематические модели PLSA и LDA, векторные представления текста fastText и Glove, три нейросетевые модели — свёрточная CNN и рекуррентые MaLSTM, TreeLSTM. Результаты сравнения представлены на рисунке.

Точность иерархического тематического поиска оказалась на 7% выше, а полнота на 10% выше по сравнению с асессорами. Плоская тематическая модель показывает лишь сопоставимую точность и +5% полноты по сравнению с асессорами. Для 26 запросов из 100 для коллекции Habr и 29 запросов из 100 для TechCrunch тематический поиск обнаружил документы, которые были пропущены всеми асессорами. При этом асессоры тратили от 15 до 65 минут на выполнение нечётко поставленного творческого поискового задания, в то время как тематический поиск справлялся с ним более качественно и практически мгновенно.





Эксперименты с поочередным отключением модальностей и регуляризаторов показали, что все они важны для повышения качества поиска, за исключением модальности авторов. Простые и широко известные тематические модели PLSA и LDA не выдерживают конкуренции ни с регуляризованными мультимодальными иерархическими моделями, ни с современными нейросетевыми моделями векторных представлений текста.

При подборе числа тем в трёхуровневых иерархических моделях оптимальное число тем на нижнем уровне оказалось в несколько раз выше, чем при использовании плоских моделей: 1400 тем против 200 на Habr и 2800 тем против 475 на TechCrunch. Это можно объяснить тем, что постепенное дробление тем на подтемы приводит к более аккуратной мелко гранулированной модели тем, и, как следствие, к улучшению качества поиска.

Работа поддержана грантом РФФИ № 17-07-01536.

[1] *Ianina A., Vorontsov K.* Regularized multimodal hierarchical topic model for document-by-document exploratory search // The 25th Conference of Open Innovations Association FRUCT. 2019.

Regularized Multimodal Hierarchical Topic Model for Document-by-Document Exploratory Search

Anastasia Ianina[1]⋆ yanina@phystech.edu
Konstantin Vorontsov[1] vokov@forecsys.ru
[1]Moscow, Moscow Institute of Physics and Technology

In the exploratory search paradigm of information retrieval, the user has a complicated search demand that can not be formulated in a short query. The key distinguishing mark of the exploratory search is the absence of the exact query and the unique result. The user collects thematically relevant information iteratively in a "query–browse–refine" process being motivated by learning, understanding, and knowledge acquisition purposes. We consider an elementary step of this scenario in which the search intent can be expressed by a long text query.

For this case, we develop an exploratory search engine based on probabilistic topic modeling. Topic model gives a low-dimensional sparse interpretable vector representation (topical embedding) of a text. The search engine uses these embeddings for ranking documents by their similarity to the query.

Thus, we aim to change the iterative nature of exploratory search and make it a quick one-step procedure. To do this, we use the document-by-document topic-based search. Having a long text query q the system learns its topic vector $p(t|q)$ in the same way as it was done for the documents in the collection. Next, the system ranks document vectors $p(t|d)$ by their similarity to the query and presents top k results to the user.

In this work we are focused on hierarchical topic models with their ability to gradually narrow the scope of the search. A hierarchical topic model divides topics into subtopics recursively. Each level of the hierarchy is represented by a flat topic model. Topical hierarchy emulates a natural human strategy to focus on subtopics gradually discarding unnecessary information. Iterative level-by-level search simulates exploratory search nature with its step-by-step query rephrasing in order to clarify search intent.

We use a top-down level-by-level strategy within the ARTM (Additive Regularization for Topic Modeling) framework. In ARTM a topic model is learned from the collection by maximizing a weighted sum of the log-likelihood and additive regularization criteria. In our experiments we use the combination of three regularizers: (1) decorrelation of term distributions in topics, (2) sparsing topic distributions in documents, and (3) smoothing term distributions in topics. All the topic models were built using open-source library BigARTM, which allows to optimize several quality criteria simultaneously and find low-dimensional topic representations.

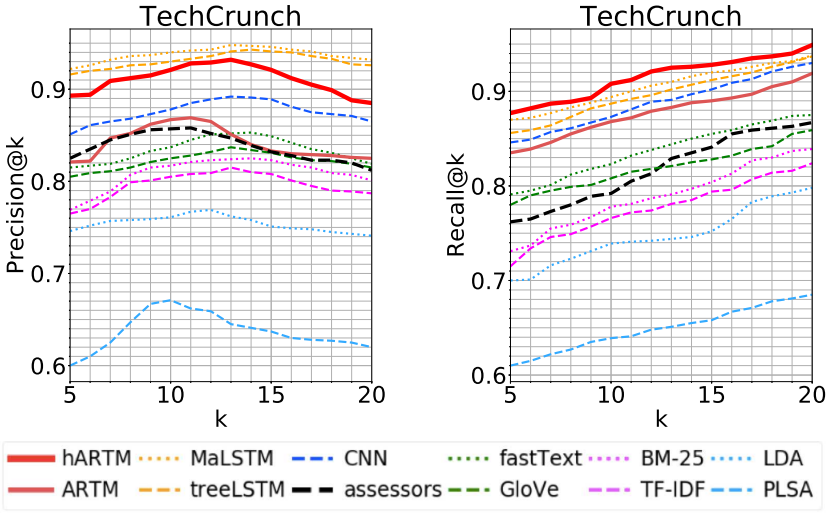
For exploratory search evaluation, we use the evaluation technique based on two-stage human assessments of relevance. First, assessor is asked to find within a given collection as many documents relevant to the query as possible. Assessor may use any search tools available. Second, assessor marks each document retrieved

by the topic-based search for the same query as relevant or irrelevant. Each query is processed by three assessors to reduce the variance of the result. For each query we measure two quality metrics: Precision@k and Recall@k.

We applied the described evaluation method to two tech news collections: Habr.com (in Russian) and TechCrunch.com (in English). For each collection we composed 100 queries by copying text fragments taken from external sources such as stackoverflow.com, ixbt.com, and other IT-oriented blogs. Next we compared precision and recall of the search performed by the assessors with the topic-based search for the best of our models (hierarchical ARTM with 3 levels). On average, precision for hierarchical topic-based search is 7% higher while recall is 10% higher than the same metric for manual human search. The same comparison between flat topic models and assessors' search shows just comparable precision and 5% higher recall.

For 26 queries out of 100 for Habr and for 29 queries out of 100 for TechCrunch our search engine found documents that were missed out by all human annotators. Moreover, it took assessors from 15 to 65 minutes to process a single query while topic-based search gives answer in less than 1 sec. Thus, topic-based exploratory search obtains higher precision and recall and performs significantly faster than humans.

To prove the competitiveness of topic-based search we have compared it with several baselines: other topic models (PLSA, LDA), TF-IDF, BM-25, fastText, Glove, CNN-based approaches, LSTM. All the results for ARTM-based models (both hierarchical and flat) and baselines are shown in the figure.



To find the best hierarchical model we compared two- and three-level models with different number of topics on each level. In the issue, our grid search covered

75 combinations of parameters. The experiment showed that increasing the number of levels in the hierarchy improves the search quality. Interestingly, the optimal number of topics at the lower level of the hierarchy can be much bigger than the number of topics in the flat model: 1400 vs. 200 topic on Habr, and 2800 vs. 475 topic on TechCrunch. This means that the hierarchy allows to produce a larger model with more fine-grained topics.

Tuning the entire set of model hyperparameters, such as the number of levels, the number of topics per level, regularization coefficients, and the set of meta-information modalities gives the topic model highly competitive in the exploratory search task.

This research is funded by RFBR, grant 17-07-01536.

[1] *Ianina A., Vorontsov K.* Regularized multimodal hierarchical topic model for document-by-document exploratory search // The 25th Conference of Open Innovations Association FRUCT. 2019.

Квантильный подход к оцениванию когнитивной сложности текста

Еремеев Максим Алексеевич[1]⋆ maks5507@yandex.ru
Воронцов Константин Вячеславович[1,2] voron@forecsys.ru
[1]Московский государственный университет им. М. В. Ломоносова [2]Московский физико-технический институт (ГУ)

Индексы удобочитаемости или когнитивной сложности текста используются для сравнения учебных текстов, веб-сайтов, деловых и рекламных материалов. Представляется перспективным их применение также в системах разведочного информационного поиска и текстовых рекомендательных системах для ранжирования поисковой выдачи в порядке «от простого к сложному», «от популярного, учебного и обзорного к специализированному и узко профессиональному». Такой принцип ранжирования может быть использован в образовательных платформах и поисково-рекомендательных системах, нацеленных на автоматизацию процесса изучения новых предметных областей пользователем.

Известные индексы удобочитаемости используют простые количественные признаки текста, такие, как средняя длина слов, доля длинных слов, средняя длина предложений, средняя длина подчинённых и сочинённых клауз, и т. д. Реже используются дискурсивные признаки: количество анафорических связей, сложность риторической структуры, и т. д.

Эти методы имеют два основных недостатка.

Во-первых, они не учитывают относительную природу самого понятия сложности. Оценка сложности текста должна зависеть от того, какие тексты мы согласны считать простыми, и для какой читательской аудитории, включая факторы языкового опыта, возраста, образования, профессии.

Во-вторых, они не позволяют учитывать одновременно и единообразно все уровни языка: фонетический, морфологический, синтаксический, дискурсивный.

Мы предлагаем квантильный подход к оцениванию когнитивной сложности текста, свободный от указанных недостатков.

Во-первых, сложность определяется относительно представительного референтного корпуса текстов, которые считаются простыми для выбранной читательской аудитории. В зависимости от целей исследования референтным корпусом может быть электронная библиотека художественной литературы, Википедия, корпус учебной литературы по заданной специальности, тема или подмножество тем из тематической модели мультидисциплинарной текстовой коллекции.

Во-вторых, для каждого уровня языка определяется свой алфавит токенов: для фонетического — фонемы или буквы; для морфологического — морфемы или слоги; для лексического — слова или термины; для синтаксического — типы

и длины синтаксических связей; для дискурсивного — типы и длины риторических структур или предложений.

Предлагаемая математическая формализация понятия сложности основана на следующих представлениях нейрофизиологии и психофизиологии. Восприятие текста или речи проходит через несколько этапов декодирования, приблизительно соответствующих уровням языка. На каждом этапе происходит распознавание и анализ токенов соответствующего уровня. Процессы декодирования происходят в различных зонах нервной системы, от зрительного и слухового анализаторов до коры головного мозга. Каждая зона специализирована на декодировании определённого кода. Завершив декодирование, зона переходит в состояние рефрактерности и некоторое время восстанавливается. Находясь в состоянии рефрактерности, зона не способна декодировать тот же код. Если он снова встретится во входном сигнале, то для декодирования будет задействована другая зона. Перераспределение ресурса может приводить к снижению эффективности анализа сигнала на последующих этапах, реализующих более сложные и эволюционно более молодые функции сознания и мышления. Если частота кода в тексте существенно превышает комфортную (эволюционно обусловленную) частоту, то мозг реципиента не успевает обрабатывать его в обычном режиме и испытывает дополнительную нагрузку.

Таким образом, текст является сложным для восприятия, нагруженным, если он содержит аномально много редких токенов — незнакомых, непонятных или непривычных для реципиента. Это интуитивное определение сложности допускает естественную статистическую формализацию.

Имея представление референтного корпуса текстов в заданном алфавите, мы вычисляем эмпирические распределения частот для каждого токена по референтному корпусу. Теперь допустим, что требуется оценить сложность некоторого текста, не обязательно из референтного корпуса. Токен считается сложным в данном тексте, если он встречается в нём аномально часто по сравнению с референтным корпусом. Аномально высокая частота токена определяется через квантиль эмпирического распределения его частоты. Доля сложных токенов (в процентах) в данном тексте принимается за оценку его сложности на заданном уровне языка относительно заданного референтного корпуса. Таким образом, на всех уровнях языка используется единый квантильный подход, и все оценки сложности измеряются в процентах от длины текста. Это облегчает построение агрегированных оценок сложности по всем уровням.

Для эмпирического сравнения различных оценок когнитивной сложности текста мы подготовили набор пар статей Википедии и разметили его на краудсорсинговой платформе Яндекс.Толока. Сначала была построена тематическая модель Википедии, с помощью которой мы отобрали 10 тысяч пар статей схожей тематики и длины (пример подходящей пары — статьи «Свинец» и «Олово»). Затем эти пары статей предъявлялись асессорам, которых просили поставить одну из четырёх отметок: «первая статья проще второй», «первая статья слож-

нее второй», «статьи примерно одинаковы по сложности», «статьи невозможно сравнить, так как они относятся к разным темам». После отбрасывания несравнимых пар осталось 8 тысяч пар статей.

Квантильные оценки сложности вычислялись по референтному корпусу из 1,5 миллионов русскоязычных статей Википедии. В серии экспериментов предложенные оценки сложности совпадали с асессорскими на 81–84% пар, тогда как известные индексы удобочитаемости ARI и Флеша–Кинкейда давали лишь от 41% до 58% совпадений.

Работа поддержана грантом РФФИ № 17-07-01536.

[1] *M.Eremeev, K.Vorontsov.* Semantic-Based Text Complexity Measure. Recent Advances inNatural Language Processing, RANLP-2019.

Quantile-base approach to measuring cognitive complexity of text

Maksim Eremeev[1]⋆ maks5507@yandex.ru
Konstantin Vorontsov[1,2] vokov@forecsys.ru
[1]Moscow State University
[2]Moscow Institute of Physics and Technology

The indexes of readability or cognitive complexity of the text are used to compare educational texts, websites, business and promotional materials. It seems promising to use them also in information retrieval and text recommendation systems for ranking search results in the order "from simple to complex", "from popular, educational, and surveys to highly specialized". This principle of ranking can be used in educational platforms, personalized exploratory search and recommendation systems aimed at automating the process of studying new subject areas by the user.

The well-known readability indices use simple quantitative features of the text, such as the average word length, the frequency of long words, the average sentence length, the average length of subordinates or composed clauses, etc. Discursive features are less commonly used: the number of anaphoric connections, the complexity of the rhetorical structure, etc.

All these estimations have two main disadvantages.

Firstly, they do not take into account the relative nature of the complexity concept itself. The complexity of the text should depend on which texts can be considered as simple, and for which readership, including language experience, age, education, and profession factors.

Secondly, they do not unify all levels of the language: phonetic, morphological, syntactic, and discursive.

We propose a quantile-based approach to the cognitive complexity of a text, free of these shortcomings.

Firstly, the complexity is determined with respect to a representative reference corpus of texts that we consider as simple for the implied readership. Depending on the objectives of the study, the reference corpus can be an electronic library of fiction, Wikipedia, educational literature for a given specialty, a topic or a subset of topics from the topic model of a multidisciplinary text collection.

Secondly, for each level of the language, its own alphabet of tokens is determined. These are: phonemes or letters for the phonetic level; morphemes or syllables for the morphological one; words or terms for the lexical one; types and lengths of syntactic links, rhetorical structures, or sentences for the discursive level.

Our mathematical formalism is based on the following ideas from neurophysiology and psychophysiology. The perception of text or speech goes through several stages of decoding, approximately corresponding to the levels of the language. At each stage, the tokens of the corresponding level are recognized and analyzed. Decoding processe takes place in various areas of the nervous system, from the visual and auditory analyzers to the cerebral cortex. Each zone is specialized in decoding a

specific code. Having completed the decoding, the zone goes into a refractoriness state and is restored for some time. In a refractory state, the zone is not able to decode the same code. If it occurs again in the input sequence, then another zone will be used for decoding. If the frequency of the code in the text significantly exceeds the comfortable (evolutionarily determined) frequency, then the recipient's brain is subjected to additional stress being unable to process it in the usual mode.

Thus, the text is complicated if it contains abnormally many rare tokens, which are unfamiliar, incomprehensible or unusual for the recipient. This intuitive definition of complexity allows for a natural statistical formalization.

Having a representation of the reference corpus of texts in a given alphabet, we calculate the empirical frequency distribution for each token over the reference corpus. Let us have to measure the complexity of a text, not necessarily from the reference corpus. A token is considered complex in the text if it appears abnormally often in the text, if compared to the reference corpus. The abnormally high frequency of the token is determined through the quantile of the empirical distribution of its frequency. We take the percentage of complex tokens in the text as a complexity measure of this text at a given language level with respect to a given reference corpus. Thus, we use a unified quantile-based approach at all levels of the language, and all level-wise complexities are measured as a percentage of the text length. This makes it easy to build aggregate complexity measures for all levels.

For the empirical comparison of complexity measures of the text, we prepared a set of pairs of Wikipedia articles and marked it up on the crowdsourcing platform Yandex.Toloka. First, we built a topic model of Wikipedia, from which we selected 10 thousand pairs of articles of similar topic distributions and lengths (an example of a suitable pair is the articles "Lead" and "Tin"). Then annotators labeled these pairs by four marks: "the first article is simpler than the second one", "the second article is simpler than the first one", "the articles are approximately the same in complexity", "the articles cannot be compared because they relate to different topics". After discarding incomparable pairs, 8 thousand of pairs of articles remained.

Complexity measures were calculated from the reference corpus of 1.5 million Russian-language Wikipedia articles. In a series of experiments, our complexity measures were agreed with assessors in 81–84% pairs, whereas the well-known readability indices ARI and Flash-Kincaid yielded only from 41% to 58% matches.

The research is funded by RFBR, grant 17-07-01536.

[1] *M.Eremeev, K.Vorontsov.* Semantic-Based Text Complexity Measure. Recent Advances inNatural Language Processing, RANLP-2019.

Когнитивные образы для визуального анализа состояний сложных объектов применительно к космической отрасли

Емельянова Юлия Геннадиевна[1]⋆ yuliya.emelyanowa2015@yandex.ru
Хачумов Вячеслав Михайлович[1] vmh48@mail.ru
[1]Переславль-Залесский, ИПС им. А.К. Айламазяна РАН

Работа направлена на поддержку принятия решений операторов наземных станций, осуществляющих контроль и диагностику подсистем космических аппаратов. Решается задача автоматического построения цветояркостных когнитивных графических образов, способствующих оперативному пониманию сложившейся ситуации. Впервые дана комплексная формальная оценка качества образов, предназначенных для наблюдения за состоянием сложной динамической системы реального времени.

Разработан подход к построению когнитивных образов для отображения и распознавания радиотехнических сигналов. Предлагается использовать интегральный контурный метод, дополненный монохромной и цветовой компонентами.

Разработан метод когнитивной визуализации состояний датчиков ориентации в условиях помех. Поведение предложенного 3D-образа, подключенного ко всем датчикам ориентации, позволяет судить о наличии сбоев в их показаниях.

Разработаны универсальные иерархические графические образы на основе циклоид, обеспечивающие представление динамической ситуации в состояниях сложных многопараметрических объектов. Обеспечивается одновременное отображение нескольких состояний параметров и объектов контролируемой системы.

Построены новые двухуровневые когнитивные образы в составе интеллектуальных интерфейсов наземных станций, существенно повышающие скорость интерпретации данных телеметрии, антенных систем и датчиков жизнеобеспечения. Полученные результаты внедрены в АО РКС и НИИ КС им. А.А. Максимова.

Работа поддержана грантами РФФИ № 18-37-00037-мол_а и № 18-07-00014-а.

Cognitive images for visual analysis of the complex objects states in relation to the space industry

Emelyanova Yulia[1]⋆ yuliya.emelyanowa2015@yandex.ru
Khachumov Vyacheslav[1] vmh48@mail.ru
[1]Pereslavl-Zalessky, Ailamazyan Program Systems Institute of RAS

The work is aimed at supporting the decision-making of earth-based station operators who monitor and diagnose spacecraft subsystems. The problem of automatic construction of colour-luminance cognitive-graphical images, which contribute to operational understanding of the current situation, is solved. For the first time, a complex formal estimation of the images quality designed to monitor the complex dynamic real-time system state has been given.

An approach to the cognitive images construction for the display and recognition of radiotechnical signals has been developed. It is proposed to use an integral contour method supplemented by monochrome and color components.

A cognitive visualization method of orientation sensor states under interference conditions has been developed. The behavior of the proposed 3D image, connected to all orientation sensors, allows you to judge whether there are failures in their readings.

Universal cycloid-based hierarchical images have been developed to represent the dynamic situation in the complex multi-variable objects states. Simultaneous display of parameters and objects several states of the monitored system is provided.

New two-level cognitive images have been constructed as part of the intelligent interfaces of earth-based stations, which significantly increase the interpretation speed of telemetry data, antenna systems and life support sensors. The obtained results were implemented in JSC RSS and A.A. Maksimov Space Systems Research Institute.

The work was performed with financial support of RFFP project No. 18-37-00037-mol_a and No. 18-07-00014-a.

Обнаружение аномальных явлений в работе службы заказа такси на базе интеллектуального анализа данных

Андриянов Никита Андреевич[1,2]⋆ nikita-and-nov@mail.ru

[1]Ульяновск, Ульяновский государственный технический университет

[2]Ульяновск, Ульяновский институт гражданской авиации имени Главного маршала авиации Б.П. Бугаева

Работа посвящена задаче обнаружения аномалий на фоне данных службы заказа такси. Такие данные описываются моделями случайных процессов. Показано, что адекватного представления данных такси можно добиться с помощью авторегрессионных дважды стохастических моделей. Кроме того, описание с прогнозированием на длительную перспективу может быть получено при переходе к случайным полям. Такой подход позволяет выстроить четкую организацию данных о работе службы заказа такси. Дважды стохастические модели позволяют модифицировать и применить алгоритмы обнаружения детерминированных аномалий к данным, описываемым с помощью такой модели. В работе службы заказа такси такими аномалиями могут быть совершенно непредвиденные спады заказов. Еще аномальные значения может принимать стоимость поездок в том случае, когда она каждый раз пересчитывается автоматически специализированной программой с учетом множества факторов. Выявление подобного рода аномалий в режиме реального времени позволит своевременно корректировать работу службы заказа такси. Следует отметить, что могут быть также применены алгоритмы прогнозирования вероятного числа заказов в конкретный период времени. Качественное решение задачи прогнозирования позволит рассчитать необходимое число операторов приема вызовов службы заказа такси и/или число водителей, обеспечивающих перевозку. Таким образом, выполнены исследования по интеллектуальному анализу данных работы службы заказа такси [1].

Работа поддержана грантом РФФИ № 18-31-000056.

[1] *Andriyanov N. A., Sonin V. A.* Using mathematical modeling of time series for forecasting taxi service orders amount // CEUR Workshop Proceedings, Volume 2258, 2018. — P. 462–472.

Detection of anomalous phenomena in the work of the taxi ordering services on the basis of data mining

Nikita Andriyanov[1,2]⋆ nikita-and-nov@mail.ru

[1]Ulyanovsk, Russia, Ulyanovsk State Technical University

[2]Ulyanovsk, Russia, Ulyanovsk Civil Aviation Institute

The work is devoted to the development and research of an anomaly detection algorithm against the background of taxi order service data. Moreover, it is proposed to use mathematical models of random sequences to describe such data. The analysis showed that an adequate representation of taxi service data can be achieved using autoregressive doubly stochastic models. In addition, a description with long-term forecasting can be obtained by using the mathematical models of random fields. This approach allows to build a clear organization of data on the work of the taxi order service with reference to weekly peaks at weekends or peak loads during special hours. The apparatus for the simplest doubly stochastic models makes it quite simple to modify and apply algorithms for detecting deterministic anomalies to the data described using such a model. Algorithms for detecting non-deterministic anomalies are also investigated. In the work of a taxi order service, such anomalies can be completely unforeseen drops in orders on holidays or weekends. The cost of trips can also be considered anomalous when it is automatically recalculated each time by a specialized program taking into account many factors. Identification of such anomalies in real time will allow to timely adjust the work of the taxi order service. Algorithms for forecsting the likely number of orders in a particular period of time can also be applied. A high-quality solution to the forecasting problem will allow us to calculate the required number of call pick-up operators of a taxi order service and / or the number of drivers providing transportation. Thus, studies have been carried out on the intellectual analysis of data from the work of a taxi order service [1].

This research is funded by RFBR, grant 18-31-00056.

[1] *Andriyanov N. A., Sonin V. A.* Using mathematical modeling of time series for forecasting taxi service orders amount // CEUR Workshop Proceedings, Volume 2258, 2018. — P. 462–472.

Метод обнаружения нештатных состояний технологических процессов

Сычугов Алексей Алексеевич* xru2003@list.ru
Анчишкин Александр Павлович alexanderanchishkin@yandex.ru
Тула, Тульский государственный университет

Технологические процессы (ТП) промышленных объектов, в первую очередь, химической, металлургической и обрабатывающей промышленности представляют собой динамические системы, функционирующие в реальном времени, описание работы которых требует большого числа контролируемых переменных, формируемых разнородными территориально-распределенными датчиками.

Технологический процесс, может находиться в одном из состояний:

$$\Omega = \{\omega_1, \omega_2, ..., \omega_m\}. \tag{1}$$

Требуется по ограниченной последовательности наблюдений, формируемых показаниями датчиков, известному текущему состоянию ω_k и управляющим воздействиям определить наиболее вероятное значение $\hat{\omega}_{k+1}$.

Функционирование технологического процесса описывается моделью:

$$\omega_{k+1} = J(\omega_k), u_k = B(\omega_k, L(w_k)), y_k = H(\omega_k), ... \rightarrow ....., \tag{2}$$

где J – оператор перехода из состояния ω_k в состояние ω_{k+1}, u_k — управляющее воздействие, B — закон управления ТП с учетом статистических характеристик случайных воздействий внешней среды, H – матрица, отображающая состояние ω_k в вектор измерений y_k.

Предложен и исследован следующий метод обнаружения аномалий.

1. Все возможные состояния технологического процесса упорядочиваются по степени опасности. При этом выделяются граничные состояния ω_j^*,
2. Накапливается массив наблюдений Y_k,
3. На каждом шаге функционирования по полученному набору значений ω_k, Y_k, u_k определяется оценка $\hat{\omega}_{k+1}$,
4. Полученное на предыдущем шаге значение сравнивается со значениями наиболее близких граничных состояний ω_j^*. Если результат сравнения меньше заданного допустимого значения, то выдается сообщение об обнаружении аномалии. При этом сохраняется набор значений ω_k, Y_k, u_k,
5. Процесс продолжается начиная с п. 2.

Появлению аномальных состояний технологического процесса всегда предшествует появление отклонений от штатного функционирования, то есть нештатные состояния. Множество состояний (1) разделяется на множество штатных $\Omega^+ \subset \Omega$ и нештатных $\Omega^- \subset \Omega$. Учитывая, что варианты возможных

отклонений от штатного функционирования могут быть различны и не определены, множество Ω^- заранее неизвестно, в то время как Ω^+ детерминировано. Тогда возникает задача обнаружения нештатных состояний технологического процесса как начальной стадии появления аномального состояния. Предложен метод решения данной задачи, основанный на следующих рассуждениях.

Технологический процесс представлен в виде множества операций $A = \{a_1, a_2, ..., a_n\}$, каждая из которых описывается множеством переменных X^{a_i}, которое разделяется на контролируемые $\tilde{X}^{a_i} = \left\{\tilde{x_1}^{a_i}, \tilde{x_2}^{a_i}, ..., \tilde{x_p}^{a_i}\right\}$, используемые для функционирования системы управления или анализа состояния, и неконтролируемые $\hat{X}^{a_i} = \left\{\hat{x_1}^{a_i}, \hat{x_2}^{a_i}, ..., \hat{x_p}^{a_i}\right\}$, то есть определяемые при проектировании технологического процесса и не изменяющиеся в процессе его работы. Множества контролируемых и неконтролируемых переменных не пересекаются. Контролируемые переменные являются временными рядами.

Для современных производств характерно множество особенностей. Режим функционирования в реальном времени характеризуется быстрой сменой значений большого числа контролируемых переменных. Значения контролируемых переменных в некоторый момент времени определяют состояние технологического процесса в этот момент времени, то есть фактически представляет собой точку в многомерном пространстве. Не все контролируемые переменные имеют одинаковую значимость.

Учитывая детерминированность множества Ω^+ и неопределенность множества Ω^-, задача обнаружения нештатных состояний технологического процесса рассматривается как отделение штатных состояний от всех остальных и сводится к задаче классификации в бесконечных потоках данных с большим количеством параметров, представляющих собой временные ряды с изменяющимися свойствами. В теории машинного обучения данная задача известна как задача одноклассовой классификации для решения которой разработаны различные подходы. В данной работе использован одноклассовый метод опорных векторов.

Учитывая, что не все элементы множества $\tilde{X}^A$ имеют одинаковую значимость, для ускорения анализа текущего состояния технологического процесса возникает задача отбора переменных. Применение классических методов отбора (фильтрация, обертка, встроенные методы) в ходе процедуры обучения в данном случае невозможно.

Для оценки значимости переменных используется вероятностная модель.

Проведенные экспериментальные исследования на модельных и реальных данных показали применимость и адекватность предложенной модели.

Работа поддержана грантом РФФИ № 19-07-01107.

[1] *Сычугов А. А., Французова Ю. В., Анчишкин А. П.* A method for analyzing the condition of technological processes // Asia Life Sciences, Philippines: ©Rushing Water Publishers Ltd., 2019. — С. 241–250.

Method of Abnormal States of Technological Processes Detection

Alexander Sychugov⋆ xru2003@list.ru
Alexander Anchishkin alexanderanchishkin@yandex.ru
Tula, Tula State University

Technological processes (TP) of industrial facilities, primarily chemical, metallurgical and manufacturing industries are dynamic systems that operate in real time, the description of which requires a large number of controlled variables formed by heterogeneous geographically distributed sensors.

The technological process can be in one of the following states:

$$\Omega = \{\omega_1, \omega_2, ..., \omega_m\} \tag{1}$$

It is required to determine the most probable value by a limited sequence of observations formed by sensor readings, the known current state and control actions.

The functioning of the technological process is described by the model:

$$\omega_{k+1} = J(\omega_k), u_k = B(\omega_k, L(w_k)), y_k = H(\omega_k), ... \rightarrow ..... \tag{2}$$

where J – is the operator of state transition from ω_k state into ω_{k+1} state, u_k is control action, B is TP control law taking into account the statistical characteristics of random environmental influences $L(\omega_k)$, H is the matrix reflecting the condition of ω_k into the measurement vector y_k.

The following anomaly detection method is proposed and investigated.

1. All possible states of the technological process are ordered according to the degree of danger. The boundary ω_j^* states are highlighted,
2. The array of observations Y_k is accumulated,
3. At each step of the operation on the resulting set of values ω_k, Y_k, u_k the assessment $\hat{\omega}_{k+1}$ is defined,
4. The value obtained in the previous step is compared with the values of the closest boundary states ω_j^*. If the result of the comparison is less than the specified allowable value, an anomaly detection message is displayed. Here the set of values is saved ω_k, Y_k, u_k,
5. The process continues starting from Point 2.

The appearance of abnormal states of the technological process is always preceded by the appearance of deviations from the normal operation, that is, abnormal states. The set of states (1) is divided into the set of normal states $\Omega^+ \subset \Omega$ and abnormal $\Omega^- \subset \Omega$. Given that the options for possible deviations from the normal functioning may be different and are not defined, the set of is unknown in advance, whereas Ω^+ is determined. Then there is a problem of detection of abnormal states of technological process as an initial stage of occurrence of an abnormal state. A method for solving this problem based on the following reasoning is proposed.

The technological process is presented in the form of a set of operations $A = \{a_1, a_2, ..., a_n\}$, each of them is described by the set of variables X^{a_i}, which is divided into controlled $\tilde{X}^{a_i} = \left\{\tilde{x_1}^{a_i}, \tilde{x_2}^{a_i}, ..., \tilde{x_p}^{a_i}\right\}$, used for the operation of the control system or analysis of the state, and not controlled $\hat{X}^{a_i} = \left\{\hat{x_1}^{a_i}, \hat{x_2}^{a_i}, ..., \hat{x_p}^{a_i}\right\}$, that is defined at designing of technological process and not changing in the course of its work. Sets of controlled and uncontrolled variables do not intersect. Controlled variables are time series.

For modern productions the following features are characteristic. The mode of operation in real time is characterized by rapid change of values of a large number of controlled variables. The values of the controlled variables at some point in time determine the state of the process at this point in time, that is, in fact, is a point in multidimensional space. Not all controlled variables have the same significance.

Given the determinism of the set Ω^+ and uncertainty of the set Ω^-, the problem of detection of abnormal states of the technological process is considered as a separation of normal states from all others and is reduced to the problem of classification in infinite data streams with a large number of parameters representing time series with changing properties. In machine learning theory this problem is known as the one class classification problem for which different approaches have been developed. In this paper, the one-class support vector method is used.

Given that not all elements of the set $\tilde{X}^A$ have the same importance, to accelerate the analysis of the current state of the process there is a problem of selection of variables. The application of classical methods of selection (filtering, wrapping, built-in methods) during the training procedure in this case is impossible.

A probabilistic model is used to assess the significance of variables.

Experimental studies on model and real data have shown the applicability and adequacy of the proposed model.

This research is funded by RFBR, grant 19-07-01107.

[1] *Sychugov A., Frantsuzova Y, Anchishkin A.* A method for analyzing the condition of technological processes // Asia Life Sciences, Philippines: ©Rushing Water Publishers Ltd., 2019. — p. 241–250.

Анализ объема церебральных структур пациентов с гипоксически-ишемической энцефалопатией

Ерохин Михаил Владимирович[1] erokhin_m_v@mail.ru
Плоткин Артем Владимирович[2]⋆ avplotkin@gmail.com
[1]Санкт-Петербург, ИМЧ РАН
[2]Санкт-Петербург, СПбГУ

Метод МР-морфометрии позволяет получить количественные показатели объемов большинства структур головного мозга. В работе приводится анализ собранных данных МР-морфометрии в применении к ранней диагностики гипоксически-ишемической энцефалопатии (ГИЭ) у детей раннего возраста (до 3-х лет) [1].

Первым этапом работы являлся сбор медицинских данных. После отбора историй болезни и МР-томограмм, с помощью программы `Freesurfer` были получены морфометрические данные двух групп пациентов: с признаками ГИЭ (43 человека) и без нее (50 человек).

В результате были получены численные показатели 71 признака для каждого из пациентов. После проведение предобработки полученных данных, с использованием подхода [2] был построен линейный классификатор с точностью 77.1% на кросс-валидации.

В связи с трудоемкостью сбора данных, далее была рассмотрена задача выделения наиболее важных признаков. С использованием комбинированного подхода из исходного списка были выделены 16 признаков. Было показано, что дальнейшее сокращение количества признаков невозможно без нарушения линейной отделимости множеств.

Отобранные признаки соответствуют современным представлениям о природе поражений при ГИЭ — уменьшение объемов субкортикального серого вещества, коры, истончение мозолистого тела и некоторых других, уязвимых для гипоксии структур. Одним из отобранных признаков оказался пол пациента. Данная зависимость не описана в литературе и требует дальнейшего изучения.

По выделенным признакам был построен новый линейный классификатор, продемонстрировавший на кросс-валидации точность в 82%.

[1] *Ерохин М. В., Тащилкин А. И.* О возможностях МР-морфометрии при диагностике гипоксически-ишемических поражений головного мозга у детей раннего возраста // Детская медицина Северо-Запада, 2018, Т. 7, №1. — С. 116–116.

[2] *Малоземов В. Н., Плоткин А. В.* Строгое полиномиальное отделение двух множеств // Вестник Санкт-Петербургского университета. Математика. Механика. Астрономия. 2019. Т. 6 (64). Вып. 2. — С. 232–240.

Analysis of cerebral structures volume of patients with hypoxic-ischemic encephalopathy

Erokhin Mikhail[1]⋆ erokhin_m_v@mail.ru
Plotkin Artem[2] avplotkin@gmail.com
[1]Saint Petersburg, IHB RAS
[2]Saint Petersburg, SPbSU

The method of voxel-based morphometry allows getting quantitative values of the volume of most brain structures. The paper presents an analysis of the collected voxel-based morphometry data applied to the early diagnosis of hypoxic-ischemic encephalopathy (HIE) of young children (up to 3 years old) [1].

The first step was to collect medical data. After medical history and MRI results analysis, the morphometric data of two groups of patients were obtained using `Freesurfer` program: with signs of HIE (43 people) and without it (50 people).

As a result, numerical values of 71 features were obtained for each of the patients. After pre-processing the data, a linear classifier was built using [2] approach with an accuracy of 77.1% on cross-validation.

Due to the time cost of data collection, the task of identifying the most important features was further considered. Using a combined approach, 16 features were identified from the initial list. It was shown that a further reduction is impossible without violating the linear separability of the sets.

The selected features correspond to modern ideas about the nature of lesions in HIE — subcortical and cortical grey matter volume loss, thinning of corpus callosum and some other changes in susceptible to injury areas. One of the selected features was the patient's sex. Such relation is not yet described in literature and requires further investigation.

Based on the selected features, a new linear classifier was built, which demonstrated cross-validation accuracy of 82%.

[1] *Erokhin M., Tashilkin A.* Voxel-based morphometry capabilities in diagnosing hypoxic-ischemic encephalopathy in infants // Children's Medicine of the North-West, 2018, Vol. 7, issue 1. — p. 116–116.

[2] *Malozemov V., Plotkin A.* Strict polynomial separation of two sets. // Vestnik of Saint Petersburg University. Mathematics. Mechanics. Astronomy, 2019, vol. 6 (64), issue 2. — p. 232–240.

Множественное выравнивание геномов на основе спектрально-аналитического подхода

Панкратов Антон Николаевич[1]⋆ pan@impb.ru

[1]Пущино, ИМПБ РАН - филиал ИПМ им.М.В.Келдыша РАН

Выравнивание полных геномов организмов одного или разных видов является критически важной задачей в связи большим объемом данных по секвенированию и аннотации геномов. В настоящей работе рассматривается проблема построения вычислительной системы сравнения геномов, призванной решать как эволюционные, так и популяционные задачи.

Сравнение геномов базируется на определении всевозможных неточных повторов в геномах, с учетом перестановок и ориентации этих повторов, а также масштаба, на котором они возникают, и последующего анализа связности между семействами повторов, отвечающей гомологии или ортологии. Сравнение геномов является квадратичной по сложности задачей в зависимости от суммарной длины сравниваемых геномов, что является серьезным вызовом для разрабатываемых методов, алгоритмов и программ.

Развиваемые методы анализа последовательностей и структур основаны как на классическом динамическом программировании [1], так и на спектральном сравнении объектов на подходящем масштабе [2]. При этом показано, что задача множественного выравнивания или кластеризации семейств повторов решается на этапе спектрально-аналитического метода, поскольку он основан на пороговом решающем правиле и находит все кандидаты на повторы, а метод динамического программирования является оптимизирующим и с помощью него находятся наилучшие повторы или верифицируются найденные повторы.

[1] *Pankratov A. N., Tetuev R. K., Pyatkov M. I.* LSCGAT: Long Sequences Customizable Global Alignment Tool // Journal of Bioinformatics and Genomics, No 1 (10) 2019

[2] *Панкратов А. Н., Пятков М. И., Тетуев Р. К., Назипова Н. Н., Дедус Ф. Ф.* Поиск протяженных повторов в геномах на основе спектрально-аналитического метода // Математическая биология и биоинформатика, 2012, Т.7, N.2, с.476–492

Multiple genome alignment based on a spectral-analytical approach

Anton Pankratov[1]⋆ pan@impb.ru

[1]Pushchino, IMPB RAS - Branch of KIAM RAS

Alignment of the whole genomes of organisms of one or different species is a critical task due to the large amount of data on sequencing and annotation of genomes. In this paper, we consider the problem of constructing a computer system for comparing genomes, designed to solve both evolutionary and population problems.

Comparison of genomes is based on the determination of all kinds of inexact repeats in the genomes, taking into account the transpositions and orientations of these repeats, as well as the scale at which they arise, and the subsequent analysis of the connectivity between families of repeats corresponding to homology or orthology. Comparing genomes is a quadratic task of complexity depending on the total length of the compared genomes, which is a challenge for the developed methods, algorithms, and programs.

Developed methods for the analysis of sequences and structures are based both on classical dynamic programming [1], and on spectral comparison of objects on a suitable scale [2]. It has been shown that the problem of multiple alignment or clustering of repeat families is solved at the stage of the spectral-analytical method, since it is based on a threshold decision rule and finds all candidates for repeats, while the dynamic programming method is optimizing and finds the best repeats or verifies the found ones.

[1] *Pankratov A. N., Tetuev R. K., Pyatkov M. I.* LSCGAT: Long Sequences Customizable Global Alignment Tool // Journal of Bioinformatics and Genomics, No 1 (10) 2019

[2] *Pankratov A. N., Pyatkov M. I., Tetuev R. K., Nazipova N. N., Dedus F. F.* Search for Extended Repeats in Genomes Based on the Spectral-Analytical Method //Math. Biol. Bioinf., 2012, V.7, N.2, pp.476–492

Применение метода функциональной томографии к экспериментальным данным электрической активности головного мозга при психических расстройствах

Панкратова Наталья Михайловна[1]⋆ natpan1974@mail.ru
Рыкунов Станислав Дмитриевич[1] stanislavrykunov@gmail.com
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

Современные методы исследования электрической активности головного мозга, такие как электроэнцефалография (ЭЭГ) и магнитная энцефалография (МЭГ), позволяют определять и локализовать искомую активность, в том числе и патологическую, с высокой точностью. Рассматривается вопрос изменения спектральных и пространственных характеристик ЭЭГ и МЭГ при психических расстройствах [1]. Представленный обзор литературы показывает, что электрическая активность головного мозга в этих случаях имеет как спектральные особенности, так и особенности локализации источников патологической активности в разных частотных диапазонах. Вопрос о пространственном расположении источников патологической активности является ключевым при изучении работы мозга и решается с помощью различных методов локализации. Предложен метод для точного количественного анализа активности по данным энцефалографии. Результаты локализации отображаются на магнитно-резонансной томограмме субъекта, на основе чего строятся гипотезы о нейрофизиологическом механизме изучаемой патологии. Метод, предложенный в данной работе, опирается на преобразование Фурье многоканальных данных энцефалографии и локализацию отдельных спектральных компонент. Это позволяет детально изучать те или иные частотные признаки патологической активности мозга и отвечать на вопросы об их связи с анатомией мозга. Ритмическая активность головного мозга при психических расстройствах отличается от нормальной в нескольких частотных диапазонах, но чёткой картины и понимания происходящих изменений, которое можно бы было использовать в диагностических целях, пока нет. В работе анализируются данные энцефалографии пациентов с синдромом дефицита внимания и гиперактивности (СДВГ). Использовались экспериментальные данные МЭГ из банка данных с открытым доступом Open MEG Archive (OMEGA) [2]. Об актуальности такого анализа говорят литературные источники, в которых говорится, с одной стороны, о наблюдаемых спектральных изменениях на энцефалограммах при СДВГ, а с другой стороны, об отсутствии какой-либо системы диагностических признаков данной патологии. Кроме того, необходимо отличать СДВГ от аутизма, который имеет похожие симптомы.

Работа выполнена при поддержке Программы Президиума РАН №2 «Механизмы обеспечения отказоустойчивости современных высокопроизводительных

и высоконадежных вычислений» и грантов РФФИ 17-29-02178, 18-00-00619, 19-07-00964, 17-07-00677, 17-07-00686.

[1] *Панкратова Н. М., Рыкунов С. Д., Бойко А. И., Молчанова Д. А., Устинин М. Н.* Локализация спектральных особенностей энцефалограмм при психических расстройствах // Математическая биология и биоинформатика, 2018. Т. 13. № 2. С. 322–336.

[2] Open MEG Archive, `https://www.mcgill.ca/bic/resources/omega`

Application of the functional tomography method to experimental data on the electrical activity of the brain in mental disorders

Natalia Pankratova[1]⋆ natpan1974@mail.ru
Stanislav Rykunov[1] stanislavrykunov@gmail.com
Anna Boyko[1] a.boyko@list.ru
Mikhail Ustinin[1] u_m_n@mail.ru

[1]Pushchino, IMPB RAS- Branch of KIAM RAS

Modern methods for studying the electrical activity of the brain, such as electroencephalography (EEG) and magnetic encephalography (MEG), allow you to determine and localize the desired activity, including pathological, with high accuracy. The question of changes in the spectral and spatial characteristics of the EEG and MEG in mental disorders is considered [1]. The presented review of the literature shows that the electrical activity of the brain in these cases has both spectral features and features of localization of pathological activity sources in different frequency ranges. Spatial position of the sources of pathological activity is a key issue of brain studies and it is solved by various localization methods. A method is proposed for an accurate quantitative analysis of activity from encephalography data. Localization results are displayed on the magnetic resonance tomogram of the subject, on the basis of which hypotheses about the neurophysiological mechanism of the pathology under study are constructed. The method, proposed in this paper, is based on Fourier transform of multichannel encephalography data and on the localization of spectral components. Such approach permits to study in detail some or other frequency features of the brain pathological activity and to reveal their connections with the brain anatomy. The rhythmic activity of the brain in mental disorders differs from normal in several frequency ranges, but there is no clear picture and understanding of the changes that could be used for diagnostic purposes. The paper analyzes the patient data of encephalography with attention deficit hyperactivity disorder (ADHD). The experimental MEG data from the Open MEG Archive (OMEGA) open access database were used [2]. The relevance of such an analysis is indicated by literary sources, which say, on the one hand, the observed spectral changes in the encephalograms in ADHD, and on the other hand, the absence of any system of diagnostic signs of this pathology. In addition, it is necessary to distinguish ADHD from autism, which has similar symptoms.

This research is funded by Programs of the Presidium of the Russian Academy of Sciences No.2 ≪Mechanisms for ensuring fault tolerance of modern high-performance and highly reliable computing≫ and RFBR, grants 17-29-02178, 18-00-00619, 19-07-00964, 17-07-00677, 17-07-00686.

[1] *Pankratova N. M., Rykunov S. D., Boyko A. I., Molchanova D. A., Ustinin M. N.* Localization of Encephalogram Spectral Features in Psychic Disorders // Mathematical Biology and Bioinformatics, 2018;13(2):322-336.

[2] Open MEG Archive, https://www.mcgill.ca/bic/resources/omega

Исследование магнитных энцефалограмм пациентов с синдромом дефицита внимания и гиперактивности методом виртуальных электродов

Рыкунов Станислав Дмитриевич[1]⋆ rykunov@impb.ru
Устинин Михаил Николаевич[1] u_m_n@mail.ru
Бойко Анна Ивановна[1] a.boyko@list.ru
Панкратова Наталья Михайловна[1] natpan1974@mail.ru
[1]Пущино, ИМПБ РАН - филиал ИПМ им. М.В. Келдыша РАН

Был разработан новый метод анализа данных магнитной энцефалографии – метод виртуальных электродов. По данным магнитной энцефалографии строится функциональная томограмма — пространственное распределение источников поля на дискретной сетке. Функциональная томограмма отображает на пространство головы информацию, содержащуюся в многоканальных временных рядах энцефалограммы. Это достигается решением обратной задачи для всех элементарных осцилляций, выделяемых с помощью преобразования Фурье. Каждой частоте осцилляции соответствует узел трехмерной сетки, в котором располагается источник. Пользователь задает местоположение, размер и форму области мозга для детального изучения частотной структуры функциональной томограммы – виртуальный электрод. Совокупность осцилляций, попавших в заданную область, представляет собой парциальный спектр данной области. По данному спектру восстанавливаются временные ряды энцефалограммы, измеренной виртуальным электродом. Метод был применен к анализу магнитных энцефалограмм спонтанной активности для испытуемых с синдромом дефицита внимания и гиперактивности и для контрольных испытуемых. Были использованы виртуальные электроды различных размеров и пространственных конфигураций. Электроды размещались в областях максимума альфа ритма в правом и левом полушариях и в прилегающих к ним областях. Было произведено сравнение мощностей и спектральных особенностей исследуемых сигналов.

Работа выполнена при поддержке Программы Президиума РАН №2 «Механизмы обеспечения отказоустойчивости современных высокопроизводительных и высоконадежных вычислений» и грантов РФФИ 17-29-02178, 18-00-00619, 19-07-00964, 17-07-00677, 17-07-00686.

[1] *Рыкунов С. Д., Рыкунова Е. Д., Бойко А. И., Устинин М. Н.* Программный комплекс «ВиртЭл» для анализа данных магнитной энцефалографии методом виртуальных электродов // Математическая биология и биоинформатика, 2019. — Т. 14(1)— С. 340–354.

The study of magnetic encephalograms of patients with attention deficit hyperactivity disorder using virtual electrodes

Rykunov Stanislav[1]⋆ rykunov@impb.ru
Ustinin Mikhail[1] u_m_n@mail.ru
Boyko Anna[1] a.boyko@list.ru
Pankratova Natalia[1] natpan1974@mail.ru

[1]Pushchino, IMPB RAS - Branch of KIAM RAS

A new method for analyzing magnetic encephalography data was developed - the virtual electrode method. According to magnetic encephalography, a functional tomogram is constructed — the spatial distribution of field sources on a discrete grid. Functional tomogram displays information contained in multichannel time series of the encephalogram onto the head space. This is achieved by solving the inverse problem for all elementary oscillations extracted using the Fourier transform. Each oscillation frequency corresponds to a node of a three-dimensional grid in which the source is located. The user sets the location, size and shape of the brain area for a detailed study of the frequency structure of a functional tomogram - a virtual electrode. The set of oscillations that fall in a given region is a partial spectrum of this region. From this spectrum, the time series of the encephalogram measured by a virtual electrode are restored. The method was applied to the analysis of magnetic encephalograms of spontaneous activity for subjects with attention deficit hyperactivity disorder and for control subjects. Virtual electrodes of various sizes and spatial configurations were used. The electrodes were located in the areas of maximum alpha rhythm in the right and left hemispheres and in the areas adjacent to them. The power and spectral features of the studied signals were compared.

This research is funded by Program No.2 of the Presidium of the Russian Academy of Sciences and RFBR, grants 17-29-02178, 18-00-00619, 19-07-00964, 17-07-00677, 17-07-00686.

[1] *Rykunov S., Rykunova E., Boyko A., Ustinin M.* VirtEl - Software for Magnetic Encephalography Data Analysis by the Method of Virtual Electrodes // Mathematical Biology and Bioinformatics, 2019. —V. 14(1)— p. 340–354.

Распознавание, отбор структурных мотивов, образованных двумя спиралями в белковых молекулах, и исследование межспиральных углов в спиральных парах

Тихонов Дмитрий Анатольевич[1,3]⋆ dmitry.tikhonov@gmail.com
Куликова Людмила Ивановна[1] likulikova@mail.ru
Ефимов Александр Васильевич[2] Efimov@protres.ru
[1]Пущино, Институт математических проблем биологии РАН филиал ИПМ им. М.В. Келдыша РАН
[2]Пущино, Институт белка РАН
[3]Пущино, Институт теоретической и экспериментальной биофизики РАН

В данной работе проведен анализ распределения межспиральных углов в парах связанных между собой перетяжками спиралей в пространственных структурах белковых молекул. Были разработаны правила отбора спиральных пар в структурах белковых молекул Protein Data Bank. Полученное множество спиральных пар было проанализировано с целью его классификации и установления закономерностей структурной организации. Все отобранные спиральные пары были разделены по типу спиралей, образующих исследуемые структуры. Также по критерию пересечения проекций спиралей на параллельные плоскости, проходящие через оси спиралей, полученное множество было разбито на три подмножества. Проведен анализ распределения углов между осями спиралей всех типов спиральных пар в каждом подмножестве. Показано, что распределение всех типов спиральных пар, не имеющих пересечения проекций спиралей, охватывает практически весь диапазон углов с одним максимумом в области прямого угла. Большинство пар этого множества составляют спиральные пары, состоящие из α и 3_{10} — спиралей, а множества с пересечением только проекций спиралей – спиральные пары, образованные двумя α-спиралями. Также показано, что образованные двумя α-спиралями спиральные пары составляют абсолютное большинство пар подмножества с пересечением проекций и осей спиралей. При этом значительная часть указанных пар имеет острый угол от 20^o до 60^o между осями спиралей. Межплоскостное расстояние для всех этих структур равно 10 A. Проведен анализ распределения всех типов спиральных пар, принадлежащих различным множествам, в зависимости от длины перетяжки. Показано, что во всех множествах больше всего исследуемых структур с короткой перетяжкой. В работе также исследована зависимость торсионных углов между осями α-спиралей от их длины. В результате показано, что пары длинных α-спиралей в большинстве случаев имеют торсионный угол между их осями $\Omega = 20^o$. Большинство спиральных пар, в которых одна спираль короткая, а другая – длинная или обе спирали короткие, образуют мотив с ортогональной ($-90^o < \Omega < -70^o$) или скошенной ($\Omega = -50^o$) упаковкой спиралей. Полученные результаты имеют очень большое значение для определения

взаимной ориентации спиралей при моделировании и предсказании структуры белков.

Работа поддержана грантом РФФИ № 18-07-01031.

[1] *Tikhonov D.A., Kulikova L.I., Efimov A.V.* The Study of Interhelical Angles in the Structural Motifs Formed by Two Helices. // Mathematical Biology and Bioinformatics, 2019, V 14, N S, PP t1–t17.

Recognition, selection of structural motifs formed by two helices in protein molecules, and the study of inter-helical angles in helical pairs

Dmitry Tikhonov[1,3]⋆ dmitry.tikhonov@gmail.com
Ludmila Kulikova[2] likulikova@mail.ru
Alexander Efimov[2] Efimov@protres.ru
[1]Pushchino, Institute of Mathematical Problems of Biology Branch of Keldysh Institute of Applied Mathematics of RAS
[2]Pushchino, Institute of Protein Research RAS
[3]Pushchino, Institute of Theoretical and Experimental Biophysics of RAS

In this paper a statistical analysis of distributions of inter-helical angles in pairs of consecutive and connected α - helices in spatial structures of proteins is presented. A number of rules for selection of the helical pairs from a set of protein structures obtained from the Protein Data Bank (PDB) were developed. The set of helical pairs has been analyzed for the purpose of classification and finding out the features of protein structural organization. All selected helical pairs were divided by the type of helices forming the studied structures. Also all pairs of connected helices were divided into three subsets according to the criterion of crossing of projections of the helices on parallel planes, which pass through the axes of the helices. The analysis of the distribution of angles between the axes of the helices of all types of helical pairs in each subset is carried out. It is shown that the distribution of all types of helical pairs, whose projections do not cross each others, covers almost the entire range of inter-helical angles. The distribution have a single maximum which is close to right angle. Most pairs in this set constitute helical pairs consisting of α- and 3_{10}-helices, and most pairs with the crossing projections of helices are helical pairs formed by two α-helices. It is also shown that a great amount of the pairs of connected α-helices has acute angle $20^o < \varphi < 60^o$ between the axes of the helices. The interplanar distances for all of these structures is 10 A. The distribution of all the types of helical pairs depending on the length of the inter-helical connections is also analyzed. It is shown that the structures with short connections occur most often in all the subsets. The work also investigated relationship between the interhelical packing angles and the length of α-helices in proteins. Analysis of the database has shown that most helical pairs in which both the helices are long form structures having interhelical packing angles $\Omega = 20^o$. Most helical pairs in which one α-helix is long and the other is short or both the helices are short form structures having the orthogonal ($-90^o < \Omega < -70^o$) or slanted ($\Omega = -50^o$)packing of α-helices. These results are of great importance in protein modeling and prediction since enable to determine the mutual arrangement of α-helices in protein molecules. This research is funded by RFBR, grant 18-07-01031.

[1] *Tikhonov D.A., Kulikova L.I., Efimov A.V.* The Study of Interhelical Angles in the Structural Motifs Formed by Two Helices. // Mathematical Biology and Bioinformatics, 2019, V 14, N S, PP t1–t17.

Интерфейс мозг-компьютер: Распознавание визуальных электроэцефалографических потенциалов врача при чтении маммограмм

Сулимова Валентина Вячеславовна[1]⋆ vsulimova@yandex.ru
Красоткина Ольга Вячеславовна[2] okrasotkina@markovprocesses.com
Виндридж Дэвид[3] d.windridge@mdx.ac.uk
Моттль Вадим Вячеславович[4] vmottl@yandex.ru
Морозов Алексей Олегович[5] ao.morozov@phystech.edu

[1]Тула, Тульский государственный университет
[2]Саммит, США, Markov Processes International
[3]Лондон, Великобритания, Middlesex University
[4]Москва, Вычислительный центр РАН
[5]Москва, Московский физико-технический институт

Электроэнцефалография изначально появилась как средство изучения механизмов управления активностью человека, в частности, с целью диагностики заболеваний мозга. В последние десятилетия электроэнцелография все чаще используется как базис для построения различных интерфесов мозг-компьютер, преобразующих отклик нейронов мозга на определенные стимулы в команды управления внешними устройствами. В нашем исследовании [1] анализ откликов многоканальной электроэнцефалограммы (ЭЭГ) на внешние стимулы направлен на достижение иной цели. Предполагается, что анализируется ЭЭГ опытного врача-маммолога, владеющего искусством отличать рентгеновские маммограммы женщин, больных раком груди, от маммограмм здоровых женщин. Целью исследования является поиск путей повышения производительности редких выдающихся диагностов путем, во-первых, ускорения просмотра маммографических изображений до десяти в секунду и, во-вторых, немедленного распознавания потенциала, вызванного в ЭЭГ эксперта раковой рентгенограммой, появившейся среди множества здоровых, раньше, чем сам эксперт успеет осознать этот факт и отреагировать не него.

Работа поддержана грантом РФФИ №18-07-01087.

[1] *V. Sulimova, O. Krasotkina, at al.* Regularized SVMs for classification of image evoked EEG potentials captured from an observer. Proceedings of the 15th Int. Conf. on Machine Learning and Data Mining MLDM 2019, Vol. I, pp. 355-366.

Brain-computer interface: Detecting visual electroencephalogradic potentials of the physician evoked by his reading of XR mammograms

Valentina Sulimova[1]⋆ vsulimova@yandex.ru
Olga Krasotkina[3] okrasotkina@markovprocesses.com
David Windridge[3] d.windridge@mdx.ac.uk
Vadim Mottl[4] vmottl@yandex.ru
Alexey Morozov[5] ao.morozov@phystech.edu
[1]Tula State University
[2]Summit, NJ, USA, Markov Processes International
[3]London, GB, Middlesex University
[4]Moscow, Computing Center of the Russian Academy of Sciences
[5]Moscow Institute of Physics and Technology

Electroencephalography is a method of testing the electrical activity of the brain by jointly processing electrical signals registered at several points on the surface of the skull. It was originally invented to study mechanisms by which human behavior is generated, in particular, for brain diseases diagnosis. However, in the past decades, electroencephalography has become the basis of many brain-computer interfaces, which decode neural response to different stimuli into commands that, for instance, operate external devices. Our research [1] is concerned with another purpose of analyzing responses of a multi-channel electroencephalogram (EEG) to outward stimuli. It is assumed that the person whose EEG is processed is an experienced mammologist able to reliably distinguish between X-ray mammograms of women with breast cancer and those of healthy women. These studies pursue the aim to essentially improve productivity of rare pronounced experts by way of, first, accelerating the screening of mammographic images up to ten pictures per second, and, second, immediately detecting the eventual potentials evoked in the expert's EEG by a target (cancer) image among a crowd of non-target ones before the expert becomes aware of this fact.

This research is funded by RFBR, grant No18-07-01087.

[1] *V. Sulimova, O. Krasotkina, at al.* Regularized SVMs for classification of image evoked EEG potentials captured from an observer. Proceedings of the 15th Int. Conf. on Machine Learning and Data Mining MLDM 2019, Vol. I, pp. 355-366.

Основы создания прикладной интеллектуальной системы персонифицированного предсказания проявления аутоиммунных заболеваний и шизофрении

Янковская Анна Ефимовна[1]⋆ ayyankov@gmail.com
Часовских Наталия Юрьевна[2] nch03@mail.ru
Пеккер Яков Семенович[2] pekker@ssmu.ru
Гречишникова Александра Юрьевна[2] grechishnikova.al@mail.ru
[1]Россия, Томск НИ ТГУ
[2]Россия, Томск, СибГМУ
[3]Россия, Томск, НИ ТПУ

Важнейшим аспектом исследований медико-биологической науки является этиопатогенез мультифакториальных заболеваний. При изучении комплексных (мультифакториальных) патологий наиболее значимыми являются факторы генетической предрасположенности. В настоящее время большое внимание уделяется механизмам развития социально-значимого заболевания шизофрении [1], особенно в сочетании с другими мультифакториальными патологиями, в том числе аутоиммунными. В публикации [2] приведены исследования, направленные на установление связи сочетанного развития шизофрении и отдельных аутоиммунных заболеваний (шизофрении и целиакии, шизофрении с ревматоидным артритом и шизофрении с рассеянным склерозом). Однако, комплексных исследований, позволяющих с применением методов выявления различного рода закономерностей и принятия решений по персонифицированному предсказанию возможного проявления (в том числе сочетанного) аутоиммунных заболеваний и шизофрении не проводилось. Такие исследования в целях персонифицированного предсказания возможного проявления (в том числе совместного) аутоиммунных заболеваний и шизофрении целесообразны осуществить с использованием методов тестового распознавания образов с применением средств когнитивной графики, реализованных в прикладной интеллектуальной системы (ИС ПРЕДПАТ), предназначенной для выявления возможных сочетанных проявлений генетических факторов мультифакториальных заболеваний. Ниже вкратце излагаются проведённые нами предварительные исследования по рассматриваемой проблемной области, основы создания прикладной ИС ПРЕДПРАТ и дальнейшие исследования.

На основе проведённого нами анализа информации из каталога GWAS (каталог ассоциаций однонуклеотидных полиморфизмов с заболеваниями) по результатам широкомасштабных полногеномных исследований [3] были выявлены гены предрасположенности к ревматоидному артриту, целиакии, рассеянному склерозу и шизофрении: 927 гена предрасположенности для шизофрении, 434 - для ревматоидного артрита, 243 – для рассеянного склероза и 102 – для целиакии. С целью выявления генетических факторов таких социально-значимых заболеваний, как аутоиммунные (ревматоидный артрит, целиакия, рассеянный

склероз) и шизофрения будут сформированы и заполнены матрицы описаний и различений изучаемых заболеваний [4], служащие основой для построения базы данных и знаний прикладной ИС ПРЕДПРАТ.

Впервые предлагается прикладная ИС ПРЕДПРАТ, позволяющая эффективно проводить анализ имеющихся результатов различных полногеномных исследований, сформировать комплексную картину генетических взаимодействий при исследуемых патологиях, прогнозировать индивидуальные факторы сочетанного проявления рассматриваемых заболеваний.

Математический аппарат прикладной ИС ПРЕДПАТ, базируемый на конвергенции нескольких наук и научных направлений, основан на матричном способе представления данных и знаний (матрицы описаний и различений), оригинальных тестовых методах распознавания образов; выявлении различного рода закономерностей (общее число 10), включая отказоустойчивые безызбыточные и смешанные диагностические тесты и их весовые коэффициенты; принятии решения и их обоснования с использованием графических, включая когнитивные, средств [4,5]. ИС ПРЕДПАТ создается на базе интеллектуального инструментального средства ИМСЛОГ [6], предназначенного для выявления различного рода закономерностей, включая отказоустойчивые диагностические тесты; принятия и обоснования решений с использованием средств когнитивной графики.

Анализ имеющихся результатов различных полногеномных исследований, содержащихся в созданных нами базах данных и знаний, впервые позволит выявить комплексную картину генетических взаимодействий при исследуемых патологиях (заболеваний), а также для обследуемых пациентов прогнозировать заболевания, индивидуальные факторы сочетанного проявления которых представлены матрицей различений исследованных заболеваний. Предложенный подход для анализа данных полногеномных исследований послужит основой для разработки методологии диагностики сочетанного проявления исследуемых заболеваний в рамках исследований персонализированной медицины. На основании индивидуальных геномных данных будет прогнозироваться наличие совместного проявления аутоиммунных заболеваний и шизофрении у обследуемых лиц.

[1] *Bush W. S., Moore J. H* Genome-wide association studies // PLoS Comput Biol, No 8, 2012. — Pp. 1–11.

[2] *Eaton W. W., Nielsen P R., Pedersen M G.* Autoimmune disease, bipolar disorder, and non-affective psychosis // Bipolar disorder, No 12, 2010. — Pp. 638–646.

[3] *Eaton W. W., Nielsen P. R., Pedersen M. G.* The NHGRI GWAS Catalog, a curated resource of SNPtrait associ-ations // Nucleic Acids Research, No 42, 2014. — Pp. 1001–1006.

[4] *Янковская А. Е.* Логические тесты и средства когнитивной графики. Издательский Дом: LAP LAMBERT Academic Publishing, 2011. – 92 с.

[5] *Yankovskaya A.* 2-Simplex Prism as a Cognitive Graphics Tool for Decision-Making // Encyclopedia of Computer Graphics and Games. Springer Nature Switzerland AG 2019 N. Lee (ed.). 2019. – 13 p.

[6] *Yankovskaya A. E., Gedike A I., Ametov R V., Bleikher A M.* IMSLOG-2002 Software Tool for Supporting Information Technologies of Test Pattern Recognition // Pattern Recognition and Image Analysis, Vol. 13, No 4, 2003. — Pp. 650–657.

Fundamentals of construction of applied intelligent system for personalized prediction of autoimmune deceases and schizophrenia

Ann Yankovskaya [1]⋆ ayyankov@gmail.com
Natalia Chasovskikh[2] nch03@mail.ru
Yakov Pekker[2] pekker@ssmu.ru
Alexandra Grechishnikova[2] grechishnikova.al@mail.ru
[1]Russia, Tomsk NR TSU
[2]Russia, Tomsk, SSMU
[3]Russia, Tomsk, NR TPU

The most important aspect of medical and biological science research is the etiopathogenesis of multifactor diseases. The special role taken to the factors of genetic predisposition at the study of complex (multifactorial) pathologies. Paid attention presently large to the mechanisms of development of socially-meaningful disease of schizophrenia in combination with other multifactorial pathologies, including autoimmune. The researches directed to establishment of communication of the combined development of schizophrenia and individual autoimmune diseases are given in the publication [2] (schizophrenia and celiac disease, schizophrenia with rheumatoid arthritis and schizophrenia with multiple sclerosis). However, no comprehensive studies have been carried out using methods of identifying various patterns and deciding on the personalized prediction of the possible manifestation (including the combined) of autoimmune diseases and schizophrenia. Such researches are possible only based on application of the applied intelligent system (IS) for the personified prediction of possible manifestation (including joint) autoimmune diseases and schizophrenia (IS PREDPAT) allowing to reveal the possible combined manifestations of genetic factors of multifactorial diseases. Below the preliminary researches conducted by us on considered problem area, a basis of creation of the applied IS PREDPAT and further researches are in brief stated.

By results of the analysis of information, which carried out by us from the GWAS catalog of large-scale full-genomic researches [3] genes of predisposition to a pseudorheumatism, a Gee's disease, multiple sclerosis and schizophrenia were reveal: 927 predisposition genes for schizophrenia, 434 - for a pseudorheumatism, 243 - for the dissipated sclerosis and 102 - for celiac disease. With the purpose of revealinge of genetic factors of such socially meaningful diseases as autoimmune and schizophrenia, will be formed and filled matrices of descriptions of the studied diseases [4], for a construction data and knowledgebase of applied IS PREDPAT.

The applied first offered IS PREDPAT allowing effectively conducting analysis of present results of different full genome researches, formatiing complex picture of genetic interaction at the pathologies under investigation, prediction of individual factors of the diseases under examination.

A mathematical apparatus of applied IS PREDPAT, based on convergences of a few sciences and scientific directions is based on the matrix method of representation

of the data and knowledge, original test methods of pattern recognition; revealing different sort of regularities (sum total 10), irredundant and mixed diagnostic tests with their coefficients, decision –making with the using graphic and tool cognitive graphics [4,5].

IS PREDPAT is created on the base of intelligent tool of IMSLOG [6] different sort of conformities to law intended for an exposure including fault-tolerant diagnostic tests and acceptance and ground of decisions with the use of cognitive means.

Analysis of the results of various full genome studies for the first time will allow to reveal a complex picture of genetic interactions in the investigated pathologies, as well as to predict diseases, individual factors of combined manifestation of which are included in the matrix of distinguishing of diseases under investigation. The proposed approach for the analysis of full genome research data will serve as a basis for the development of a methodology for the diagnosis of the combined manifestation of the studied diseases within the framework of research of personalized medicine. Because of individual genomic data, the possibility of joint manifestation of autoimmune diseases and schizophrenia in the examined persons will be predict.

[1] *Bush W. S., Moore J. H* Genome-wide association studies // PLoS Comput Biol, No 8, 2012. — Pp. 1–11.

[2] *Eaton W. W., Nielsen P R., Pedersen M G.* Autoimmune disease, bipolar disorder, and non-affective psychosis // Bipolar disorder, No 12, 2010. — Pp. 638–646.

[3] *Eaton W. W., Nielsen P. R., Pedersen M. G.* The NHGRI GWAS Catalog, a curated resource of SNPtrait associations // Nucleic Acids Research, No 42, 2014. — Pp. 1001–1006.

[4] *Yankovskaya A. E.* Logical tests and cognitive graphics. LAP LAMBERT Academic Publishing, 2011. – 92 p.

[5] *Yankovskaya A.* 2-Simplex Prism as a Cognitive Graphics Tool for Decision-Making // Encyclopaedia of Computer Graphics and Games. Springer Nature, 2019, — 13 p.

[6] *Yankovskaya A. E., Gedike A I., Ametov R V., Bleikher A M.* IMSLOG-2002 Software Tool for Supporting Information Technologies of Test Pattern Recognition // Pattern Recognition and Image Analysis, Vol. 13, No 4, 2003. — Pp. 650–657.

Основы создания прикладной интеллектуальной системы диагностики качества жизни пациентов с неврологической патологией

Янковская Анна Ефимовна[1]⋆ ayyankov@gmail.com
Обуховская Виктория Борисовна[1,2] diada1991@gmail.com
[1]Россия, Томск НИ ТГУ
[2]Россия, Томск, СибГМУ

В настоящее время широкая распространенность неврологической патологии является одной из наиболее значимых медицинских и социальных проблем, поскольку патология пациента приводит к значительному снижению качества жизни. Низкое качество жизни является фактором, снижающим психологическую безопасность и приводящим к самостигматизации пациентов [1]. Для проведения мероприятий по повышению качества жизни необходимо воздействовать на весь комплекс физических и психологических факторов [2]. Актуальность диагностики качества жизни пациентов с неврологической патологией не вызывает сомнений, поскольку неврологическая патология и снижение качества жизни имеет тяжелые психологические, социальные и соматические последствия, утрату трудоспособности и инвалидизацию. Настоятельная необходимость создания прикладной интеллектуальной системы диагностики качества жизни пациентов с неврологической патологией (ИС ДИАКАЖ) не вызывает сомнений. Ниже вкратце излагаются проведённые нами исследования по структуризации проблемной области и построению базы данных и знаний, основам создания прикладной ИС ДИАКАЖ и приводятся результаты исследования.

На основе проведённого нами анализа пациентов с различной неврологической патологией (болезнь Паркинсона, рассеянный склероз, остеохондроз позвоночника, последствия инсульта, головокружения и нарушения устойчивости) были выявлены параметры (признаки), определяющие физический компонент здоровья (физическое функционирование; ролевое функционирование, обусловленное физическим состоянием; интенсивность боли; общее состояние здоровья) и психический компонент здоровья (жизненная активность; социальное функционирование; ролевое функционирование, обусловленное эмоциональным состоянием; психологическое здоровье) пациентов с неврологической патологией. Осуществлена структуризация данных и знаний с применением матрицы описаний Q объектов (пациентов) в пространстве признаков и различений R, задающих различные механизмы разбиения объектов на классы эквивалентности [3]. Строки матрицы R сопоставлены строкам матрицы Q. Сформированы и заполнены матрицы описаний и различений изучаемых заболеваний, служащие основой для построения базы данных и знаний прикладной ИС ДИАКАЖ.

Впервые создаётся прикладная ИС ДИАКАЖ, предназначенная для своевременной и эффективной (повышающей достоверность результатов и сокращающей временные затраты) диагностики качества жизни пациентов с невро-

логической патологией на основе кратких опросников. Математический аппарат прикладной ИС ДИАКАЖ базируется на конвергенции нескольких наук и научных направлений, основан на матричном способе представления данных и знаний (матрицы описаний и различений); оригинальных тестовых методах распознавания образов; выявлении различного рода закономерностей, включая отрицательные образы; альтернативные, зависимые и сигнальные признаки; отказоустойчивые безызбыточные и смешанные диагностические тесты и их весовые коэффициенты; принятии решения и их обоснования с использованием графических, включая когнитивные, средств [4]. ИС ДИАКАЖ конструируется на базе интеллектуального инструментального средства (ИИС) ИМСЛОГ [5], предназначенного для выявления различного рода закономерностей, включая отказоустойчивые диагностические тесты и их весовые коэффициенты; принятии решения и их обоснования с использованием средств когнитивной графики. База данных и знаний создаётся на основе результатов исследования пациентов с неврологической патологией, находящихся на лечении в клиниках неврологического профиля.

Впервые предложено создание прикладной ИС ДИАКАЖ. Впервые проведён анализ проблемной области, структуризация данных и знаний пациентов с различной неврологической патологией и формирование матриц описаний и различений. Поскольку сконструированные на базе ИИС ИМСЛОГ более 30 прикладных ИС показали высокую эффективность, есть все основания, что применение прикладной ИС ДИАКАЖ позволит повысить качество диагностики и сократить временные затраты на её проведение. Работа поддержана грантом РФФИ № 18-313-00195.

[1] *Соловьева С. Л.* Самостигматизация как фактор превращения личности здорового в личность больного // Неврологический вестник. Журнал им. В. М. Бехтерева. Т. 49, вып. 1. с. 49–56, 2017.

[2] *Молчанова Ж. И.* Качество жизни у больных рассеянным склерозом проживающих в северном регионе в зависимости от когнитивного статуса // Вестник новых медицинских технологий. Т. 21. с. 104–107, 2014.

[3] *Янковская А. Е.* Логические тесты и средства когнитивной графики. Издательский Дом: LAP LAMBERT Academic Publishing, 2011. – 92 с.

[4] *Yankovskaya A.* 2-Simplex Prism as a Cognitive Graphics Tool for Decision-Making // Encyclopedia of Computer Graphics and Games. Springer Nature Switzerland AG 2019 N. Lee (ed.). 2019. – 13 p.

[5] *Yankovskaya A. E., Gedike A I., Ametov R V., Bleikher A M.* IMSLOG-2002 Software Tool for Supporting Information Technologies of Test Pattern Recognition // Pattern Recognition and Image Analysis, Vol. 13, No 4, 2003. — Pp. 650–657.

Basics of creating an applied intelligent system for diagnostic the quality of life of patients with neurological pathology

Ann Yankovskaya [1]⋆ ayyankov@gmail.com
Victoria Obukhovskaya[2] diada1991@gmail.com
[1]Russia, Tomsk NR TSU
[2]Russia, Tomsk, SSMU

Currently, the widespread prevalence of neurological pathology is one of the most significant medical and social problems, since the patient's pathology leads to a significant decrease in the quality of life. Poor quality of life is a factor that reduces psychological safety and leads to self-stigmatization of patients [1]. To improve the quality of life, it is necessary to influence the whole complex of physical and psychological factors by performing certain types of activities [2]. The relevance of the quality of life diagnostics for patients with neurological pathology is beyond doubt, since the pathology and the decrease in the quality of life have serious psychological, social and somatic consequences, and disabilities. The urgent need for an applied intelligent system development of the quality of life diagnostics for patients with neurological pathology (IS DIAQOL) is currently undoubtedly relevant. Below, we briefly outline our studies on structuring the problem area and construction a data and knowledge base, the fundamentals of creating an applied IS DIAQOL, and further studies.

Based on our analysis of patients with various neurological pathologies (Parkinson's disease, multiple sclerosis, spinal osteochondrosis, the effects of stroke, dizziness and impaired stability), we revealed the parameters (features) that determine the Physical Health (physical functioning; role-physical functioning; bodily pain; general health) and the Mental Health (vitality; social functioning; role-emotional, general mental health) of patients with neurological disorders. The data and knowledge were structured using the matrix of description Q of objects (patients) in the space of features and the distinguishing matrix R, defining various mechanisms of dividing objects into equivalent classes [3]. The rows of the matrix R correspond to the rows of the matrix Q. Creating and filling the above mentioned matrixes of the diseases under study serve as the basis for construction of the data and knowledge base for the applied IS DIAQOL.

For the first time, the applied IS DIAQOL is designed for timely and effective (increased reliability of results and minimized time costs) life quality diagnostics for patients with neurological pathology based on brief questionnaires. The mathematical apparatus of the applied IS DIAQOL is based on the convergence of several sciences and scientific researches, based on the matrix method of data and knowledge representation (matrixes of description and distinguishing); original test methods for pattern recognition; revealing regularities, including negative patterns; alternative, dependent and signal features; fault-tolerant irredundant and mixed diagnostic tests and their weight coefficients; decision making and their justification using graphic,

including cognitive, tools [4]. Applied intelligent system is constructed on the base of intelligent instrumental software (ISS) IMSLOG, designed to revealing various of regularities, including fault-tolerant diagnostic tests and their weight coefficients; decision making and justification using cognitive graphics tools. The data and knowledge base is created using the results of research involving patients with neurological pathology being treated in neurological clinics.

For the first time, the creation of an applied intelligent system for quality of life diagnostics of patients with neurological pathology is proposed. For the first time, analysis, structurization of data and knowledge of patients with various neurological pathologies were carried out. Since more than 30 applied IS designed on the basis of the ISS IMSLOG have shown high efficiency, there is every reason that the use of the applied IS DIAQOL can improve diagnostics of the quality of life. That will reduce the time and cost on the diagnostics of patients with neurological deceases.

The reported study was funded by RFBR according to the research project No 18-313-00195.

[1] *Solovieva S. L.* Self-Stigmatization as Factor in Transformation of Healthy Person's Personality into Patient's Personality // Neurological Bulletin. Vol. 49, No 1, Pp. 49–56, 2017.

[2] *Molchanova Zh. I.* The Quality of Life in Multiple Sclerosis Patients Living in the North Region, Depending on Cognitive Status // Journal of New Medical Technologies. Vol. 21, Pp. 104–107, 2014.

[3] *Yankovskaya A. E.* Logical tests and cognitive graphics. LAP LAMBERT Academic Publishing, 2011. – 92 p.

[4] *Yankovskaya A.* 2-Simplex Prism as a Cognitive Graphics Tool for Decision-Making // Encyclopedia of Computer Graphics and Games. Springer Nature Switzerland AG 2019 N. Lee (ed.). 2019. – 13 p.

[5] *Yankovskaya A. E., Gedike A I., Ametov R V., Bleikher A M.* IMSLOG-2002 Software Tool for Supporting Information Technologies of Test Pattern Recognition // Pattern Recognition and Image Analysis, Vol. 13, No 4, 2003. — Pp. 650–657.

Кластерный анализ в задаче дооперационного прогнозирования метастатического поражения регионарных лимфоузлов у больных раком молочной железы

Аснина Наталия Георгиевна[1]⋆ andrey050569@yandex.ru
Азарнова Татьяна Васильевна[2] ivdas92@mail.ru
[1]Воронеж, ФГБОУ ВО «Воронежский государственный технический университет»
[2]Воронеж, ФГБОУ ВО «Воронежский государственный университет»

Выраженный рост заболеваемости женщин раком молочной железы инициировал поиск новых путей лечения этой патологии, в частности, широкое внедрение экономных оперативных вмешательств.

Ключевым фактором определения прогноза и выбора оптимальной лечебной тактики является своевременная диагностика поражения регионарных лимфатических узлов. Однако, следует отметить, что неинвазивные методы диагностики метастазов в лимфоузлы характеризуются низкой специфичностью и недостаточной чувствительностью, а процедура биопсии сигнальных лимфоузлов (БСЛУ), которая считается методом выбора во многих клинических ситуациях, из-за технических сложностей рутинно применяется не во всех стационарах.

Таким образом, дооперационный статистический прогноз метастатического поражения регионарных лимфоузлов у больных раком молочной железы, даст возможность уже до начала операции оценить возможность оптимизации лимфодиссекции.

В работе проведен статистический анализ клинических данных и морфологических факторов, полученных в онкологическом отделении патологии молочной железы Воронежского областного клинического онкологического диспансера в ходе проведения комплексного исследования с участием 223 пациенток. Учитывались только стадии I T1N0M0 и IIa T2N0M0.

В таблице 1 представлен фрагмент клинико-морфологических данных 223 пациенток, разделенных на группы по признаку «отсутствие/наличие метастазов в лимфоузлы».

На первом этапе в качестве переменных для кластеров были взяты возраст пациентов, индекс пролиферации Ki67 (фактор определения биологического подтипа) и размер опухоли в миллиметрах. В результате применения кластерного анализа методом k-средних было получено три кластера. В первый кластер попали пациенты со средним возрастом 56,9 лет, средним Ki67=74,9 и средним размером опухоли 33,9 мм во второй со средними значениями 58,14; 34,25; 24,45 и в третий 64,60; 16,52; 18,23 соответственно.

Далее оценим частоту метастазирования в регионарные лимфоузлы для каждого кластера. Такую оценку проведем с использованием таблиц сопряженности. Анализ показал, что на уровне значимости $p < 0{,}05$ частота метастазирования в регионарные лимфоузлы оценивается более чем в 2 раза ниже у

Таблица 1

Клинико-морфологические данные больных раком молочной

Категория	Отсутствие метастазов N = 154		Наличие метастазов N = 69		Всего	
Средний возраст	61,32±10,38		61±12,02		61,25±11,22	
Средний размер опухоли (мм)	20,43±9,28		27,43±12,4		22,61±10,83	
Гистологический диагноз	N	%	N	%	N	%
инвазивная карцинома G2	98	63,63	52	75,36	150	67,26
Биологический подтип	N	%	N	%	N	%
Люминальный А	70	45,45	20	28,98	90	40,36
Люминальный В Her -	46	29,87	31	44,92	77	34,53

пациентов со средним возрастом 65 лет, средним Ki-67 16,52 % (кластер 3) и средним размером опухоли 18,2 мм, чем у пациентов со средним возрастом 57 лет, средним Ki-67 75 % и средним размером опухоли 34мм (кластер 1).

На следующем этапе анализа в набор переменных для кластеров нами были включены показатели гормональных характеристик клеток опухоли Er, PgR, HER2, гистологический диагноз. В результате было получено пять кластеров. Остановимся подробнее на структуре полученных кластеров. В пятый кластер попали пациенты с самым низким средним Ki67(8,9), с отрицательным HER2 и высокой активностью Er и PgR, а в третий кластер пациенты с высоким Ki67(82), положительным HER2 и низкой экспрессией Er и PgR.

Оценка частоты метастазирования в регионарные лимфоузлы для каждого из пяти кластеров, с использованием таблицы сопряженности показала следующие результаты: на уровне значимости $p < 0{,}05$ частота метастазирования в регионарные лимфоузлы оценивается более чем в 2 раза ниже у пациентов 4 кластера по сравнению с пациентами из первого кластера. Именно для этого кластера наблюдаются самые благоприятные прогнозы по поводу отсутствия метастазов в лимфоузлах (78 %; $p < 0{,}05$), а для 1 кластера характерно наиболее частое метастазирование в регионарные лимфоузлы (51,85 %).

Проведенные исследования показали, что существует возможность дооперационного прогнозирования метастатического поражения регионарных лимфатических узлов у больных раком молочной железы [1].

[1] *Исмагилов А. Х., Аснина Н. Г, Азаров Г. А., Мошуров И. П.* Прогнозирование метастатического поражения регионарных лимфатических узлов у больных раком молочной железы // Опухоли женской репродуктивной систем, 2017. — 13(2). — С. 13–19.

Table 1

Clinical and morphological data of patients with breast cancer

Category	Lack of metastases N = 154		The presence of metastases N = 69		Total	
Average age	61,32±10,38		61±12,02		61,25±11,22	
The average tumor size (mm)	20,43±9,28		27,43±12,4		22,61±10,83	
Histological diagnosis	N	%	N	%	N	%
invasive carcinoma G2	98	63,63	52	75,36	150	67,26
Biological subtype	N	%	N	%	N	%
Luminal A	70	45,45	20	28,98	90	40,36
Luminal B Her -	46	29,87	31	44,92	77	34,53

Cluster analysis in the task of preoperative prognosis of metastatic lesions of regional lymph nodes in patients with breast cancer

Natalia Asnina[1] andrey050569@yandex.ru
Tatiana Azarnova[2]⋆ ivdas92@mail.ru
[1]Voronezh, Federal State Budgetary Educational Institution of Higher Education "Voronezh State Technical University"
[2]Voronezh, Federal State Budgetary Educational Institution of Higher Education "Voronezh State University"

A marked increase in the incidence of women with breast cancer initiated the search for new ways to treat this pathology, in particular, the widespread introduction of cost-effective surgical interventions.

A key factor in determining the prognosis and choosing the optimal treatment tactics is the timely diagnosis of regional lymph node lesions. However, it should be noted that non-invasive methods for the diagnosis of lymph node metastases are characterized by low specificity and insufficient sensitivity, and the signal lymph node biopsy procedure (BSL), which is considered the method of choice in many clinical situations, is not routinely used in all hospitals due to technical difficulties. In this way, preoperative statistical prognosis of metastatic lesions of regional lymph nodes in patients with breast cancer will make it possible to evaluate the possibility of optimizing lymphadenectomy already before the start of surgery.

A statistical analysis of clinical data and morphological factors obtained in the oncology department of breast pathology of the Voronezh Regional Clinical Oncology Center during a comprehensive study involving 223 patients was performed. Only stages I T1N0M0 and IIa T2N0M0 were taken into account.

In Table 1 are clinical and morphological data of 223 patients is presented and divided into groups on the basis of "the absence/presence of metastases in the lymph nodes".

At the first stage, the age of the patients, the Ki67 proliferation index (factor for determining the biological subtype), and tumor size in millimeters were taken as

variables for the clusters. As a result of applying cluster analysis using the k-means method, three clusters were obtained. The first cluster included patients with an average age of 56.9 years, an average Ki67 = 74.9, and an average tumor size of 33.9 mm in the second, with an average of 58.14; 34.25; 24.45 and the third 64.60; 16.52; 18.23.

Next, we estimate the frequency of metastasis to regional lymph nodes for each cluster. We carry out such an assessment using contingency tables.

The analysis showed that at a significance level of $p < 0.05$ h, the rate of metastasis to regional lymph nodes is estimated to be more than 2 times lower in patients with an average age of 65 years, an average Ki -67 of 16.52 % (cluster 3) and an average tumor size of 18, 2 mm, than in patients with an average age of 57 years, an average Ki -67 of 75 % and an average tumor size of 34 mm (cluster 1).

At the next stage of the analysis, we included in the set of variables for clusters indicators of hormonal characteristics of tumor cells Er, PgR, HER2, and histological diagnosis. As a result, five clusters were obtained. Let us dwell in more detail on the structure of the resulting clusters. So, for example, patients with the lowest average Ki 67 (8.9), with negative HER2 and high Er and PgR activity were included in the fifth cluster , and patients with high Ki 67 (82), positive HER2 and low expression were in the third cluster Er and PgR.

Estimation of the frequency of metastasis to regional lymph nodes for each of the five clusters using the contingency table showed the following results: at a significance level of $p < 0.05$, the frequency of metastasis to regional lymph nodes is estimated to be more than 2 times lower in patients of 4 clusters compared with patients from the first a cluster. It is for this cluster that the most favorable prognoses are observed regarding the absence of metastases in the lymph nodes (78 %; $p < 0.05$), and for 1 cluster the most frequent metastasis to regional lymph nodes is characteristic (51.85 %).

Studies have shown that there is the possibility of preoperative prognosis of metastatic lesions of regional lymph nodes in patients with breast cancer not by one attribute (histological or morphological), but by their combination.

[1] *Ismagilov A., Asnina N, Azarov G, Moshurov I.* Prediction of metastatic lesion of regional lymph nodes in patients with breast cancer// Tumors of the female reproductive system, 2017. — 13(2). — p. 13–19.

Ранжирование и анализ моделей белок-белкового докинга онлайн мета-сервером QASDOM

Кузнецов Евгений Николаевич[1] kuznetsov.eugene@gmail.com
Кравацкий Юрий Викторович[2] jiri@eimb.ru
Туманян Владимир Гайевич[2] tuman@eimb.ru
Аджубей Алексей Алексеевич[2] alexei.adzhubei@eimb.ru
Анашкина Анастасия Андреевна[2]⋆ nastya@eimb.ru
[1]Москва, Институт проблем управления им. В.А.Трапезникова РАН
[2]Москва, Институт молекулярной биологии им. В.А.Энгельгардта РАН

Предсказание взаимодействия белков является важным инструментом для изучения процессов, происходящих в живой клетке. В настоящее время существуют несколько серверов, использующих разные алгоритмы докинга. Результаты моделирования выдаются в виде наборов разных вариантов взаимодействия рецептора и лиганда. Однако оценка и сравнение всех моделей такого набора экспертом весьма трудоемка.

Сервер QASDOM [1] (Quality ASsessment of Docking Models, qasdom.eimb.ru) разработан для пользователей, которым необходимо проанализировать набор моделей докинга большого объема, полученный разными способами и инструментами предсказания, оценить вероятность вовлечения тех или других остатков в процесс узнавания, ранжировать модели по критериям качества и выбрать наилучшую модель в режиме реального времени. Сервер позволяет визуализировать участки последовательности рецептора и лиганда, вовлеченные во взаимодействия, и отобразить в трехмерном пространстве структуры моделей комплекса рецептор-лиганд.

Описаны критерии оценки и ранжирования моделей докинга, определение линейных и структурных кластеров взаимодействия молекул, статистика обращений к серверу QASDOM и примеры его использования в конкретных исследованиях.

Работа поддержана грантом РФФИ №17-04-02105.

[1] *Anashkina A, Kravatsky Y, Kuznetsov E, Makarov A, Adzhubei A* Meta-server for automatic analysis, scoring and ranking of docking models // Bioinformatics, 2018. — Volume 34, Issue 2, p.297—299.

Ranking and analysis of protein-protein docking models online by QASDOM meta-server

Eugene Kuznetsov[1] kuznetsov.eugene@gmail.com
Yury Kravatsky[2] jiri@eimb.ru
Vladimir Tumanyan[2] tuman@eimb.ru
Alexei Adzhubei[2] alexei.adzhubei@eimb.ru
Anastasia Anashkina[2]⋆ nastya@eimb.ru

[1]Moscow, V. A. Trapeznikov Institute of Control Sciences of RAS
[2]Moscow, Engelhardt Institute of Molecular Biology of RAS

Prediction of protein interactions is an important tool for studying the processes occurring in a living cell. There are currently several servers using different docking algorithms. The simulation results are given in the form of sets of different variants of the interaction of the receptor and the ligand. However, the assessment and comparison of all models of such a set by an expert is very laborious.

The QASDOM [1] server (Quality ASsessment of Docking Models, qasdom.eimb.ru) is designed for users who need to analyze a set of large docking models obtained in different ways and prediction tools, to assess the probability of involving certain residues in the process recognition, rank models according to quality criteria and choose the best model in real time mode. The server allows you to visualize portions of the receptor and ligand sequences involved in the interaction and display the structure of the models of the receptor-ligand complex in the three-dimensional space.

Criteria for assessing and ranking docking models, determining linear and structural clusters of molecular interactions, statistics of calls to the QASDOM server, and examples of its use in specific studies are described.

This work was supported by the RFBR grant 17-04-02105.

[1] *Anashkina A, Kravatsky Y, Kuznetsov E, Makarov A, Adzhubei A* Meta-server for automatic analysis, scoring and ranking of docking models // Bioinformatics, 2018. — Volume 34, Issue 2, p.297—299.

Подход к детектированию эпилептиформной активности в сигналах ЭЭГ и способы дифференциации эпилептических приступов от артефактов жевания

Кершнер Иван Андреевич[1]⋆ ivan.kershner@gmail.com
Синкин Михаил Владимирович[2] mvsinkin@gmail.com
Обухов Юрий Владимирович[1] yuvobukhov@mail.ru

[1]Москва, Институт радиотехники и электроники им. В.А. Котельникова РАН
[2]Москва, Научно-исследовательский институт скорой помощи имени Н. В. Склифосовского

Разработан новый метод, позволяющий автоматически детектировать различную активность в длительных (сутки и более) сигналах электроэнцефалограмм. Для дифференциации эпилептиформной активности от артефактов жевания предложены методы, основанные на анализе вейвлет-спектрограмм электроэнцефалограмм. Изучались свойства хребтов вейвлет-спектрограмм электроэнцефалограмм, а также периодичность широкополосных пиков в моменты времени, соответствующие пик-волновой эпилептиформной активности, с одной стороны, и пикам миографической активности в электроэнцефалограмме при жевании с другой стороны.

Работа выполнена в рамках государственного задания и частично поддержана Российским фондом фундаментальных исследований (проект № 18-29-02035 мк).

[1] *Kershner I., Obukhov Yu., Sinkin M.* A new approach to the detection of epileptiform activity in EEG signals and methods to differentiate epileptic seizures from chewing artifacts // Rensit: Radioelectronics. Nanosystems. Information techologies, 2019. — Vol. 11, No. 2. — pp. 237-242.

Approach to the detection of epileptiform activity in EEG signals and methods to differentiate epileptic seizures from chewing artifacts

Ivan Kershner[1]⋆ ivan.kershner@gmail.com
Mikhail Sinkin[2] mvsinkin@gmail.com
Yury Obukhov[1] yuvobukhov@mail.ru

[1]Moscow, Kotel'nikov Institute of Radio Engineering and Electronics of Russian academy of sciences

[2]Moscow, Sklifosovsky Scientific Research Institute for Emergency Medicine

The new approach based on Morlet wavelet spectrograms ridges analysis and allowing automatic detecting different activity in long term EEG signals is developed. To distinguish epileptiform activity from chewing artifacts two approaches are proposed. The quantitative characteristics of events wavelet spectrogram ridges were studied, as well as the frequency of broadband peaks at time points corresponding to peak-wave epileptiform activity on the one hand, and the peaks of myographic activity during chewing on the other hand. Signs by which one can qualitatively divide the group containing epileptic discharges from chewing artifacts were found.

The work was carried out within the framework of the state task and partially was supported by the Russian Foundation for Basic Research (project No. 18-29-02035).

[1] *Kershner I., Obukhov Yu., Sinkin M.* A new approach to the detection of epileptiform activity in EEG signals and methods to differentiate epileptic seizures from chewing artifacts // Rensit: Radioelectronics. Nanosystems. Information techologies, 2019. — Vol. 11, No. 2. — pp. 237-242.

О емкости семейств характеристических функций, обеспечивающих корректное решение задач диагностического типа

Забежайло Михаил Иванович[1]⋆ m.zabezhailo@yandex.ru

[1]Вычислительный центр им. А. А. Дородницына ФИЦ ИУ РАН

Обсуждаются возможности оценки достаточности данных в обучающей выборке - совокупности описаний прецедентов наличия анализируемого целевого эффекта (примеров) и описаний его отсутствия (контрпримеров) – для корректного решения задач диагностического типа («диагностики» по описанию нового прецедента наличия или, наоборот, отсутствия у него целевого эффекта, принимая во внимание опыт ранее изученных примеров и контрпримеров анализируемого эффекта). Особое внимание уделено анализу малых (не являющихся статистически значимыми) коллекций эмпирических данных. При этом предполагается, что анализируемый эффект имеет детерминистскую «причинную» природу, а в используемых описаниях прецедентов (языке представления знаний) отражаются те или иные необходимые для возникновения такого эффекта факторы влияния. Рассматриваются порождаемые исходной обучающей выборкой прецедентов значимые комбинации подобных факторов влияния (неподвижные точки некоторого замыкания Галуа, которое порождается сходствами описаний прецедентов, формализуемыми средствами бинарной алгебраической операции). С учетом выделяемых таким образом совокупностей факторов влияния строятся специального вида функции многозначной логики, каждая из которых на всех примерах обучающей выборки принимает значение «истина», а на всех контрпрпримерах этой же выборки – значение «ложь» (характеристические функции соответствующей обучающей выборки эмпирических данных). Непустота семейства таких функций для конкретной выборки прецедентов – критерий корректного (исключающего наличие «причин» в контрпримерах) разделения собранных в ней примеров и контрпримеров, и, как следствие, - аргумент в пользу надежности «диагностики» новых прецедентов, которая использует результаты обучения на данной выборке. Сохранение тех или иных функций из данного семейства при расширении обучающей выборки новыми прецедентами (примерами и контрпримерами) – дополнительный аргумент в использовании именно этих характеристических функций в дальнейших диагностических процедурах.

Демонстрируется экспоненциально быстрый (по отношению к линейному росту параметров исходной обучающей выборки) рост размеров семейства рассматриваемых характеристических функций. Доказана так называемая перечислительная полнота некоторых возникающих здесь переборных задач.

Возможности предлагаемого инструментария интеллектуального анализа данных иллюстрируются примерами его использования при решении ряда задач высокотехнологичной медицинской диагностики.

To the complexity of characteristic function sets providing correct diagnostic solutions

Michael Zabezhailo[1]⋆ m.zabezhailo@yandex.ru

[1]Moscow, FRCCSC of the Russian Academy of Sciences

The possibilities to evaluate the sufficiency of data in the training sample (TS) are discussed. Every TS is formed by examples (where analyzed effect is fixed) and counterexamples (where the absence of analyzed effect is fixed). Special attention is payed to the analysis of small (e.g. - statistically insignificant) TSs. The sufficiency of data in a concrete TS is necessary to provide correct diagnostics for new cases: are we able to extrapolate the existence (or, on the contrary, the absence) of the analyzed effect from TS to a new case (example or counterexample)? It's supposed that analyzed effect enforced by deterministic "causality": at least some enforcing "causality" factors are presented in TS-cases descriptions (i.e. are described by the used knowledge representation language). There are analyzed all significant combinations of such enforcing "causality" factors (corresponding to fix points of the Galois closure formed by the similarity of TS-cases descriptions that is formalized as a binary algebraic operation). Special functions (of many-valued logics) – so called characteristic functions (ChF) - are designed basing on such combinations of "causality" factors. Every ChF has truth-value "true" on every example and truth-value "false" for every counterexample from TS. Non-emptiness of ChF set is forming a criteria for correct (excluding the presence of any designed combination of "causality" factors enforcing the analyzed effect in all TS-counterexamples) extrapolation of TS-learned empirical dependencies to a new diagnostic case. (Correctness of this type extrapolation is based on a consistent separability of all examples and all counterexamples in TS made by learned ChFs. Stability (heritability) of some ChFs with respect to extensions of current TS by new cases provides additional important reason to implement these heritable ChFs to "diagnose" the analyzed effect for new cases.

It is demonstrated an exponentially high (with respect to linear-complex increase of TS) speed of growth for the set of all ChFs learned by concrete TS. So-called enumeration completeness (#P-completeness) of some ChF-related combinatorial problems is proven.

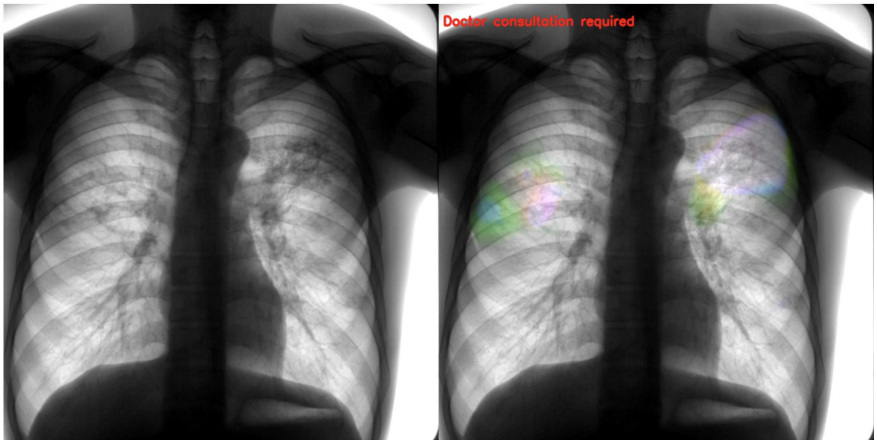
Some productive abilities of the presented tools of intelligent data analysis are illustrated by examples of their applications in high-technology medical diagnostics.

Особенности имплементации систем искусственного интеллекта в задаче анализа двухмерных радиологических изображений

Гогоберидзе Юрий Тенгизович[1]⋆ gut@vector.ru
Классен Виктор Иванович[1] klassen@vector.ru
Натензон Михаил Яковлевич[2] mnatenzon4@gmail.com
Просвиркин Илья Александрович[1] pia@vector.ru
Сафин Артем Альбертович[3] saal@vector.ru

[1]Чистополь, АО Радиокомпания "Вектор"
[2]Москва, Научно-производственное объединение "Национальное телемедицинское агентство"
[3]Чистополь, ООО "Фтизисбиомед"

Доклад посвящен первой в России научно-апробированной системе автоматизированного анализа флюорографических снимков, позволяющей выявлять легочные патологии различных типов. Система работает в режиме реального времени и способна производить анализы флюорографических изображений, полученных из любой точки земного шара. Результат анализа система выдает в виде модифицированного первоначального изображения с выделенными на нем участками с предполагаемым наличием патологий, а также рекомендацией обратить внимание врача-специалиста или же без выделенных участков и с заключением об отсутствии патологий на снимке. Важно отметить, что на снимках выделяются области с изменениями структуры легких, которые свидетельствуют о наличии патологий, как опасных для здоровья человека, так и неопасных.



Диагностическая точность системы оценена в 2018г. ГБУЗ г. Москвы «НПЦ медицинской радиологии ДЗМ». Оценка проводилась путем бинарной класси-

фикации изображений: «норма» или «патология». Описание дизайна и результатов выполнено в соответствии с методологией «STARD 2015». Оценка показала, что система обладает высокой прогностической ценностью отрицательного результата (82,4-97,5%).

Фундаментом представленной системы является ансамбль 3-х сверточных нейронных сетей (СНС) спроектированных по принципам архитектуры U-net (8 сворачивающих слоёв, 8 разворачивающих слоёв, 32 стартовых фильтра, каждый слой x1.5 фильтров на свёртке, x1.5 фильтров на разворачивании. Вход - 1 канал, выход - 1 канал). Ключевыми особенностями технического решения являются способ обучения и структура СНС. Для достижения технического результата применяются:

1. Размеченная специальным образом база (около 300000) флюорографических снимков для обучения с классификацией каждой области.
2. Перевзвешивание классов в соответствии с их важностью в выборке.
3. Комбинация выходных изображений для увеличения обучающей базы.

В работе рассмотрены вопросы построения архитектуры медицинского ИИ на базе сверточных нейронных сетей, вопросы подборки обучающей выборки и собственно обучения системы. Кроме того, затрагиваются вопросы точности системы, ее внедрения в медицинскую практику и повседневной эксплуатации специалистами.

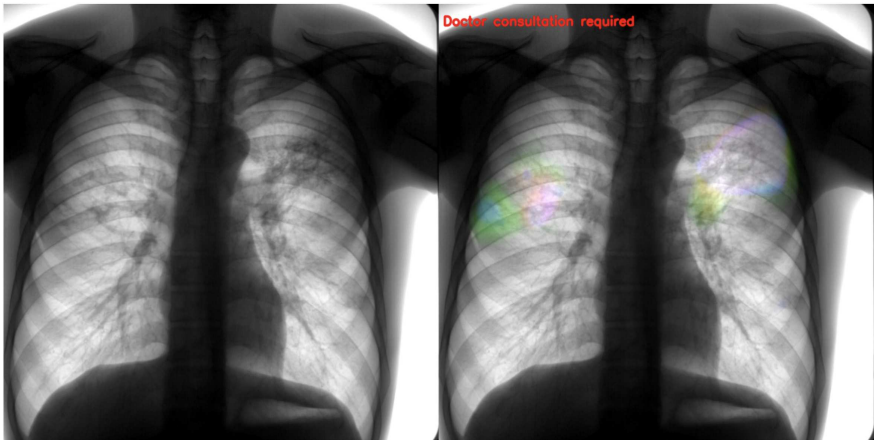
[1] *Klassen V. I., Safin A. A., Maltsev A. V., Andrianov N. G., Morozov S. P., Vladzymyrskyy A. V., Ledikhova N. V., Sokolina I. A., Kulberg N. S., Gombolevsky V. A, Kuzmina E. S.* AI-based screening of pulmonary tuberculosis: diagnostic accuracy // Journal of eHealth Technology and Application, Vol.16, No 1, 2018. P. 28 – 32.

Features of the implementation of artificial intelligence systems in the task of analysis of two-dimensional radiological images

Yuriy Gogoberidze[1]⋆ gut@vector.ru
Victor Klassen[1] klassen@vector.ru
Michael Natenzon[1] mnatenzon4@gmail.com
Ilya Prosvirkin[1] pia@vector.ru
Artem Safin[1] saal@vector.ru

[1]Chistopol, Vector Radio Company JSC
[2]Moscow, "National Telemedicine Agency" Research–and–Production Union
[3]Chistopol, "FtizisBioMed"

The report is devoted to the first in Russia scientifically tested system of automated analysis of fluorographic images, which allows to identify pulmonary pathologies of various types. The system operates in real time and is capable of analyzing fluorographic images obtained from anywhere in the world. The system gives the result of the analysis in the form of a modified initial image with selected areas on it with the alleged presence of pathologies, as well as a recommendation to consult with specialist doctor or without selected areas and with a conclusion about the absence of pathologies in the picture. It is important to note that the images highlight areas with changes in the structure of the lungs, which indicate the presence of pathologies, both dangerous to human health and non-dangerous.



The diagnostic accuracy of the system was evaluated in 2018 by the Moscow state medical radiology research center. The assessment was carried out by binary classification of images: "norm" or "pathology". The description of the design and results is carried out in accordance with the methodology "STARD 2015". The evaluation showed that the system has a high predictive value of a negative result (82.4- 97.5%).

The Foundation of the presented system is an ensemble of 3 convolutional neural networks (CNN) designed according to the principles of U-net architecture (8 convolutional layers, 8 deconvolutional layers, 32 starting filters, each layer x1.5 filters on convolution, x1.5 filters on deconvolution. Input - 1 channel, output - 1 channel). The key features of the technical solution are the method of training and the structure of the CNN. To achieve technical results are used:

1. Specially marked training set (about 300,000) of fluorographic images for training with the classification of each marked area.
2. Reweigh classes according to their importance in the sample.
3. Combination of output images to increase the training set.

The paper deals with the construction of the architecture of medical AI on the basis of convolutional neural networks, the selection of training samples and the actual training system. In addition, the issues of accuracy of the system, its implementation in medical practice and daily operation by specialists are discussed.

[1] *Klassen V. I., Safin A. A., Maltsev A. V., Andrianov N. G., Morozov S. P., Vladzymyrskyy A. V., Ledikhova N. V., Sokolina I. A., Kulberg N. S., Gombolevsky V. A, Kuzmina E. S.* AI-based screening of pulmonary tuberculosis: diagnostic accuracy // Journal of eHealth Technology and Application, Vol.16, No 1, 2018. P. 28 – 32.

Построение графовых нейронных сетей в задаче синтеза химических молекул

Никитин Филипп Александрович[1]⋆ filipp.nikitin@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

Решается задача прямого синтеза химических элементов. Требуется восстановить молекулярную структуру основного продукта химической реакции по молекулам исходных веществ. Данная задача является одной из ключевых для автоматизации процессов разработки лекарств и входит в более общую проблему ретросинтеза химических элементов.

Поставленная задача сводится к классификации атомов исходных молекул: определяется вероятность принадлежности атома основному продукту и вероятность атомов образовать центр реакции. Используется графовое представление молекул. Предложены модификации графовой свёрточной сети, повышающие качество работы модели для несвязанных графов, которым является набор молекулярных графов. Предложенная модель использует экспертных знания о структуре и особенностях молекулярного графа.

Качество решения задачи прямого синтеза измеряется по полному совпадению предсказанного и оригинального молекулярных графов основного продукта. Показана адекватность предложенной модели. Предложенная модель провалидирована на выборке реакций, собранных из паттентов США [2].

Работа выполнена при поддержке РФФИ (проекты 19-07-1155, 19-07-0875) и НТИ (проект 13/1251/2018).

[1] *Butler K. T. et al.* Machine learning for molecular and materials science // Nature, Nature. – 2018. – Т. 559. – №. 7715. – С. 547-555.

[2] *Lowe D. M.* Extraction of chemical structures and reactions from the literature. // PhD thesis – University of Cambridge, 2012.

Graph neural network learning for chemical compounds synthesis

Nikitin Filipp[1]⋆ filipp.nikitin@phystech.edu
Vadim Strijov[1] strijov@phystech.edu
[1]Moscow, Moscow Institute of Physics and Technology
[2]Moscow, FRCCSC of the Russian Academy of Sciences

The study is devoted to the problem of chemical compounds synthesis. It determines the molecular structure of the main product of a chemical reaction, molecules of the reagents and reactants. This problem is one of the most important problems in automation drug development processes. The solution to the problem is necessary for building the retrosynthesis path for a chemical compound.

Two node classification problems are solved in the work: the probability of the atom belonging to the main product and the probability of the atom forming the center of the reaction. Both probabilities are estimated for each atom. A graph representation of the molecules is used. Modifications of a graph convolutional network which increase the efficiency of the model on unrelated graphs are proposed. The proposed models use expert knowledge about the structure and features of molecular graphs.

The quality of the models is measured as the complete match of the predicted and original molecular graphs of the main product. The adequacy of the proposed models is shown. The proposed models are validated on a reaction dataset which was collected from US patents [2].

This research was supported by RFBR (projects 19-07-1155, 19-07-0875) and NTI (project 13/1251/2018).

[1] *Butler K. T. et al.* Machine learning for molecular and materials science // Nature, Nature. – 2018. – Т. 559. – №. 7715. – С. 547-555.

[2] *Lowe D. M.* Extraction of chemical structures and reactions from the literature. // PhD thesis – University of Cambridge, 2012.

Исследование признаков раннего паркинсонизма и эссенциального тремора в низкочастотном диапазоне 0.5–4 Гц 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
Карабанов Алексей Вячеславович[3] doctor.karabanov@mail.ru
Чигалейчик Лариса Анатольевна[3] chigalei4ick.lar@yandex.ru

[1]Москва, ИРЭ им. В.А. Котельникова РАН
[2]Москва, ИВНД РАН
[3]Москва, ФГБНУ «Научный центр неврологии»

Проведено исследование малоизученного частотного диапазона 0.5–4 Гц сигналов электромиограммы (ЭМГ) мышц у пациентов с болезнью Паркинсона (БП) и эссенциальным тремором (ЭТ). Для исследования был применён метод анализа всплескообразной электрической активности мышц, основанный на вейвлет-анализе и ROC-анализе. Идея метода заключается в поиске локальных максимумов («всплесков») на вейвлет-спектрограмме и вычислении различных характеристик, описывающих эти максимумы. Анализируется степень отличия группы пациентов с БП и ЭТ от контрольной группы в пространстве этих параметров. Для этого используется ROC-анализ. Исследуется функциональная зависимость AUC от значений границ диапазонов рассматриваемых параметров. Применение метода позволило выявить новые закономерности в ЭМГ-сигналах, которые ранее не удавалось выявить с помощью стандартных спектральных методов, основанных на анализе спектральной плотности мощности сигналов.

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта No. 18-37-20021.

[1] *Sushkova O. S., Morozov A. A., Gabova A. V., Karabanov A. V., Chigaleychik L. A.* An investigation of the 0.5–4 Hz low-frequency range in the wave train electrical activity of muscles in patients with Parkinson's disease and essential tremor. — RENSIT: Radioelectronics. Nanosystems. Information technologies, 2019. — V. 11, No. 2. — P. 225–236.

Investigation of features of early parkinsonism and essential tremor in the 0.5–4 Hz low-frequency range of wave train electrical activity of muscles

Olga Sushkova [1]⋆ o.sushkova@mail.ru
Alexei Morozov[1] morozov@cplire.ru
Alexandra Gabova[2] agabova@yandex.ru
Alexei Karabanov[3] doctor.karabanov@mail.ru
Larisa Chigaleychik [3] chigalei4ick.lar@yandex.ru

[1]Moscow, Kotel'nikov IRE RAS
[2]Moscow, IHNA&NPh RAS
[3]Moscow, FSBI "Research Center of Neurology"

An investigation of electromyograms (EMG) in the 0.5–4 Hz little-studied frequency range was performed in patients with Parkinson's disease (PD) and essential tremor (ET). A method for analyzing wave train electrical activity of the muscles based on the wavelet analysis and ROC analysis was used. The idea of the method is to find local maxima (that correspond to the wave trains) in the wavelet spectrogram and to calculate various characteristics describing these wave trains. The degree of difference of the group of patients from the control group of subjects is analyzed in the space of these parameters. ROC analysis is used for this purpose. The functional dependence of AUC on the values of the bounds of the ranges of the parameters under consideration is investigated. The application of the method allowed revealing new statistical regularities in EMG signals, which previously were not detected using standard methods based on the analysis of the power spectral density of the signals.

The reported study was funded by RFBR according to the research project No. 18-37-20021.

[1] *Sushkova O. S., Morozov A. A., Gabova A. V., Karabanov A. V., Chigaleychik L. A.* An investigation of the 0.5–4 Hz low-frequency range in the wave train electrical activity of muscles in patients with Parkinson's disease and essential tremor. — RENSIT: Radioelectronics. Nanosystems. Information technologies, 2019. — V. 11, No. 2. — P. 225–236.

Оценка межканальной фазовой синхронизации сигналов ЭЭГ в хребтах их вейвлет-спектрограмм у пациентов с черепно-мозговой травмой до и после реабилитации

Толмачева Рената Алексеевна[1]⋆ tolmatcheva@ya.ru
Обухов Юрий Владимирович[1] yuvobukhov@mail.ru
Жаворонкова Людмила Алексеевна[2] lzhavoronkova@hotmail.com
[1]Москва, Институт радиотехники и электроники им. В.А. Котельникова РАН
[2]Москва, Институт высшей нервной деятельности и нейрофизиологии РАН

В рамках разработанного нами нового подхода к оценке фазовой синхронизации сигналов электроэнцефалограмм в различных парах отведений выделены одинаковые межканальные связи у группы здоровых испытуемых при когнитивных и моторном тестах. Подход основан на вычислении и сравнении фаз в точках хребтов вейвлет-спектрограмм ЭЭГ. Установлены области интересов (пары фазово-связанных отведений) коры головного мозга при двух типах когнитивных и моторном тестах у группы контрольных здоровых испытуемых. Представлены результаты анализа межканальной фазовой синхронизации у испытуемых с черепно-мозговой травмой. Рассмотрены фазово-связанные пары каналов ЭЭГ, полученные по результатам ЭЭГ-записей пациентов с черепно-мозговой травмой, имевших повторные ЭЭГ-записи. Определение фазово-связанных пар сигналов ЭЭГ может быть полезно для мониторинга лечения людей с черепно-мозговой травмой. Восстановление фазово-связанных пар ЭЭГ может быть использовано в качестве индикатора правильного лечения.

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта No. 18-07-00609.

[1] *Tolmacheva R. A., Obukhov Y. V., Zhavoronkova L. A.* The estimation of inter-channel phase synchronization of EEG signals in the ridges of their wavelet spectrograms in patients with traumatic brain injury. — RENSIT: Radioelectronics. Nanosystems. Information technologies, 2019. — V. 11, No. 2. —P. 243–248.

The estimation of inter-channel phase synchronization of EEG signals in the ridges of their wavelet spectrograms in patients with traumatic brain injury before and post the rehabilitation

Renata Tolmacheva[1]⋆ tolmatcheva@ya.ru
Yury Obukhov[1] yuvobukhov@mail.ru
Ludmila Zhavoronkova[2] lzhavoronkova@hotmail.com
[1]Moscow, Kotelnikov Institute of Radioengineering and Electronics of RAS
[2]Moscow, Institute of Higher Nervous Activity and Neurophysiology of RAS

The new approach, which we developed to the estimation of phase synchronization of electroencephalogram (EEG) signals in different pairs of channels is proposed. Identical inter-channel phase coherency of EEG signals is determined for healthy subjects during cognitive and motor tests. EEG signal phase is evaluated at the points of its wavelet spectrogram ridge. Areas of interest of the cortex at cognitive and motor tests for group of healthy subjects are determined. The phase-coupled pairs of EEG channels obtained from results of EEG records of patients with traumatic brain injury who had repeated EEG records are considered. The definition of phase-coupled pairs of EEG signals can be useful for the monitoring the treatment of people with traumatic brain injury. The recovery of phase-coupled EEG pairs can be used as indicator of proper treatment.

The reported study was funded by RFBR according to the research project No. 18-07-00609.

[1] *Tolmacheva R. A., Obukhov Y. V., Zhavoronkova L. A.* The estimation of inter-channel phase synchronization of EEG signals in the ridges of their wavelet spectrograms in patients with traumatic brain injury. — RENSIT: Radioelectronics. Nanosystems. Information technologies, 2019. — V. 11, No. 2. —P. 243–248.

Оценка направлений движения магнитных наночастиц методом функциональной томографии

Устинин Михаил Николаевич[1]⋆ u_m_n@mail.ru
Рыкунов Станислав Дмитриевич[1] rykunov@impb.ru
Бойко Анна Ивановна[1] a.boyko@list.ru
[1]Пущино, ИМПБ РАН - филиал ИПМ им. М.В. Келдыша РАН

Создан метод функциональной томографии для изучения электромагнитной активности различных сложных систем с помощью измерения производимых ими магнитных полей. По многоканальным магнитным измерениям строится функциональная томограмма, отображающая информацию, содержащуюся во временных рядах, на пространство эксперимента. Это достигается с помощью решения обратной задачи для всех элементарных осцилляций, выделенных преобразованием Фурье. Для каждой частоты определяется узел трехмерной сетки, в котором расположен ее источник. Также возможно оценить доминирующее направление источников в этом узле. В данной работе метод использовался для оценки направлений движения магнитных наночастиц. Все существующие методы визуализации магнитных наночастиц в биологических объектах основаны на использовании внешних магнитных полей, которые могут значительно изменить пространственное положение и свойства изучаемого наномагнитного ансамбля. В работе использовался СКВИД-магнетометр для измерения магнитного шума, генерируемого суперпарамагнитными наночастицами в стационарном сосуде без приложения внешнего магнитного поля. Было показано, что феррожидкость генерирует спонтанные магнитные поля, достаточные для локализации предложенным методом функциональной томографии. Было найдено, что возможно использовать модель стационарного магнитного диполя с переменным током для описания механического движения магнита, образованного в феррожидкости магнитным полем Земли. Физическая причина возможности такого моделирования состоит в том, что СКВИД-сенсоры регистрируют не само поле магнита, а только его изменения. Оказалось, что спонтанные магнитные поля в определенной полосе частот обладают сильной анизотропией. Найденный эффект может существенно повысить пространственное разрешение предложенного метода визуализации магнитных наночастиц в биологических объектах без использования внешнего магнитного поля.

Исследование выполнено за счет гранта Российского научного фонда (проект № 18-11-00178).

[1] *Polikarpov M. A., Ustinin M. N., Rykunov S. D., Yurenya A. Y., Naurzakov S. P., Grebenkin A. P., Panchenko V. Y.* Study of anisotropy of magnetic noise, generated by magnetic particles in geomagnetic field // Study of anisotropy of magnetic noise, generated by magnetic particles in geomagnetic field, 2019. — V. 475— p. 620–626.

Estimation of the movement directions of magnetic nanoparticles by the method of functional tomography

Stanislav Rykunov[1]⋆ rykunov@impb.ru
Mikhail Ustinin[1] u_m_n@mail.ru
Anna Boyko[1] a.boyko@list.ru
Natalia Pankratova[1] natpan1974@mail.ru

[1]Pushchino, IMPB RAS - Branch of KIAM RAS

The method of functional tomography for studying the electromagnetic activity of various complex systems using external measurements of the magnetic field was developed. From the multichannel magnetic data, a functional tomogram is calculated, projecting onto the space of experiment the information contained in time series. This is achieved by solving the inverse problem for all the elementary oscillations that are separated by the Fourier transform. For each oscillation frequency the node of a three-dimensional grid is found, where the source is located. Also, it is possible to estimate the dominant direction of the sources in this node. In this study the method was used to estimate directions of the magnetic nanoparticles movement. Any existing method of visualization of magnetic nanoparticles in biological objects is based on using of external magnetic field, which can considerably change the spatial distribution and properties of the nanomagnetic ensemble under study. In our work a SQUID-based magnetometry device was used for the measurement of a magnetic noise generated by superparamagnetic nanoparticles in the stationary standing vial without imposing of an external magnetic field. It was demonstrated that the ferrofluid generates spontaneous magnetic fields sufficient for its localization by the proposed method of functional tomography. It was shown, that it is possible to use static magnetic dipole with alternating current as a model for the description of mechanical movement of the magnet, formed in ferrofluid by the geomagnetic field. Physical reason for such modeling lays in the fact, that SQUID sensors register not the field of this magnet, but only the changes of the field. It was revealed that the spontaneous magnetic fields in certain frequency band have a strong spatial anisotropy. The detected effect can essentially increase the spatial resolution of the proposed method of visualization of magnetic nanoparticles in biological objects without using the external magnetic field.

The research was supported by the Russian Science Foundation (grant 18-11-00178).

[1] *Polikarpov M. A., Ustinin M. N., Rykunov S. D., Yurenya A. Y., Naurzakov S. P., Grebenkin A. P., Panchenko V. Y.* Study of anisotropy of magnetic noise, generated by magnetic particles in geomagnetic field // Study of anisotropy of magnetic noise, generated by magnetic particles in geomagnetic field, 2019. — V. 475— p. 620–626.

Эксперименты с нейросетевой классификацией суб-терагерцовых изображений скрытого под одеждой оружия и других опасных предметов

Морозов Алексей Александрович ⋆ morozov@cplire.ru
Сушкова Ольга Сергеевна o.sushkova@mail.ru
Кершнер Иван Андреевич ivan_kershner@mail.ru
Москва, ИРЭ им. В.А. Котельникова РАН

Проведены эксперименты с нейросетевой классификацией суб-терагерцовых изображений оружия и других опасных предметов, скрытых под одеждой человека. Цель экспериментов — выяснить, содержит ли терагерцовое видеоизображение достаточное количество информации, чтобы научить нейросеть различать опасные и неопасные предметы. Обучающая выборка включала изображения людей в домашней и уличной одежде. Под одеждой были спрятаны оружие и опасные предметы, такие как автомат Калашникова, топор, бутылки, нож, резиновая дубинка и пистолеты различных марок, а также обычные бытовые предметы, такие как телефоны и USB-диски. Нейронные сети, обученные на этом множестве данных, были применены для анализа другого множества данных, включающего винтовку M16. Нейронные сети сумели успешно определить, что винтовка является опасным предметом. Результаты экспериментов показывают, что нейронные сети могут быть использованы для обобщения свойств терагерцовых видеоизображений и могут успешно предсказывать, является ли скрытый под одеждой предмет опасным.

Для экспериментов с терагерцовым и многомодальным видеонаблюдением были разработаны специальные средства логического программирования, включающие набор встроенных классов языка Акторный Пролог для получения, записи и чтения терагерцовых, тепловых, визуальных и 3D видеоизображений.

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 16-29-09626-офи-м (www.fullvision.ru).

[1] *Morozov A. A., Sushkova O. S., Kershner I. A., Polupanov A. F.* Development of a method of terahertz intelligent video surveillance based on the semantic fusion of terahertz and 3D video images. — CEUR, 2019. — Vol. 2391, Paper 19.

On experiments with the neural-network-based classification of sub-terahertz images of concealed weapons and other dangerous objects

Alexei Morozov ⋆ morozov@cplire.ru
Olga Sushkova o.sushkova@mail.ru
Ivan Kershner ivan_kershner@mail.ru
Moscow, Kotel'nikov IRE RAS

The experiments with the neural-network-based classification of sub-terahertz images of concealed weapons and other dangerous objects are described. The goal of the experiments is to check whether the terahertz videos contain information enough to teach a convolutional network to distinguish the dangerous and safe objects. The learning set includes terahertz images of people dressed in home clothes and outer clothing. The set of hidden objects include the Kalashnikov sub-machine-gun, an ax, bottles, a knife, a baton, and handguns of different brands. Some images contain ordinary objects like smartphones and USB disks. Convolutional networks of several standard architectures were trained using the data sets. After that, an additional test data set that includes only the images of a person that keeps the M16 automatic rifle and the images of the person without extra objects was prepared. Then, the trained networks were used to analyze the video images. The networks recognized successfully the M16 automatic rifle as a dangerous object. These results demonstrate that the neural network can make generalizations of the terahertz images of hidden objects and successfully predict that the hidden object is a kind of a weapon and/or dangerous object.

Special logic programming means were developed for the experimenting with the terahertz/multimodal video surveillance including a set of built-in classes of the Actor Prolog language for terahertz, thermal, RGB, and 3D video data acquisition, writing, and reading.

This research was supported by the Russian Foundation for Basic Research, project 16-29-09626-ofi-m (www.fullvision.ru).

[1] *Morozov A. A., Sushkova O. S., Kershner I. A., Polupanov A. F.* Development of a method of terahertz intelligent video surveillance based on the semantic fusion of terahertz and 3D video images. — CEUR, 2019. — Vol. 2391, Paper 19.

Об одном подходе к статистическому моделированию транспортных потоков

Старожилец Всеволод Михайлович[1]⋆ starvsevol@gmail.com
Чехович Юрий Викторович[1] chehovich@forecsys.ru
[1]Москва, Вычислительный центр им. А. А. Дородницына ФИЦ ИУ РАН

Моделирование транспортных потоков зачастую основано на их сходстве с жидкой или газовой средой. В частности, базовая модель Лайтхилла–Уизема–Ричардса (Lighthill–Whitham–Richards, LWR) [1, 2] основана на предположении о существовании взаимно-однозначной зависимости между скоростью и плотностью потока автомобильно-транспортных средств(АТС) и сохранении числа АТС в транспортной сети. В современном макроскопическом подходе транспортный поток описывается нелинейной системой гиперболических дифференциальных уравнений в частных производных второго порядка в различных постановках [3, 4, 5]. Исследователи также пытаются учесть тот факт, что поток автомобилей состоит из различных типов транспортных средств [6, 7]: легковых автомобилей, мотоциклов, грузовых автомобилей — каждый из которых обладает отличающимися друг от друга характеристиками.

Все вышеизложенные подходы, однако, вычислительно сложны. По этой причине мы пошли по пути упрощения, предположив что автомобили в транспортной сети в среднем обладают одинаковыми характеристиками и могут быть объединены в относительно однородные группы АТС скорость всех транспортных средств в которой одинакова.

В работе предлагается статистическая модель транспортных потоков для моделирования движения транспортных средств на автомагистралях идентифицируемая на данных из гетерогенных источников. Модель симулирует движение групп транспортных средств по магистрали с использованием фундаментальной диаграммы для определения ключевых характеристик потока автомобильно-транспортных средств.

Группы АТС в рамках предложенной модели могут разделяться и объединяться, но такие изменения происходят относительно редко. Следует отметить, что авторы рассматривают задачу моделирования движения АТС по автомагистрали, где возможно пересечение дорог в одном уровне. В рамках рассматриваемой задачи сделанные предположения считаем достаточно разумными.

Для определения скорости групп АТС и значений оптимального потока на рассматриваемом участке автомобильно-транспортной сети используется фундаментальная диаграмма на ней. Динамические габариты же автомобилей считаются линейно зависящими от скорости.

Расчёт по предложенной модели проводится на графе отождествленном с некоторым участком транспортной сети по ветвям которого движутся группы АТС. Состояние системы в каждый момент времени определяется совокупностью скорости, положения и размера каждой группы АТС на каждой ветви.

Сами же расчёты проводимые для перехода между двумя состояниями сводится с трем процедурам: перемещения групп АТС по ветви, перемещения групп АТС между ветвями и объединение двух групп АТС.

Хотя модель использует довольно грубые приближения из-за использования группы АТС как базовой единицы моделирования, в работе показано, что модель показывает адекватные результаты при любых режимах автомагистрали. Проводятся несколько вычислительных эксперимента на синтетических данных для подтверждения работоспособности модели симулирующих как возникновения заторов в следствие аварий на автомагистрали, так и проблемы связанные с недостаточной пропускной способностью перекрестков.

Работа поддержана грантом РФФИ №17-07-01574.

[1] *Lighthill M.J., Whitham G.B.* On kinematic waves. II. A theory of traffic flow on long crowded roads // P. Roy. Soc. Lond. A Mat., 1955. Vol. 229. P. 317–345. doi: 10.1098/rspa.1955.0089.

[2] *Whitham J.B.* Linear and nonlinear waves. — Wiley, 1974. 656 p.

[3] *Daganzo C.F.* Requiem for second-order fluid approximations of traffic flow // Transport. Res. B Meth., 1995. Vol. 29. No. 4. P. 277–286. doi: 10.1016/0191-2615(95)00007-Z.

[4] *Siebel F., Mauser W.* On the fundamental diagram of traffic flow // SIAM J. Appl. Math., 2006. Vol. 66. No. 4. P. 1150–1162. doi: 10.1137/050627113.

[5] *Siebel F., Mauser W.* Synchronized flow and wide moving jams from balanced vehicular traffic // Phys. Rev. E, 2006. Vol. 73. No. 6. P. 066108. doi: 10.1103/PhysRevE.73.066108.

[6] *Dey P.P., Chandra S., Gangopadhyay S.* Simulation of mixed traffic flow on two-lane roads // Journal of Transportation Engineering, 2008. Vol. 134. No. 9. P. 361–369.

[7] *Lan L.W., Chang C. -W., Gangopadhyay S.* Inhomogeneous cellular automata modeling for mixed traffic with cars and motorcycles // Journal of advanced transportation, 2005. Vol. 39. No. 3. P. 323–349.

About one approach to traffic flows statistical modeling

Vsevolod Starozhilets[1]⋆ starvsevol@gmail.com
Yuriy Chehovich[1] chehovich@forecsys.ru
[1]Dorodnicyn Computing Centre FRC CSC RAS, Vavilov st. 40, 119333 Moscow, Russia

Modeling of traffic flows is often based on their similarity with a liquid or gas. In particular, the basic Lighthill–Witham–Richards (LWR) model [1, 2] is based on assumption that there are one-to-one relationship between speed and density of cars and conservation of its number in transport network. In a modern macroscopic approach, transport flow is described by a nonlinear system of second-order hyperbolic partial differential equations in various statements [3, 4, 5]. Researchers are also trying to take into account the fact that traffic flow consists of various types of vehicles [6, 7]: cars, motorcycles, trucks — each has different characteristics.

All the above approaches, however, are computationally complex. For this reason, we took the path of simplification, assuming that cars in transport network, on average, have the same characteristics and can be combined into groups, speed of all vehicles in which are the same.

In this paper we propose a statistical model of traffic flows for modeling speed and number of cars on highways identified on data from heterogeneous sources. The model simulates movement of car groups along the highway using fundamental diagram to determine key characteristics of traffic flow.

Vehicle groups in this model can be united and divided, but such events are relatively rare. It should be noted that the authors consider simulation of highway with only crossing of roads at the same level is possible. Within the boundaries of studied problem we consider assumptions that are made are reasonable enough.

We use a fundamental diagram to determine the speed of groups and optimal flow values on the considered section of transport network. Dynamic length of cars is considered linearly dependent on speed.

First of all, we identify simulated part of road with graph on the branches of which cars groups are moved. State of system in every moment of time is determined by combination of speed, position and size of each group on all branches. In order to calculate current system state based on previous one we determine three procedures: moving cars groups along the branches, moving groups between the branches and union of two groups.

Although the model uses rather rough approximations due to use of cars groups as the basic unit of simulation, it will be shown that the model presents adequate results in all transport conditions. Several computational experiments are conducted on automatically generated data to confirm the performance of the model, which stimulate both traffic jams as a result of accidents on the highway, and problems associated with insufficient traffic capacity of crossroads.

This research is funded by RFBR, grant 17-07-01574.

[1] *Lighthill M.J., Whitham G.B.* On kinematic waves. II. A theory of traffic flow on long crowded roads // P. Roy. Soc. Lond. A Mat., 1955. Vol. 229. P. 317–345. doi: 10.1098/rspa.1955.0089.

[2] *Whitham J.B.* Linear and nonlinear waves. — Wiley, 1974. 656 p.

[3] *Daganzo C. F.* Requiem for second-order fluid approximations of traffic flow // Transport. Res. B Meth., 1995. Vol. 29. No. 4. P. 277–286. doi: 10.1016/0191-2615(95)00007-Z.

[4] *Siebel F., Mauser W.* On the fundamental diagram of traffic flow // SIAM J. Appl. Math., 2006. Vol. 66. No. 4. P. 1150–1162. doi: 10.1137/050627113.

[5] *Siebel F., Mauser W.* Synchronized flow and wide moving jams from balanced vehicular traffic // Phys. Rev. E, 2006. Vol. 73. No. 6. P. 066108. doi: 10.1103/PhysRevE.73.066108.

[6] *Dey P.P., Chandra S., Gangopadhyay S.* Simulation of mixed traffic flow on two-lane roads // Journal of Transportation Engineering, 2008. Vol. 134. No. 9. P. 361–369.

[7] *Lan L.W., Chang C. -W., Gangopadhyay S.* Inhomogeneous cellular automata modeling for mixed traffic with cars and motorcycles // Journal of advanced transportation, 2005. Vol. 39. No. 3. P. 323–349.

Моделирование процесса организации грузоперевозок

Бекларян Лева Андреевич[1] beklar@cemi.rssi.ru
Хачатрян Нерсес Карленович[2]⋆ nerses@cemi.rssi.ru
Акопов Андраник Сумбатович[3] akopovas@umail.ru
[1]Москва, ЦЭМИ РАН
[2]Москва, ЦЭМИ РАН
[3]Москва, ЦЭМИ РАН

Изучается модель организации железнодорожных грузоперевозок на протяженном участке пути с большим количеством промежуточных станций и расположенных между ними перегонов для временного хранения части грузов. Предполагается, что между произвольными соседними станциями существует межстанционный перегон, где временно может храниться часть грузов. Такая модель позволяет прогнозировать динамику загруженности станций и потоков возникающих в транспортной сети, при заданной процедуре организации грузопотока. Сама процедура организации грузопотока использует две технологии, единые для всех станций. Первая технология основана на взаимодействии соседних станций и формируется по определенному правилу. Согласно этому правилу, каждая из станций должна принимать груз с предыдущей станции, если количество задействованных путей на ней меньше чем на предыдущей станции, и отправлять на следующую станцию, если количество задействованных путей на ней больше чем на следующей станции. При этом как интенсивность приема, так и интенсивность отправки грузов пропорциональна разности чисел задействованных путей на соседних станциях. Вторая технология призвана использовать инфраструктурные возможности станций и обеспечить бесперебойное движение грузов. Она основана на взаимодействии станций с соседними перегонами. Организация грузоперевозок включает в себя систему контроля, которая заключается в том, что объемы грузов на соседних станциях должны совпадать с единым лагом времени.

Рассматривается два типа моделей: первый тип моделей описывает динамику загруженности станций при условии неограниченности емкостей перегонов, соответственно, второй - при их ограниченности. Каждый из двух типов моделей представлен двумя моделями: модель организации грузоперевозок между двумя узловыми станциями и модель организации грузоперевозок по замкнутой цепочке станций. Для указанных моделей описаны режимы грузоперевозок, исследована из зависимость от параметров модели, начального состояния системы. Стационарные режимы были исследованы на устойчивость.

Реальный процесс организации грузоперевозок может подвергнуться воздействию случайных факторов различной природы. В связи с этим возникает естественный вопрос: насколько режимы грузоперевозок указанных моделей устойчивы к случайным воздействиям? Для ответа на данный вопрос указанные вы-

ше модели были модифицированы с учетом влияния на процесс грузоперевозок случайных воздействий.

Работа поддержана грантом РФФИ № 19-01-00147.

[1] *Бекларян Л. А., Хачатрян Н. К., Акопов А. С.* Моделирование процесса организации грузоперевозок // Машинное обучения и анализ данных, 2019.

Modeling of cargo transportation organization process

Leva Beklaryan[1] beklar@cemi.rssi.ru
Nerses Khachatryan[2]⋆ nerses@cemi.rssi.ru
Andranik Akopov[3] akopovas@umail.ru
[1]Moscow, CEMI RAS
[2]Moscow, CEMI RAS
[3]Moscow, CEMI RAS

The model for organizing railway transportation on a long stretch of road with a large number of intermediate stations and railway tracks located between them is investigated. It is assumed that between arbitrary stations there is interexchange railway track, where part of the cargo can be temporarily stored. Such model allows to predict dynamics of load of the stations and flows arising in transport network at the set procedure of the movement of freight traffic. The procedure of the movement of freight traffic uses two technologies uniform for all stations. The first technology is based on interaction of the neighboring stations and is formed by a certain rule. According to this rule, each of stations has to take the cargo from the previous station if the quantity of the involved roads on it are less than at the previous station, and to send on the next station if the quantity of the involved roads on it are more than at the next station. In this case, both the intensity of reception, and intensity of shipment cargo is proportional to the difference of numbers of the involved roads at the neighboring stations. The second technology uses technical capabilities of the station and is based on interaction of the station with the neighboring railway tracks. The organization of cargo transportation includes a control system, which is that the volumes of cargo at neighboring stations should coincide with a single period of time.

Two types of models are considered: the first type of models describes the dynamics of stations load under the condition of unlimited capacities of railway tracks, respectively, the second-with their limitations. Each of the two types of models is represented by two models: a model of organization of cargo transportation between two nodal stations and a model of organization of cargo transportation in a closed chain of stations. The modes of cargo transportation are described for these models, the dependence on the model parameters and the initial state of the system is studied. Stationary regimes were investigated for stability.

The real process of organization of cargo transportation may be affected by random factors of different nature. In this regard, there is a natural question: how modes of transportation of these models are resistant to random influences? To answer this question, the above models have been modified to take into account the influence on the process of transportation of random influences.

This research is funded by RFBR, grant 19-01-00147.

[1] *Beklaryan L., Khachatryan N., Akopov A.* Modeling of cargo transportation organization process // Machine Learning and Data Analysis, 2019.

Модель «кочевников» и «землепашцев» с учетом ограничений на перемещения агентов по ареалу

Бекларян Лева Андреевич[1] beklar@cemi.rssi.ru
Белоусов Федор Анатольевич[2]⋆ belousovfedor@gmail.com
[1]Москва, ЦЭМИ РАН
[2]Москва, ЦЭМИ РАН

Рассматривается модель некоторого сообщества с простой социальной структурой, в которой агенты разделены на два типа – «кочевников» и «землепашцев», каждый из которых отличается своим отношением к способу производства продукта. Если землепашцы умеют самостоятельно воспроизводить продукт, то кочевники такими навыками не наделены, вместо этого они обладают способностью находить такой продукт на ареале, в том числе и отнимая его у землепашцев. Изучается вопрос как вымирания одного из рассматриваемых сообществ, так и вопрос сосуществования этих сообществ на едином пространстве на протяжении наблюдаемого периода времени.

В рамках данного исследования дополнительно введено ограничение на перемещение агентов, которое состоит в том, что «кочевники» на протяжении всей своей жизни не могут уходить от места своего рождения дальше некоторого экзогенно заданного расстояния. Значения данного экзогенного параметра было обозначено через radius. Для такой модификации проведено множество экспериментов, на основе которых осуществлен анализ параметра radius и выявлено при каких его значениях наблюдается та или иная качественная динамика всей популяционной системы.

Длительность каждого эксперимента составила 15000 периодов. Проведена вариация параметра radius в пределах от 4 до 16 с шагом 1, также проведено сравнение полученных результатов со случаем, когда никаких ограничений на перемещения агентов нет. Для каждого из значений параметра radius выявлены процентные соотношения тех случаев, когда выживают только «землепашцы», когда выживают только «кочевники», а также процентное соотношение случаев, при которых выживают оба сообщества.

Полученные результаты согласуются с интуицией, а именно, чем в большей степени кочевники ограничены в перемещениях, тем меньше шансов для них выжить как виду.

Работа поддержана грантом РФФИ № 19-01-00147.

[1] *Бекларян Л. А., Белоусов Ф. А.* Модель «кочевников» и «землепашцев» с учетом ограничений на перемещения агентов по ареалу // Машинное обучения и анализ данных, 2019.

Model of nomads and plowmen with restrictions on the movement of agents in the area

Leva Beklaryan[1] beklar@cemi.rssi.ru
Fedor Belousov[2]⋆ belousovfedor@gmail.com
[1]Moscow, CEMI RAS
[2]Moscow, CEMI RAS

The model with the simple social structure is considered. In the model agents are divided into two types - nomads and plowmen, each of which is distinguished by its attitude to the method of production of the product. If plowmen can independently reproduce product, then nomads are not endowed with such skills, instead they have the ability to find such a product in the area, including taking it away from plowmen. The question of extinction of one of the communities is studied, and the question of the coexistence of these communities in a single space during the observed period of time also is studied.

In this study, an additional restriction on the movement of agents is introduced, which is that nomads throughout their lives are not able to move away from their place of birth beyond some exogenously specified distance. The values of this exogenous parameter were denoted by radius. For this modification, many experiments were carried out, on the basis of which the analysis of the radius parameter was carried out and it was revealed at what values one or another qualitative dynamics of the entire population system was observed.

The duration of each experiment was 15,000 periods. A variation of the radius parameter in the range from 4 to 16 with a step of 1 is carried out, and the results are compared with the case when there are no restrictions on the movement of agents. For each of the values of the radius parameter, the percentages of cases in which only plowmen survive, when only nomads survive, and the percentage of cases in which both communities survive are revealed.

The results are consistent with intuition, namely, the more restricted nomads are in their movements, the less likely they are to survive as a species.

This research is funded by RFBR, grant 19-01-00147.

[1] *Beklaryan L., Belousov F.* Model of nomads and plowmen with restrictions on the movement of agents in the area // Machine Learning and Data Analysis, 2019.

Трехмерные морфометрические модели рельефа дна Северного Ледовитого океана

Флоринский Игорь Васильевич[1]⋆ iflor@mail.ru
Филиппов Сергей Валерьевич[1] fsv141@mail.ru
[1]Пущино, ИМПБ РАН – филиал ИПМ им. М.В. Келдыша РАН

В работе представлены результаты первого этапа выполнения проекта по созданию системы трехмерного (3D) геоморфометрического моделирования подводного рельефа Северного Ледовитого океана [1]. Разработана тестовая настольная версия системы. В качестве исходных данных использована ЦМР низкого разрешения (шаг сетки 5 км), выделенная из International Bathymetric Chart of the Arctic Ocean (IBCAO) version 3.0. По сглаженной ЦМР рассчитаны следующие морфометрические величины: горизонтальная, вертикальная, минимальная и максимальная кривизна, водосборная и дисперсивная площадь, а также индекс мощности потоков. Для создания и визуализации 3D моделей применена разработанная нами ранее методика 3D моделирования рельефа в среде пакета Blender. Методика включает в себя следующие этапы: 1) Создание полигонального объекта; 2) Выбор алгоритма моделирования 3D геометрии; 3) Выбор вертикального преувеличения масштаба; 4) Выбор типа, параметров, количества и расположения виртуальных источников освещения; 5) Выбор методов моделирования теней; 6) Выбор метода шейдинга 3D модели; 7) Выбор материала поверхности 3D модели; 8) Драпировка 3D модели текстурами; 9) Выбор количества и расположения виртуальных камер; 10) Рендеринг 3D модели. Представлена серия 3D морфометрических моделей с ракурсами со стороны Атлантики, Евразии, Тихого океана и Северной Америки. Использованный подход работоспособен и может служить основой для создания следующих версий системы – настольной и онлайновой, которые обеспечат работу с моделями высокого разрешения.

Работа поддержана грантами РФФИ №18-07-00223 и 18-07-00354.

[1] *Florinsky I. V., Filippov S. V.* Three-dimensional desktop morphometric models for the Arctic Ocean floor // Proceedings of the International Cartographic Association, 2019. — Vol. 2. — 32. — doi:10.5194/ica-proc-2-32-2019.

Three-dimensional morphometric models of the Arctic Ocean submarine topography

Florinsky Igor[1]⋆ iflor@mail.ru
Filippov Sergei[1] fsv141@mail.ru
[1]Pushchino, IMPB KIAM RAS

We present results of the first phase of an ongoing project to create a system for three-dimensional (3D) geomorphometric modeling of the Arctic Ocean submarine topography [1]. In this phase, we developed a testing, low-resolution desktop version of the system. We utilized a small, 5-km gridded digital elevation model (DEM) extracted from the International Bathymetric Chart of the Arctic Ocean (IBCAO) version 3.0. From the smoothed DEM, we derived digital models of several morphometric variables: horizontal, vertical, minimal, and maximal curvatures, catchment and dispersive areas, as well as stream power index. To construct and visualize 3D morphometric models, we applied an original approach for 3D terrain modeling in the environment of the Blender package. The approach includes the following main steps: (1) Automatic creating a polygonal object from a DEM; (2) Selecting an algorithm to model the 3D geometry; (3) Selecting a vertical exaggeration scale; (4) Selecting types, parameters, a number, and positions of light sources; (5) Selecting methods for generating shadows; (6) Selecting a shading method for the 3D model; (7) Selecting a material for the 3D model surface; (8) Overlaying a texture on the 3D model; (9) Setting a virtual camera(s); and (10) Rendering the 3D model. Finally, we present a series of 3D morphometric models with perspective views from the Atlantic, Eurasia, the Pacific, and North America. The experiment showed that the approach is efficient and can be used for creating next, desktop and web versions of the system for visualizing 3D morphometric models of higher resolutions.

This research is funded by RFBR, grants 18-07-00223 and 18-07-00354.

[1] *Florinsky I. V., Filippov S. V.* Three-dimensional desktop morphometric models for the Arctic Ocean floor // Proceedings of the International Cartographic Association, 2019. — Vol. 2. — 32. — doi:10.5194/ica-proc-2-32-2019.

Качество прогнозирования потока релятивистский электронов на геостационарной орбите с помощью метдов машинного обучения

Ефиторов Александр Олегович[1,2] a.efitorov@sinp.msu.ru
Широкий Владимир Романович[1,2]⋆ shiroky@sinp.msu.ru
Мягкова Ирина Николаевна[1,2] irina@sinp.msu.ru
Доленко Сергей Анатольевич[1,2] dolenko@sinp.msu.ru

[1]Московский государственный университет им. М. В. Ломоносова
[2]Научно-исследовательский институт ядерной физики имени Д.В.Скобельцына

Представлены результаты прогнозирования максимальных за сутки среднечасовых значений потока релятивистских электронов (E>2 МэВ) во внешнем радиационном поясе Земли на 1-3 суток вперед. В качестве входных признаков для прогнозирования использовались значения геомагнитных индексов, индукции межпланетного магнитного поля, скорости солнечного ветра, плотности протонов, ультра-низкочастотного индекса волновой активности (ULF) и среднечасовые значения потока релятивистских электронов. Фазовое пространство динамической системы было реконструировано путем погружения временного ряда каждой физической величины. Для прогнозирования использовались следующие модели машинного обучения: многомерная автогрегрессионная модель, ансамбли деревьев решений в рамках беггинг-подхода, искусственные нейронные сети типа многослойный персептрон. Результаты сравниваются с аналогичными, полученными другими авторами. Показано, что при решении данной задачи наилучший результат дают ансамбли деревьев решений. Также показано, что использование глубины погружения каждой компоненты временного ряда, рассчитанной по спаду автокорреляционной функции, значительно улучшает качество прогнозирования на горизонте прогноза в 1 час.

Исследование выполнено за счет гранта РНФ № 16-17-00098.

[1] *Irina Myagkova, Alexander Efitorov, Vladimir Shiroky, Sergey Dolenko* Quality of Prediction of Daily Relativistic Electrons Flux at Geostationary Orbit by Machine Learning Methods // Lecture Notes in Computer Science, V.11730. Springer Nature, 2019. — P. 556–565.

Quality of Prediction of Daily Relativistic Electrons Flux at Geostationary Orbit by Machine Learning Methods

Alexander Efitorov[1,2] a.efitorov@sinp.msu.ru
Vladimir Shiroky[1,2]⋆ shiroky@sinp.msu.ru
Irina Myagkova[1,2] irina@sinp.msu.ru
Sergey Dolenko[1,2] dolenko@sinp.msu.ru
[1]Moscow, M. V. Lomonosov Moscow State University
[2]Moscow, D. V. Skobeltsyn Institute of Nuclear Physics

This study presents the results of prediction 1-3 days ahead for the daily maximum of hourly average values of relativistic electrons flux (E>2 MeV) in the outer radiation belt of the Earth. The input physical variables were geomagnetic indexes, interplanetary magnetic field, solar wind velocity and proton density, special ultra-low frequency (ULF) indexes and hourly average values of relativistic electron flux. The phase-space for each physical component was reconstructed by time delay vectors with their own different embedding dimensions, and all of these vectors were concatenated. Next, various adaptive models were trained on this multivariate dataset. The following models were used for prediction: multi-dimensional autoregressive model, ensembles of decision trees within bagging approach, artificial neural networks of multi-layer perceptron type. The obtained results are analyzed and compared to the results of similar predictions by other authors. The best prediction quality was demonstrated by ensembles of decision trees. Also it has been demonstrated that using embedding depth based on autocorrelation function significantly improves prediction quality for one day prediction horizon.

This study conducted at the expense of RSF, grant 16-17-00098.

[1] *Irina Myagkova, Alexander Efitorov, Vladimir Shiroky, Sergey Dolenko* Quality of Prediction of Daily Relativistic Electrons Flux at Geostationary Orbit by Machine Learning Methods // Lecture Notes in Computer Science, V.11730. Springer Nature, 2019. — P. 556–565.

Разработка алгоритма позиционирования мобильного устройства на основе сенсорных сетей из BLE-маяков для построения систем автономной навигации

Астафьев Александр Владимирович[1]⋆ Alexandr.Astafiev@mail.ru
Демидов Антон Александрович[1] AADemidov@mail.ru
Макаров Михаил Вячеславович[1] Nauka-Murom@ya.ru
Привезенцев Денис Геннадьевич[1] DGPrivezencev@mail.ru

[1]Муром, Муромский институт (филиал) федерального государственного бюджетного образовательного учреждения высшего образования «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых»

В статье приведен анализ рынка, который показал, что подобного рода системы необходимы как крупным компаниям, так и компаниям малого бизнеса, что обуславливает актуальность рассматриваемой проблемы. Целью работы является разработка алгоритма позиционирования средств малой механизации на основе сенсорных сетей из BLE-маяков для построения RTLS-систем и систем автономной навигации.

В настоящее время развитие мобильных технологий и интернета вещей (Internet of Things) дало развитие технологиям, позволяющим более эффективно организовывать позиционирование внутри помещений. Взамен более ресурсоемким технологиям, таким как: GPS, ГЛОНАСС, Wi-Fi, UHF, Bluetooth пришли технологии: ZigBee, Z-Wave, Thread и Bluetooth Low Energy (BLE).

В статье предлагается использовать радиомаяки, основанные на технологии BLE для построения беспроводных сенсорных сетей, что позволит разработать алгоритм позиционирования в режиме реального времени. Сформулированы требования, предъявляемые к сенсорным сетям на основе BLE-маяков. Приведено описание алгоритма внутреннего позиционирования и проведены экспериментальные исследования, показавшие увеличение точности определения координат считывающего устройства по сравнению с аналогами.

[1] *Astafiev A. V., Privezentsev D. G.* Development of automated identification technology objects during their movement along not typed routes using multi-code labeling // 2018 International Russian Automation Conference, RusAutoCon 2018, Sochi; Russian Federation, 2018. .

Development of an algorithm for positioning a mobile device based on sensor networks from BLE beacons for building autonomous navigation systems

Alexandr Astafiev[1]⋆ Alexandr.Astafiev@mail.ru
Anton Demidov[1] AADemidov@mail.ru
Mikhail Makarov[1] Nauka-Murom@ya.ru
Denis Privezencev[1] DGPrivezencev@mail.ru

[1]Murom, Murom Institute (branch) Federal state budgetary Educational Institution of Higher Education ≪Vladimir State University named after Alexader Grigoryevich and Nickolay Grigoryevich Stoletovs≫

The article provides an analysis of the market, which showed that such systems are necessary both for large companies and small businesses, which determines the urgency of the problem under consideration. The goal of the work is to develop an algorithm for positioning small-scale mechanization tools based on sensor networks from BLE beacons for building RTLS systems and autonomous navigation systems.

Currently, the development of mobile technologies and the Internet of Things (Internet of Things) has given rise to technologies that allow more efficient organization of indoor positioning. Instead of more resource-intensive technologies, such as: GPS, GLONASS, Wi-Fi, UHF, Bluetooth, technologies came: ZigBee, Z-Wave, Thread and Bluetooth Low Energy (BLE).

The article suggests using beacons based on BLE technology for building wireless sensor networks, which will allow developing a real-time positioning algorithm. The requirements for sensor networks based on BLE beacons are formulated. A description is given of the algorithm for internal positioning and experimental studies have been carried out that have shown an increase in the accuracy of determining the coordinates of the linking device compared to analogues.

[1] *Astafiev A. V., Privezentsev D. G.* Development of automated identification technology objects during their movement along not typed routes using multi-code labeling // 2018 International Russian Automation Conference, RusAutoCon 2018, Sochi; Russian Federation, 2018. .

Применение методов интеллектуального анализа данных для построения глобальной модели полного электронного содержания ионосферы

Жуков Алексей Витальевич[1,2]⋆ zhukovalex13@gmail.com
Сидоров Денис Николаевич[1,2] d.sidorov@iszf.irk.ru
Ясюкевич Юрий Владимирович[1] yasukevich@iszf.irk.ru
[1]Иркутск, Институт солнечно-земной физики СО РАН
[2]Иркутск, Институт систем энергетики им. Л.А. Мелентьева СО РАН

Состояние ионосферы существенно влияет на работу многих критически важных систем. Одной из форм представления информации о состоянии ионосферы являются карты полного электронного содержания (ПЭС) [1]. В работе представлена актуальная задача построения таких карт по данным о геомагнитной и солнечной активности. Анализируемые карты ПЭС представляют из себя набор изображений полученных с временным разрешением в час. На первом этапе решается задача сокращения размерности. Для этого используется метод главных компонент из-за его высокой интерпретируемости. Так как наблюдаемый процесс имеет характерные для каждого времени суток особенности, разложение строится для каждого времени суток отдельно. Для получения карт ПЭС решается задача восстановления регрессии для каждой компоненты разложения. В качестве признаков используются индексы геомагнитной, солнечной активности, а также признаки отражающие сезонность процесса. Для приближения целевой зависимости использовался метод Random Forest. Апробации подхода проводилась на глобальных ионосферных картах за 1998-2017 гг., получаемых на основе GPS/ГЛОНАСС-измерений [2]. Показано, что точность работы предложенных моделей существенно выше, чем традиционных глобальных ионосферных моделей. Работа выполнена при поддержке РФФИ № 18-35-20038.

[1] *Zhukov A., Sidorov D., Mylnikova A., Yasyukevich Y.* Machine learning methodology for ionosphere total electron content nowcasting // International Journal of Artificial Intelligence, 2018, Vol. 16. P. 144-157.

[2] *Hernández-Pajares M. et al.* The IGS VTEC maps: a reliable source of ionospheric information since 1998 // J Geod., 2009, Vol. 83. P. 263-275.

Application of data mining methods for global ionosphere total electron content model building

Aleksei Zhukov [1,2]⋆ zhukovalex13@gmail.com
Denis Sidorov[1,2] d.sidorov@iszf.irk.ru
Yury Yasyukevich[1] yasukevich@iszf.irk.ru
[1]Irkutsk, Institute of Solar-Terrestrial Physics SB RAS
[2]Irkutsk, Melentiev Institute of Energy Systems SB RAS

The ionosphere condition significantly affects the operation of many critical systems. One of the representations of ionosphere condition is the total electron content (TEC) [1] maps. This paper is devoted to the urgent problem of obtaining such maps using geomagnetic and solar activity data. The analyzed TEC map is a set of images obtained with a one hour time resolution. First, the dimensionality reduction problem is stated. For these purposes the principal components method is exploited because of its high interpretability. Since the observed process has specific behaviour for each time of the day, decomposition is built for each time of the day separately. To obtain TEC maps, the regression problem for each decomposition component is solved. Indices of geomagnetic, solar activity, as well as variables refer to the seasonality of the process are used as features. Random Forest were used to approximate the target dependency. The approbation of the approach was carried out on global GIM ionospheric maps for 1998-2017, obtained using GPS/GLONASS measurements [2]. It was shown that the accuracy of the proposed models is significantly higher than traditional global ionospheric models. This study was funded by RFBR according to the project 18-35-20038.

[1] *Zhukov A., Sidorov D., Mylnikova A., Yasyukevich Y.* Machine learning methodology for ionosphere total electron content nowcasting // International Journal of Artificial Intelligence, 2018, Vol. 16. P. 144-157.

[2] *Hernández-Pajares M. et al.* The IGS VTEC maps: a reliable source of ionospheric information since 1998 // J Geod., 2009, Vol. 83. P. 263-275.

Поиск плавно меняющихся пространственных закономерностей на рынке недвижимости

Мехедов Иван Сергеевич[1] mehedov@mail.ru
Петрова Марина Алексеевна[2] marina_petrova@mail.ru
Филипенков Николай Владимирович[3]⋆ n.filipenkov@mail.ru
[1]Москва, Сбербанк России
[2]Москва, НИЯУ МИФИ
[3]Москва, НИУ ВШЭ, САС Институт

В настоящее время бурное развитие геоинформатики и большие объёмы данных, имеющих привязку к местности, позволяют применять методы интеллектуального анализа данных к поиску закономерностей в пространственных данных. В настоящей работе алгоритмы, разрабатываемые авторами для временных рядов, применяются к поиску закономерностей в пространственных данных. Таким образом авторы исследуют закономерности, плавно меняющиеся в пространстве.

Для иллюстрации подхода в работе использованы данные о стоимости недвижимости. Анализируется зависимость целевой переменной (стоимости квадратного метра жилья) от района, а также близости к центру города, станциям общественного транспорта, крупным магистралям, магазинам, спортивным и развлекательным центрам, институтам здравоохранения, образовательным учреждениям, офисам, паркам и т.д. Предложенный в работе подход к анализу пространственных данных, широко применим и в других областях, например, анализу различных данных со спутников, геотаргетированному маркетингу и т.п.

[1] *Филипенков Н. В.* Об одном методе поиска плавно меняющихся закономерностей в пучках временных рядов // Ж. вычисл. матем. и матем. физики, Москва: Наука, 2009. — Т. 49, № 11. С. 2020–2040. http://www.mathnet.ru/links/1078c9123862ff250c7fe8cdbcfcd75e/zvmmf4787.pdf.

Mining the Slightly Changing Geospatial Patterns in the Real Estate Market

Nikolay Filipenkov[1]⋆ n.filipenkov@mail.ru
Ivan Mekhedov[2]⋆ mehedov@mail.ru
Marina Petrova[3] marina_petrova@mail.ru
[1]Moscow, HSE, SAS Institute
[2]Moscow, Sberbank
[3]Moscow, MEPhI

The rapid growth of geospatial data in the world enables the implementation of data mining techniques to mine the patterns in geospatial data. In this paper the authors have applied the algorithms that were previously used for mining slightly changing patterns in time series to geospatial data of the real estate market. So the paper discusses mining the patterns that slightly change in space (instead of time).

The paper uses data on the real estate market. The predicted variable (square meter price) is analyzed respective to the district, distance to the city center, stations of public transport, highways, shops, sports, entertainment, healthcare, education centers, offices, parks etc. The proposed approach for mining slightly changing patterns in geospatial data is highly applicable to any data with geo-tag, e.g. space image recognition, geo-targeted marketing etc.

[1] *Filipenkov N.* A method for finding smoothly varying rules in Multidimensional time series // Computational Mathematics and Mathematical Physics, Moscow: Nauka, 2009. — 49:11 p. 1930–1948. https://link.springer.com/article/10.1134\%2FS0965542509110104.

Автоматизированный метод анализа данных космических лучей и выделения спорадических эффектов

Геппенер Владимир Владимирович[1] geppener@mail.ru
Мандрикова Богдана Сергеевна[2]⋆ 555bs5@mail.ru

[1]Санкт-Петербург, Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)

[2]Паратунка, Институт космофизических исследований и распространения радиоволн ДВО РАН

В работе предложен метод анализа данных нейтронных мониторов и обнаружения спорадических эффектов в космических лучах. Изучение динамики космических лучей представляет интерес в задачах солнечно-земной физики и прикладных исследованиях, связанных с космической погодой. Изменение условий на Солнце, в солнечном ветре, магнитосфере и ионосфере Земли существенно влияет на работу и надежность бортовых и наземных технологических систем, и угрожает здоровью и жизни людей. В настоящее время задача оперативного и точного прогноза космической погоды не решена. Для оперативного прогнозирования космической погоды весьма важно создание автоматизированных методов анализа регистрируемых данных космических лучей и своевременного обнаружения спорадических эффектов.

Предлагаемый в работе метод основан на применении кратномасштабных вейвлет-разложений (КМА) и нейронных сетей векторного квантования LVQ. Схема работы, разработанной по реализации метода программной системы, представлена на рис.1. Для реализации метода определены и обоснованы семейства ортогональных вейвлетов Добеши и Койфлеты. Разработан алгоритм определения «наилучшего» аппроксимирующего базиса в классе ортогональных вейвлетов.

На основе метода по данным мировой сети наземных станций изучена динамика космических лучей в периоды повышенной солнечной активности и магнитных бурь. Экспериментально подтверждена эффективность метода для режима оперативного анализа данных и выделения спорадических эффектов, в т. ч. малой амплитуды.

[1] *Геппенер В. В., Мандрикова Б. С.* Автоматизированный метод анализа данных космических лучей и выделения спорадических эффектов // Машинное обучение и анализ данных, 2019. (В процессе)

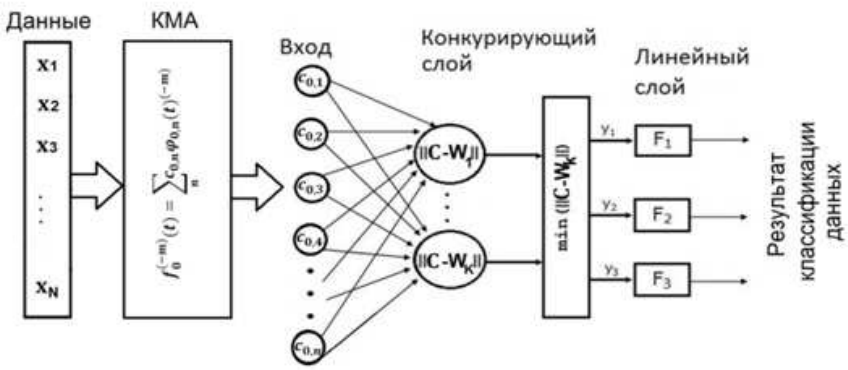


Рис. 1. Схема работы программной системы

Automated method for analyzing cosmic ray data and highlighting sporadic effects

Vladimir Geppener[1] geppener@mail.ru
Bogdana Mandrikova[2]⋆ 555bs5@mail.ru
[1]Saint-Petersburg , Saint Petersburg Electrotechnical University "LETI"
[2]Paratunka, Institute of Cosmophysical Research and Radio Wave Propagation FEB RAS

A method for analyzing neutron monitors data and detecting sporadic effects in cosmic rays is proposed. The study of the dynamics of cosmic rays is of interest in the problems of solar-terrestrial physics and applied research related to space weather. Changes in conditions on the Sun, in the solar wind, magnetosphere and Earth's ionosphere significantly affect the operation and reliability of airborne and ground-based technological systems, and threaten the health and life of people. At present, the task of prompt and accurate forecast of space weather has not been solved. It is very important to create automated methods for analyzing recorded cosmic ray data and timely detecting sporadic effects for operational forecasting of space weather.

The method proposed in the work is based on the use of multiscale wavelet decompositions (MSA) and neural networks of vector quantization LVQ. The working scheme of the software system developed for implementing the method is presented in Fig. 1. The families of orthogonal Daubechies and Coiflet wavelets are defined and grounded for the method's implementation. The algorithm for determining the "best" approximating basis in the class of orthogonal wavelets is developed.

Based on the above method, the dynamics of cosmic rays during periods of increased solar activity and magnetic storms using data from the global network of ground stations was studied. The effectiveness of the method for the operational data analysis mode and the identification of sporadic effects, including low amplitude, is experimentally confirmed.

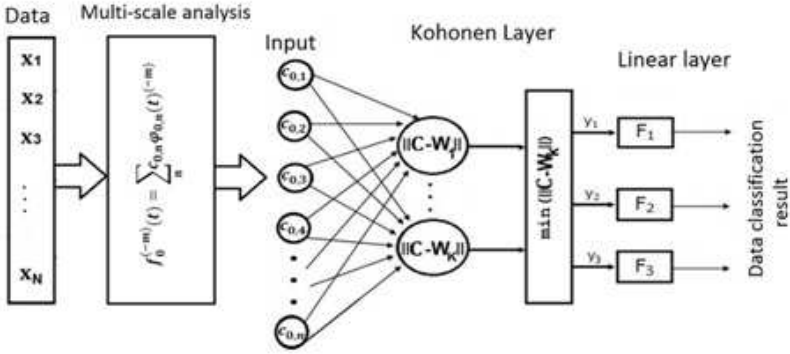


Fig. 1. Method scheme

[1] *Geppener V. V., Mandrikova B. S.* Automated method for analyzing cosmic ray data and highlighting sporadic effects // Machine Learning and Data Analysis, 2019. (In process)

Комплекс прикладных решений по построению и обучению искусственных нейронных сетей для семантической сегментации аэрокосмических изображений произвольной канально-спектральной структуры в условиях дефицита обучающих данных

Гвоздев Олег Генадиевич[1,2] gvozdev@miigaik.ru
Мурынин Александр Борисович[1,3] amurynin@bk.ru
Рихтер Андрей Александрович[1]⋆ urfin17@yandex.ru

[1]Москва, Научно-исследовательский институт аэрокосмического мониторинга "АЭРОКОСМОС"
[2]Москва, Московский государственный университет геодезии и картографии
[3]Москва, Федеральный исследовательский центр "Информатика и управление" РАН

Специфика распознавания аэрокосмических изображений обусловлена такими их особенностями как: произвольная канально-спектральная структура изображений; большой диапазон пространственных разрешений; большой объем данных каждого снимка; направление съёмки "в надир"; потребность в специальных знаниях и навыках для осуществления разметки; ограниченная доступность однородных серий снимков, пригодных для формирования обучающей выборки.

Эти особенности ограничивают или полностью исключают применение предобученных моделей (knowledge transfer), использования типовых топологий, таких как U-Net или ResNet-(18,36,...), ввиду их быстрого переобучения, загрузки изображений целиком в ОЗУ GPGPG и использование "мозаичного" подхода, ввиду наличия больших пустот (областей без данных, либо незначительных).

Показательным примером является задача распознавания открытого грунта (ОГ) и открытого мусора (ОМ) на спутниковых изображениях. Эти поверхности легко различаются человеком по характерным визуальным признакам, но, с другой, не поддаются дешифрированию классическими методами цифровой обработки.

Сложность дешефрирования ОМ обсуловлена: разнообразием геометрических форм, текстур, зернистостью, отсутствием однозначного четкого контура, смешением с объектами других типов (грунтовыми насыпятми, дорогами, инфраструктурой мусорных полигонов), случайным вкраплением частиц ОМ в фон.

Последующая оценка точности работы ИНС отягощается теми-же причинами.

Авторами разработан и реализован полностью параметризируемых технологический процесс обучения ИНС семантической сегментации, активно использующий особенности аэрокосмической природы обрабатываемых изображений.

Аугментация исходных данных представленных в растровой форме выполняется независимо, для каждого исходного изображения:

1. Определение ценности каждого пиксела изображения, с учётом: веса класса; информационной насыщенности его окрестности; дополнительнх отметок эксперта.

2. Формирование множества областей-кандидатов — 4-х-угольнх полигонов, полученные из квадратных путём последовательности случайных (в заданных пределах) аффинных преобразований.

3. Решение оптимизационной задачи выбора подмножества областей-кандидатов, наилучшим образом покрывающих изображение, пропорционально ценности пикселов внутри области-кандидата и их удалению от центра области-кандидата.

4. Извлечение и трансформация данных каждой области в квадратный растр (обеспечивает внесение геометрических искажений), внесение специфических для данного типа изображений искажений, например искажение балланса белого, гаммы, добавление шума.

В качестве каркаса топологии ИНС семантической сегментации используется U-Net. Базовые блоки кодировщика и декодировщика в котором заменяются на ResNet-подобный (выше скорость обучения) или Inception-подобный (выше точность предсказания), активационная функция заменяется на ELU (для преодоления эффектная "умирания" нейронов).

В отдельных случаях (например, при существенных колебаниях характера освещённости в выборке) к блокам кодировщика применяется техника Squeeze-Excitation.

Для особенно малых обучающих наборов, для регуляризации применяется техника Dropout.

При обучении используются плавное уменьшение learning rate.

Опционально применение техник active learning и self-supervised learning. Последняя реализуется путем добавления в обучающую выборку разметки, полученной путём усреднения результатов многократной аугментации и распознавания каждой неразмеченной области.

Совокупность рассмотренных техник позволяет достичь F1-меры > 0.7 для наборов в 20–40 исходных изображений и > 0.8 для наборов в 50–150 исходных изображений, при количестве свободных параметров ИНС в пределах $15e7$.

Перспективными направлениями развития представленного технологического процесса являются: предобучение ИНС с помощью метода deep cluster, обощенного для отдельных пикселей изображения; решение задачи сегментации экземпляров объектов с помощью методов семейства deep waterched; исследование альтернативных базовых топологий (U-Net++, DeepLab и др.); применение более сложных стратегий обучения, в частности, возврата к ранее обученным моделям, в случае выявления деградации; генерация обучающих образцов в реальном времени.

Исследования проведены при финансовой поддержке Минобрнауки России (уникальный идентификатор проекта RFMEFI58317X0061).

[1] *Гвоздев О. Г.*, *Мурынин А. Б.*, *Рихтер А. А.* Комплекс прикладных решений по построению и обучению искусственных нейронных сетей для семантической сегментации аэрокосмических изображений произвольной канально-спектральной структуры в условиях дефицита обучающих данных // Машинное обучение и анализ данных, 2019

Set of applied solution on design and training of artificial neural networks for semantic segmentation of aerospace imagery having arbitrary spectral/channel structure in case of training data deficiency

Oleg Gvozdev[1,2] gvozdev@miigaik.ru
Alexander Murynin[1,3] amurynin@bk.ru
Andrey Richter[1]⋆ urfin17@yandex.ru

[1]Moscow, State scientific Institution "Institute for Scientific Research of Aerospace Monitoring "AEROCOSMOS"
[2]Moscow, Moscow State University of Geodesy and Cartography
[3]Moscow, Federal Research Center "Informatics and Management" RAS

The specificity of aerospace imagery interpretation due to such it's peculiarities as: arbitrary spectral/channel structure; wide range of spatial resolution; large amount of data in each image, off-nadir pointing, labeling requires special qualifications; limited availability of uniform image sets, useful for building training sets.

This features significantly limits or totally rules out ability to use widely available pretrained models and knowledge transfer technique, usage of the common topoligies such as U-Net or ResNet-(18,36,...) "as-is" because of fast overfitting, loading fullsize images to GPGPU RAM or scattering with a checkered-like pattern due to large void areas (without data at all, or without significant data) in images.

Recognition of the dry soil or waste disposal areas (WDA) on satellite imagery is one of vivid examples of this complexity. This surfaces can be easily distinguished by human due to visual features, but not by using classical digital processing methods.

The WDA recognition complexity lies in variety of spatial configurations, texture patterns, texture grain, fuzzy outlines, mixings with all types of the background objects (soil mounds, roads, infrastructure objects on the territory of organized WDA) and random interspersing in it.

This also confuses evaluations of the trained ANN accuracy.

Authors designed and implemented fully-parametrized training pipeline for semantic segmentation ANN, which key feature is considering and utilization peculiarities of aerospace imagery.

Augmentation of source data performs independently for every image:

1. Evaluation of every pixel's significance due to its class weight, local image entropy and additional expert's marks.

2. Random sampling of large amount of nearly squared polygons on image area.

3. Choosing the optimal subset of sampled polygons under condition of best coverage of image pixels proportionally to its significance, and distance from polygons' center.

4. Extraction each polygon to square raster image and applying augmentation procedures, specific for image type (for RGB: white balance and gamma correction, adding noise).

As a framework for semantic segmentation ANN the U-Net topology is used. Base encoder and decoder blocks are replaced with ResNet-like (faster in training) or Inception-ResNet-like (more accurate) blocks. The ReLU nonlinearities are replaces with ELU, with the aim of reduction "dying" neurons effect.

The Squeeze-Excitation method is used on encoder blocks, on special occasions, such as wide variety of lighting conditions in imagery.

For extremely small datasets dropout regularization is used.

Learning rate decay scheduling is used.

The active learning and self-supervision methods are also used on special occasions. The self-supervision learning is implemented via extending training set with labeling, obtained via averaging of prediction for several augmented instanced of each unlabeled area.

The discussed method and techniques used jointly allows to reach F_1-score > 0.7 for datasets with 20–40 training images, and F_1-score > 0.8 for datasets with 50–150 training images, using topolgies with at most 10^7 trainable parameters.

The most promising directions of improvement for this pipeline are: pretraining ANN with deep cluster method, generalized for distinct pixels of image; applying deep watershed family methods for instance segmentation task; developing advanced training strategies, involving backtracking in case of performance regression; realtime training samples geneartion.

This research is funded by the Ministry of Science and Higher Education of the Russian Federation (unique project identifier RFMEFI58317X0061).

[1] *Gvozdev O.*, *Murynin A.*, *Richter A.* Set of applied solution on design and training of artificial neural networks for semantic segmentation of aerospace imagery having arbitrary spectral/channel structure in case of training data deficiency // Machine Learning and Data Analysis, 2019.

Полиномиальный алгоритм для нахождения нижней оценки общего времени выполнения проекта

Архипов Дмитрий Игоревич[1]⋆ miptrafter@gmail.com
Баттайя Ольга Николаевна[2] olga.battaia@kedgebs.com
Лазарев Александр Алексеевич[1,3,4] jobmath@mail.ru

[1]Москва, Институт проблем управления им. В. А. Трапезникова Российской академии наук
[2]Бордо, KEDGE Business School
[3]Москва, Московский государственный университет им. М.В. Ломоносова
[4]Москва, Национальный исследовательский университет Высшая школа экономики

Существует множество алгоритмов для нахождения нижней оценки общего времени проекта с ресурсными ограничениям (RCPSP). Однако быстрые алгоритмы обычно не дают наилучших оценок, а методы, которые позволяют найти более точные оценки, имеют высокую трудоёмкость. Предложен новый полиномиальный алгоритм нахождения нижней оценки общего времени выполнения проекта (RCPSP) для обобщённой постановки, в которой доступные количества ресурсов выражены кусочно-постоянными функциями а между работами могут быть заданы отношения предшествования с временными лагами. Основная идея алгоритма основана на последовательной оценке пар ресурсов и их совокупной рабочей нагрузки. Численные эксперименты с использованием примеров тестовой библиотеки PSPLIB и реальных промышленных данных подтвердили эффективность алгоритма и возможность его использования для задач большой размерности. Трудоемкость алгоритма составляет $O(r^2n^2(n+m))$ операций, где n — количество задач, r — количество ресурсов, m — число сегментов функции емкости ресурса.

Работа поддержана грантом РНФ № 17-19-01665.

[1] *Arkhipov D., Battaïa O., Lazarev A.* An efficient pseudo-polynomial algorithm for finding a lower bound on the makespan for the Resource Constrained Project Scheduling Problem // European Journal of Operational Research V.275 I.1, Amsterdam: Elsevier, 2019. — p. 35–44.

Polynomial algorithm for finding a lower bound on the project makespan

Dmitry Arkhipov[1]⋆ miptrafter@gmail.com
Olga Battaïa[2] olga.battaia@kedgebs.com
Alexander Lazarev [1,3,4] jobmath@mail.ru

[1]Moscow, V.A. Trapeznikov Institute of Control Sciences, Russian Academy of Sciences
[2]Bordeaux, KEDGE Business School
[3]Moscow, Lomonosov State University, Moscow, Russia
[4]Moscow, National Research University Higher School of Economics

A plenty algorithms for finding a lower bound on the makespan for the Resource Constrained Project Scheduling Problem (RCPSP) were proposed in the literature. However, fast computable lower bounds usually do not provide the best estimations and the methods that obtain better bounds are mainly based on the cooperation between linear and constraint programming and therefore are time-consuming. A new polynomial algorithm is proposed to find a makespan lower bound for RCPSP with time-dependent resource capacities and precedence relations with time lags. The main idea of the lower bound calculation is based on a consecutive evaluation of pairs of resources and their cumulated workload. Numerical experiments show that this algorithm provides good results for PSPLIB benchmark instances and it can be used for calculating lower bounds for large-scaled problem instances in reasonable time. The complexity of lower bound computation is $O(r^2n^2(n+m))$ operations, where n — number of tasks, r — number of resources, m — number of resource capacity function breakpoints.

This research is supported by the Russian Science Foundation (grant 17-19-01665).

[1] *Arkhipov D., Battaïa O., Lazarev A.* An efficient pseudo-polynomial algorithm for finding a lower bound on the makespan for the Resource Constrained Project Scheduling Problem // European Journal of Operational Research V.275 I.1, Amsterdam: Elsevier, 2019. — p. 35–44.

Распознавание, анализ и визуализация интернет-мемов

Германчук Мария Сергеевна[1]⋆ m.german4uk@yandex.ru
Козлова Маргарита Геннадьевна[1] art-inf@mail.ru
[1]Симферополь, Крымский федеральный университет им. В. И. Вернадского

Интеллектуализация обработки данных потока интернет-мемов включает задачи поиска и идентификации мемов в сети, распознавания текста и изображения мема, анализ сопутствующей информации, выяснение графовой структуры сети распространения потока мемов, задачи кластеризации и визуализации.

Разработка соответствующего инструментария и технологий направлена на создание системы анализа, синтеза и управления потоком интернет-мемов в социальных сетях.

Прежде всего в этой работе будут рассматриваться вирусные изображения, которые состоят из некоторого образа и текста. Даже в таком узком представлении мемов задача остается сложной. Мемы, как изображение и текст, представимы в социальных сетях в большом разнообразии. В моделировании процессов используется как непосредственно содержащаяся в меме информация, так и сопутствующая, сопровождающая, комментирующая. Интеллектуализация обработки такой специфичной информации сопровождается экспертными оценками [1].

В качестве среды распространения была выбрана социальная сеть «Вконтакте». Ожидаемый ввод мема в систему обработки – это ссылка на мем социальной сети «Вконтакте» или непосредственно найденное изображение с текстом. Одной из целей проекта является система, которая может правильно маркировать интернет-мемы и предсказывать их распространение. Из-за открытости вирусных мем-образов не существует списка, в котором описан (или будет описан) каждый мем, когда-либо созданный, нет единого управляющего интернет-органа, который индексирует новые мемы в соответствии с какой-либо конвенцией. Это означает, что модель не может быть обучена на каждом классе, который существовал ранее или может появиться в будущем. Требуются адаптивные модели и алгоритмы для анализа и прогнозирования.

Для разработки специального инструментария должны быть, в частности, решены следующие задачи:

1. Извлечение текста из мема-изображения.
2. Классификация текста, изображения и мема в целом.
3. Использование различных метрик сети для корректировки классификации интернет-мемов.
4. Применение различных метрик для разработки прогнозирующей системы распространения интернт-мемов.

В данной работе для решения задачи распознавания текста был выбран алгоритм EAST (An Efficient and Accurate Scene Text Detector). Алгоритм использует

полносвязную свёрточную нейронную сеть, которая принимает решения основываясь на уровне слов и строк, исключая промежуточные шаги. По сравнению с другими алгоритмами данный алгоритм выделяется высокой точностью и малым временем работы.

Ключевым компонентом предлагаемого алгоритма является модель нейронной сети, которая обучена непосредственному прогнозированию существования текстовых экземпляров и их геометрии извлекаемых из полных изображений. Модель представляет собой полносвязную свёрточную нейронную сеть, адаптированную для обнаружения текста, которая выводит предсказания слов или текстовых строк.

Распознавание текста происходит в несколько этапов. Первый этап – обработка изображения.

Пусть имеется изображение, большую часть которого составляет текст. Задача состоит в том, чтобы выделить только текст и избавиться от шумов. Проблемой является тот факт, что цвет текста неизвестен и текст может содержать несколько оттенков одного цвета. Поэтому на данном этапе изображение кластеризируется. Центроид самого большого кластера определяется цвет текста в полученном изображении. Заключительным этапом обработки изображения является построение маски, которая полностью отделяет текст от остального изображения.

Второй этап включает в себя выделение контуров, которые будут являться буквами; распознавание строк текста и объединение букв в слова. Для выделенных контуров находятся центры масс, после чего строятся основные строки, на которых лежит текст. Это позволяет не только определять правильную последовательность обработки символов, но и избавляет от шумов, которые могли остаться после обработки изображения. Далее, основываясь на интервалах между буквами, контуры объединяются в слова.

Третий этап включает в себя непосредственное распознавание символов. Для этого использовалась свёрточная нейронная сеть, обученная на 47653 изображениях русских букв различных шрифтов. Точность распознавания нейросети достигает 98.22% на представительной выборке и 98.8% на тестовой.

Разработанная OCR хорошо выделяет текст и классифицирует небольшие слова, однако испытывает проблемы при классификации больших объёмов текста. По этой причине финальная версия алгоритма по извлечению текста из изображения включает в себя два этапа. На первом этапе алгоритм EAST выделяет текстовые блоки. Если таковых много и они образуют большую группу, то для их распознавания используется Tesseract.

На первом этапе в рамках поставленной задачи собрана выборка, состоящая из 63 политических мемов и 44 неполитических. В качестве дополнительных объектов использовался текст комментариев из социальной сети «ВКонтакте», содержащий как политический, так и не политический контекст. В общей слож-

ности для обучения использовалось 168 предложений. Для валидации бралась выборка, состоящая из 11 политических мемов и 7 не политических мемов.

На основе представленной технологии разрабатывается система социальных опросов и прогнозов, которая должна работать в реальном времени на основе постоянно пополняемой, распознанной и классифицированной базе интернет-мемов.

[1] Германчук М. С., Козлова М. Г., Лукьяненко В. А. Проблематика моделирования процессов распространения интерет-мемов // Анализ, моделирование, развитие социально-экономических систем: сборник научных трудов XII Международной школы-симпозиума АМУР-2018 гг. Симферополь-Судак, 14-27 сентября 2018 / Под ред. А. В. Сигала. – Симферополь: ИП Корниенко А. А., 2018. – С. 136-139.

Recognition, analysis and visualization of Internet memes

Mariya Germanchuk[1]⋆ m.german4uk@yandex.ru
Margarita Kozlova[1] art-inf@mail.ru
[1]Simferopol, V. I. Vernadsky Crimean Federal University

Intellectualization of Internet meme stream data processing includes tasks of search and identification of memes in the network, recognition of meme text and image, analysis of related information, figuring out the graph structure of the meme stream distribution network, clustering and visualization tasks.

The development of appropriate tools and technologies is aimed at creating a system of analysis, synthesis and management of the flow of Internet memes in social networks.

First of all, this work will consider viral images, which consist of some image and text. Even in such a narrow view of memes, the task remains difficult. Memes, like image and text, are represented in social networks in a wide variety. In process modeling, both the information directly contained in the meme and the accompanying, accompanying, commenting information are used. The intellectualization of the processing of such specific information is accompanied by expert assessments [1].

The social network "Vkontakte" was chosen as the distribution medium. The expected input of the meme into the processing system is a link to the meme of the social network "Vkontakte" or directly found image with text. One of the goals of the project is a system that can correctly label Internet memes and predict their spread. Because of the openness of viral memes, there is no list that describes (or will describe) every meme ever created, there is no single Internet governing body that indexes new memes under any Convention. This means that the model cannot be trained on every class that existed before or may appear in the future. Adaptive models and algorithms are required for analysis and forecasting.

For the development of special tools, the following tasks should be solved, in particular:

1. Extract text from a meme image.
2. Classification of text, image and meme in General.
3. Using different network metrics to adjust the classification of Internet memes.
4. Application of various metrics for the development of a predictive system for the distribution of Internet memes.

In this paper, the algorithm EAST (An Efficient and Accurate Scene Text Detector) was chosen to solve the problem of text recognition. The algorithm uses a fully connected convolutional neural network that makes decisions based on the level of words and lines, excluding intermediate steps. Compared to other algorithms, this algorithm stands out for its high accuracy and short operation time.

A key component of the proposed algorithm is a neural network model that is trained to directly predict the existence of text instances and their geometry

extracted from complete images. The model is a fully connected convolutional neural network adapted for text detection that outputs predictions of words or text strings.

Text recognition occurs in several stages. The first stage is image processing.

Let there be an image, most of which is text. The task is to highlight only the text and get rid of the noise. The problem is the fact that the color of the text is unknown and the text may contain several shades of the same color. Therefore, at this stage, the image is clustered. The centroid of the largest cluster is determined by the color of the text in the resulting image. The final stage of image processing is to build a mask that completely separates the text from the rest of the image.

The second stage involves highlighting the outlines that will be letters; recognizing lines of text and combining letters into words. For the selected contours, the centers of mass are located, after which the main lines on which the text lies are built. This not only allows you to determine the correct sequence of character processing, but also eliminates the noise that could remain after image processing. Next, based on letter spacing, the outlines are combined into the words.

The third stage involves direct character recognition. For this purpose, a convolutional neural network was used, trained on 47653 images of Russian letters of different fonts. The accuracy of neural network recognition reaches 98.22% on a representative sample and 98.8% on a test one.

The developed OCR is good at highlighting text and classifying small words, but has trouble classifying large amounts of text. For this reason, the final version of the algorithm for extracting text from an image includes two stages. At the first stage, the algorithm AUTOMATICALLY allocates text blocks. If there are many of them and they form a large group, then tesseract is used to recognize them.

In the first stage, a sample consisting of 63 political memes and 44 non-political memes was collected as part of the task. As additional objects, the text of comments from the social network "Vkontakte" was used, containing both political and non-political context. A total of 168 sentences were used for training. For validation, a sample consisting of 11 political memes and 7 non-political memes was taken.

On the basis of the presented technology, a system of social surveys and forecasts is developed, which should work in real time on the basis of a constantly updated, recognized and classified database of Internet memes.

[1] Germanchuk M. S., Kozlova M. G., Lukyanenko V. A. Problematics of modeling of processes of distribution of Internet memes // Analysis, modeling, development of social and economic systems: collection of scientific works of XII international school-Symposium AMUR-2018 Simferopol-Sudak, September 14-27, 2018 / ed. – Simferopol: IP Kornienko A. A., 2018. – Pp. 136-139.

Задача распознавания символического образа динамической системы

Германчук Мария Сергеевна[1] m.german4uk@yandex.ru
Лукъяненко Владимир Андреевич[1]⋆ art-inf@yandex.ru
Меньшиков Альберт Олегович[1] bosmervor1@gmail.com
[1]Симферополь, Крымский федеральный университет им. В. И. Вернадского

Для компьютерного моделирования и исследования динамической системы удобно использовать символический образ [1], который представляет собой ориентированный граф и является аппроксимацией дискретной систем. Он строится по заданному покрытию фазового пространства ячейками. Многие задачи исследования динамических систем могут быть сведены к задачам исследования построенного ориентированного графа – символического образа.

Определение. Символический образ H – это ориентированный граф G, построенный по покрытию C фазового пространства исследуемой динамической системы и отражающий переходы из одного элемента разбиения другой.

Пусть $C = (C_i, \ldots, C_k)$ – конечное покрытие области M замкнутыми множествами, которые назовем ячейками. Имеется область M, динамика F и построенное покрытие C, тогда можно построить такой граф G, что каждая его вершина i соответствует ячейке C_i, а ребра соответствуют узлам между ячейками, то есть направленное ребро $i \to j$ существует тогда и только тогда, когда $F(C_i) \cap C_j \neq \varnothing$. Этот ориентированный граф и будет называться символическим образом.

Алгоритм построения символического образа

1. Для исходного множества M_0, в котором лежат значения дискретной динамической системы, строим ячеечное покрытие C. В данном конкретном случае, ячейки имеют форму квадрата с заранее заданным размером ребра d_0 (ячейки могут иметь, как произвольную форму, так и произвольный размер).

2. Для покрытия C строим граф G, такой, что он является символическим образом.

3. Используя один из алгоритмов поиска сильно связных вершин, например, алгоритмы Тарьяна [3], [4] или алгоритм поиска сильно связных вершин на основе поиска путей [5], выделяем сильно связные вершины $\{i_k\}$. Если $\{i_k\} = \varnothing$, то в области M_0 нет аттрактора и локализуемое цепно-рекуррентное множество является пустым и процесс его локализации прекращается [2]. Иначе, удаляем из графа невозвратные вершины и от области M_0 переходим к новой области M_1, такой что $M_1 = \{x \in M_0^{i_k} : i_k \in G\}$.

4. Строим для M_1 покрытие C_1, так что размер ребра новой ячейки $d_1 = \frac{d_0}{2}$. Переходим к пункту 2 в том случае, если $d_1 > \varepsilon$, где ε – заранее заданный предельный размер ячейки.

Задача программной реализации алгоритма построения символического образа основывается на теории предложенной Г. С. Осипенко [1], [2]. Для реализации данного проекта выбран язык программирования Python, что обосновано широким выбором библиотек, удобных для работы с графами и другими математическими структурами, его стабильностью и переносимостью. Данный язык хорошо подходит для реализации поставленной задачи. Для отображения графики и отрисовки изображений, необходимых для визуализации символического образа, выбрана графическая среда OpenGL. OpenGL (Open Graphics Library) – спецификация, определяющая платформонезависимый программный интерфейс для написания приложений, использующих двумерную и трёхмерную компьютерную графику. Используется библиотека OpenGL GLUT. OpenGL Utility Toolkit (GLUT) – библиотека утилит для приложений под OpenGL, которая в основном отвечает за системный уровень операций ввода-вывода при работе с операционной системой. Используются следующие функции: создание окна, управление окном, мониторинг за вводом с клавиатуры и событий мыши. Подключена библиотека NetworkX, созданная специально для выбранного языка программирования Python. Пакет NetworkX предназначен для создания и анализа изучения структур, динамики и функций сложных сетей, а следовательно направленных графов. Библиотека включает в себя специальные структуры данных для графов, как направленных, так и ненаправленных, множество стандартных алгоритмов на графах, генераторы классических и случайных графов, гибкую работу с вершинами и ребрами, которые могут быть практически любым объектом и иметь любые параметры и открытый код, что позволяет свободно пользоваться любыми алгоритмами из библиотеки.

Для работы с массивами данных, была подключена библиотека NumPy. Данная библиотека предоставляет широкие возможности для работы с двумерными массивами: множество удобных методов для работы с массивами, возможность создания больших массивов за одну строку и прочее.

Также была используется SciPy. Из нее была импортирована часть, относящаяся к дифференциальным уравнениям. Пакет SciPy позволяет решать, как простые системы, так и системы дифференциальных уравнений с параметрами. Средой разработки была выбрана Anaconda. Удобство работы с Anaconda состоит в том, что большинство широко используемых библиотек, работающих на Python, уже установлены в дистрибутивы, что значительно упрощает их подключение.

В качестве тестовых примеров построены отображения Хенона, Заславского, аттракторы Питера де Йона и др.

[1] Осипенко Г. С. Введение в символический анализ динамических систем / Г. С. Осипенко, Н. Б. Ампилова. – Изд. СПбГУ, 2005. – 217 с.

[2] Петренко Е. И. Разработка и реализация алгоритмов построения символического образа / Е. И. Петренко // Электронный Журнал Дифференциальные Уравнения и Процессы Управления. – 2006. – № 3. – С. 42.

[3] Tarjan R. E. Depth-first search and linear graph algorithms / Robert E. Tarjan // SIAM Journal on Computing. – 1972. – Vol. 1, no. 2. – P. 146-160.

[4] Седжвик Р. Алгоритмы на графах / Р. Седжвик. – 3-е изд. – Россия, Санкт-Петербург: «ДиаСофтЮП», 2002. – С. 496.

[5] Sedgewick R. Algorithms in Java, Part 5 – Graph Algorithms / R. Sedgewick – 3rd ed. – Cambridge MA: Addison-Wesley. – 528 p.

The problem of recognizing the symbolic image of a dynamic system

Mariya Germanchuk[1] m.german4uk@yandex.ru
Vladimir Lukyanenko[1]⋆ art-inf@yandex.ru
Albert Menshikov[1] bosmervor1@gmail.com
[1]Simferopol, V. I. Vernadsky Crimean Federal University

For computer modeling and dynamic system research, it is convenient to use the symbolic image [1], which is a directed graph and is an approximation of discrete systems. It is constructed by a given covering of the phase space by cells. Many problems of the study of dynamical systems can be reduced to the problems of the study of the constructed oriented graph-a symbolic image.

Definition. The symbolic image H is a directed graph G constructed by covering C of the phase space of the dynamical system under study and reflecting transitions from one partition element to another.

Let $C = (C_i, \ldots, C_k)$ be the finite coverage of the M region by closed sets, which we call cells. There is a M dynamics F and built a cover C, then we can build a graph G that each vertex i corresponds to the cell C_i, and edges correspond to nodes between cells, that is, a directed edge $i \to j$ exists only when $F(C_i) \cap C_j \neq \varnothing$. This oriented graph will be called symbolically.

Algorithm for the construction of a symbolic image

1. For the initial set M_0, in which the values of a discrete dynamical system lie, we construct a cell cover C. In this particular case, the cells have the shape of a square with a predetermined edge size d_0 (cells can have either an arbitrary shape or an arbitrary size).

2. To cover C, we construct a graph G, such that it is a symbolic image.

3. Using one of the algorithms for finding strongly connected vertices, for example, Tarjan's algorithms [3], [4] or the algorithm for finding strongly connected vertices based on the path search [5], we allocate strongly connected vertices $\{i_k\}$. If $\{i_k\} = \varnothing$, then there is no attractor in M_0 and the localizable chain-recurrent set is empty and the process of its localization stops [2]. Otherwise, we remove the irrevocable vertices from the graph and move from the M_0 region to the new M_1 region, such that $M_1 = \{x \in M_0^{i_k} : i_k \in G\}$.

4. We build for M_1 a coverage of C_1, so that the edge size of the new cell $d_1 = \frac{d_0}{2}$. Go to step 2 if $d_1 > \varepsilon$, where ε is a predefined limit cell size.

The problem of software implementation of the algorithm for constructing a symbolic image is based on the theory proposed by G. S. Osipenko [1], [2]. For the implementation of this project, the Python programming language was chosen, which is justified by a wide choice of libraries that are convenient for working with graphs and other mathematical structures, its stability and portability. This

language is well suited for the implementation of the task. The OpenGL graphics environment is selected to display the graphics and render the images needed to render the symbolic image. OpenGL (Open Graphics Library) – a specification that defines a platform-independent programming interface for writing applications that use two-dimensional and three-dimensional computer graphics. The OpenGL GLUT library is used. OpenGL Utility Toolkit (GLUT) is a library of utilities for OpenGL applications that is mainly responsible for the system level of I / o operations when working with the operating system. The following functions are used: window creation, window management, monitoring of keyboard input and mouse events. The NetworkX library created specifically for the selected Python programming language is connected. The NetworkX package is designed to create and analyze the study of structures, dynamics and functions of complex networks, and therefore directed graphs. The library includes special data structures for graphs, both directed and non-directed, a lot of standard algorithms on graphs, generators of classical and random graphs, flexible work with vertices and edges, which can be almost any object and have any parameters and open code, which allows you to freely use any algorithms from the library.

To work with arrays of data, the NumPy library was connected. This library provides ample opportunities for working with two-dimensional arrays: a lot of convenient methods for working with arrays, the ability to create large arrays in one line, and more.

Was also used by SciPy. The part relating to differential equations was imported from it. The SciPy package allows you to solve both simple systems and systems of differential equations with parameters. Anaconda was chosen as the development environment. The convenience of working with Anaconda is that most of the widely used Python libraries are already installed in distributions, which greatly simplifies their connection.

As test cases built display Hanona, Zaslavsky, attractors of Peter de Jon etc.

[1] Osipenko G. S. Introduction to symbolic analysis of dynamical systems / G. S. Osipenko, N. B. Ampilova. – Ed. St. Petersburg state University, 2005. – 217 p.

[2] Petrenko E. I. Development and implementation of algorithms for constructing a symbolic image / E. I. Petrenko // electronic journal differential equations and control processes. – 2006. – No. 3. – P. 42.

[3] Tarjan R. E. Depth-first search and linear graph algorithms / Robert E. Tarjan // SIAM Journal on Computing. – 1972. – Vol. 1, no. 2. – P. 146-160.

[4] Sedgwick R. Algorithms on graphs / R. Sedgwick. – 3rd ed. – Russia, St. Petersburg: "Diasoftyup", 2002. – Pp. 496.

[5] Sedgewick R. algorithms in Java, Part 5 – Graph Algorithms / R. Sedgewick – 3rd ed. – Cambridge, MA: Addison-Wesley. – 528 p.

Использование машинного обучения в задачах количественной металлографии

Ковун Владислав Анатольевич[1] sidav94@gmail.com
Каширина Ирина Леонидовна[1]⋆ kash.irina@mail.ru
Бондаренко Юлия Валентиновна[1] bond.julia@mail.ru
[1]Воронеж, ВГУ

Металлография — важное направление в металловедении, классический метод исследования и контроля качества металлических материалов. Количественная металлография изучает количественные характеристики микроструктуры сталей и сплавов. Важная часть структурного анализа образца заключается в выделении зёрен на микрофотографиях продольных и поперечных шлифов металлического образца и измерении их абсолютных и относительных размеров[1].

В последнее время количественная металлография претерпела значительные изменения благодаря появлению автоматических анализаторов изображений (AAI), но задача полностью автоматизированного определения структуры металла еще не решена. Сложность создаёт наличие на исходном изображении артефактов в виде горизонтальных линий, вызванных особенностями технического процесса создания шлифов и местами плохо различимые границы зерен.

В предлагаемой статье с использованием машинного обучения решаются такие задачи, как: выделение замкнутых контуров зерен металла на исходном изображении; автоматическое вычисление площадей зёрен и построение гистограммы их распределения.

Предлагается подход к металлографическому исследованию, основанный на сегментации с помощью обучаемого классификатора. Исходные данные (100 микрофотографий) были предоставлены лабораторией металловедения и металлофизики компании НЛМК. Для решения задач использовалась открытая библиотека компьютерного зрения OpenCV, а также нейросетевая модель U-Net, разработанная для сегментации биомедицинских изображений и хорошо проявившая себя при решении данной задачи [2].

Для обучения нейросетевой модели были вручную размечены 20 шлифов - как продольных, так и поперечных. 14 размеченных шлифов использовались для обучения, 6 использовались для тестирования. Для увеличения обучающей выборки дополнительно использовалась аугментация с помощью эластичной сети (elastic_transform), а также повороты и отражения.

Сеть U-Net полностью сверточная, в ней нет полносвязных слоев. Она проста концептуально, плохо переобучается и из-за относительного небольшого количества параметров ей нужно меньше данных для обучения. Для данной задачи использовались следующие размеры слоев: 256-128-64-32-16-32-64-128-256. Для подачи на вход сети исходное изображение разрезалось на фрагменты (батчи) размера 256x256. На выходе сети получаются изображения такого же размера.

Итоговые предсказания склеиваются. В качестве функции ошибок (loss) использовалась binary_crossentropy (так как сеть должна различить границы зерен и фон, то есть разделять пиксели изображения на два класса), используемая метрика – accuracy. После 100 итераций точность сети на обучающей выборке составила loss: 0.2479 – acc: 0.9083. Точность на тестовой выборке составила val_loss: 0.2867 - val_acc: 0.8946.

Примеры получаемых изображений представлены на рис. 1.

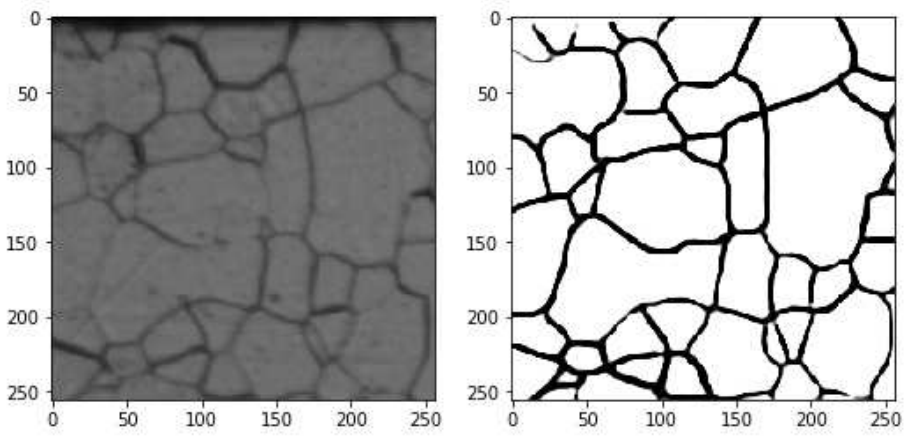


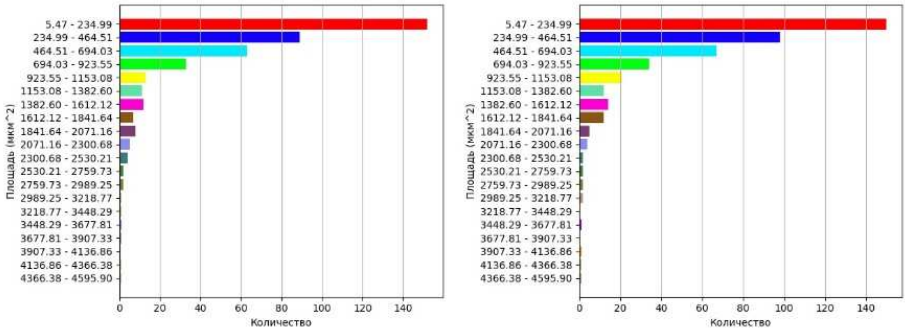
Рис. 1. Вход сети U-Net (слева) и выход (справа).

К сожалению, некоторые изображения, получаемые на выходе сети U-Net, содержали незамкнутые границы (слабо различимые и на входном изображении) что приводило к ошибочным значениям при подсчете площадей зерен. Для замыкания границ всех зерен дополнительно был разработан алгоритм постобработки с использованием OpenCV, включающий следующие шаги:

— контрастирование пикселей, эрозия и дилатация изображения;
— использование алгоритма водораздела (watershed) для поиска сегментов на изображении;
— отыскание контуров и площадей найденных алгоритмом сегментов.

Результатом работы алгоритма является гистограмма распределения площадей зерен шлифа (рис. 2).

Для сравнения эталонных и полученных гистограмм использовалась функция compareHist() из библиотеки OpenCV. В качестве метрик качества использовались величина корреляции, и статистика Хи-квадрат. Среднее значение корреляции на тестовой выборке составило 0.996, статистики Хи-квадрат 12.11, что



(a) (b)

Рис. 2. Гистограммы площадей зёрен (слева эталон, справа восстановленное изображение).

говорит о согласованности получаемых распределений (при уровне значимости 0.05 критическое значение $\chi^2_{\text{кр}} = 22.4$).

[1] ГОСТ 5639-82 Стали и сплавы. Методы выявления и определения величины зерна (с Изменением N 1) [Текст]. – Взамен ГОСТ 5639-65; Введ. с 01.01.1983. – Москва: Изд-во стандартов, 1988. — 16 с.

[2] *Ronneberger O., Fischer P., Brox T.* U-Net: Convolutional Networks for Biomedical Image Segmentation // Medical Image Computing and Computer-Assisted Intervention, New York: Springer, Cham, 2015. — vol. 9351, C. 234–241.

Machine learning usage in the tasks of quantitative metallography

Vladislav Kovun[1] sidav94@gmail.com
Irina Kashirina[1]⋆ kash.irina@mail.ru
Yuliya Bondarenko[1] bond.julia@mail.ru
[1]Voronezh, Voronezh State University

Metallography is an important area in metal science, a classic method of research and quality control of metallic materials. Quantitative metallography studies the quantitative characteristics of the microstructure of steels and alloys. An important part of the structural analysis of the sample is the selection of grains in micrographs of the longitudinal and transverse sections of the metal sample and the measurement of their absolute and relative sizes[1].

Recently, quantitative metallography has undergone significant changes due to the advent of automatic image analyzers (AAI), but the task of fully automated determination of the metal structure is yet to be solved. Complexity is created by the presence of artifacts in the form of horizontal lines on the original image caused by the features of the technical process of creating slices and often poorly distinguishable grain boundaries.

In the proposed article problems of isolation of closed loops of metal grains in the original image, automatic calculation of grain areas and the construction of a histogram of their distribution are solved via machine learning usage.

An approach to metallographic research is proposed based on segmentation using a trained classifier. Initial data (100 micrographs) were provided by the NLMK Laboratory of Metallurgy and Metallophysics. To solve the problems the OpenCV computer vision library as well as the neural network model called U-Net were used. The U-Net model was developed for biomedical images segnemtation and has proved to be good at solving such problems. [2].

To train the neural network model, 20 metal slices, both longitudinal and transverse, were manually marked out. 14 marked sections were used for training and 6 were used for testing. To increase the training sample there was additionaly used an augmentation process using the elastic network (elastic_transform) as well as rotations and reflections of input images.

The U-Net network is completely convolutional i.e. it has no fully connected layers. It is conceptually simple and poorly retrainable, and it needs less data for training because of the relatively small number of parameters. The layer sizes used for this task are 256-128-64-32-16-32-64-128-256. To feed the network input the original image was cut into fragments (batches) of size 256x256. Images of the same size are obtained as output of the network and final predictions are glued together. As a function of errors (loss) binary_crossentropy was used (since the network must distinguish between grain boundaries and background thus dividing the image pixels into two classes) an accuracy metric was used. After 100 iterations the accuracy of

the network in the training set was loss: 0.2479 - acc: 0.9083. The accuracy on the test sample was val_loss: 0.2867 - val_acc: 0.8946.

Examples of the resulting images are shown in Fig. 1.

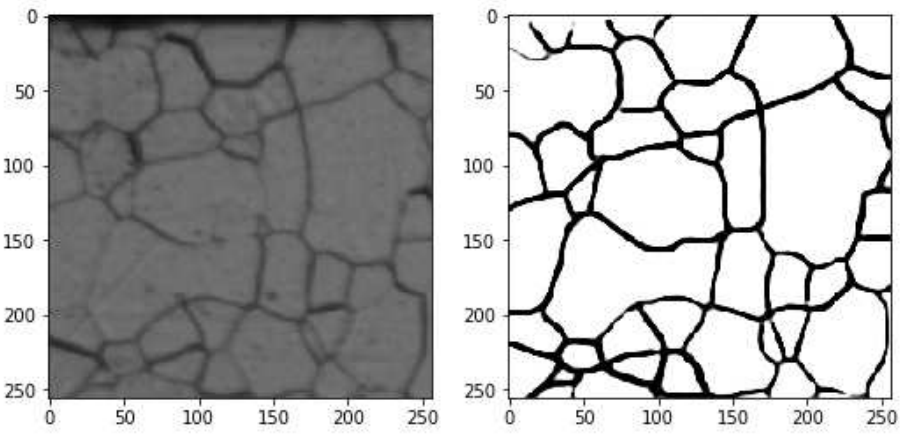


Fig. 1. U-Net input (on the left) and output (on the right) images.

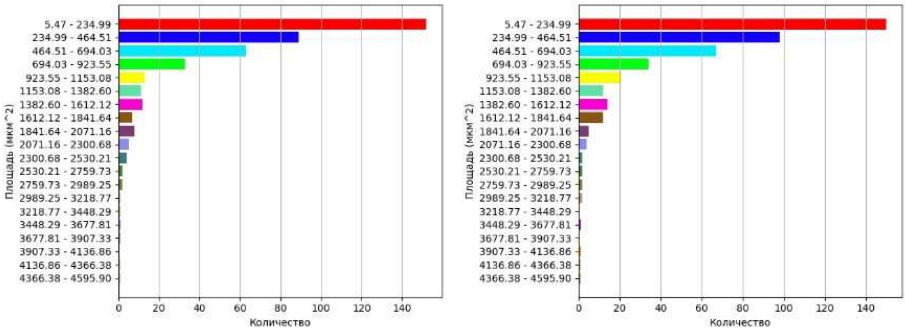
Unfortunately, some of the output images of the U-Net network contained open borders (poorly distinguishable in the input image as well), which led to erroneous values when calculating the grain areas. To close the boundaries of all grains, an additional post-processing algorithm was developed using OpenCV, which includes the following steps:

— pixel contrasting, image erosion and image dilation;
— the watershed algorithm usage for searching for segments in the image;
— approximation of the contours and areas of the segments in the algorithm output.

The output of the algorithm is the distribution histogram of the grain areas of the slice (Fig. 2).

The compareHist() function from the OpenCV library was used to compare the reference and obtained histograms. Correlation value and Chi-square statistics were used as quality metrics. The average correlation value in the test sample was 0.996, the Chi-statistic was 12.11, which indicates the consistency of the obtained distributions (at a significance level of 0.05, the critical value is $\chi^2_{\text{crit}} = 22.4$).

[1] GOST 5639-82 Steel and alloys. Methods for the identification and determination of grain size (with Change N 1) [Text]. - Replaces GOST 5639-65; Used since 01.01.1983. - Moscow: Publishing house of standards, 1988. — 16p.



(a) (b)

Fig. 2. Grain area values distribution histograms (reference is on the left, test sample is on the right).

[2] *Ronneberger O., Fischer P., Brox T.* U-Net: Convolutional Networks for Biomedical Image Segmentation // Medical Image Computing and Computer-Assisted Intervention, New York: Springer, Cham, 2015. — vol. 9351, pp. 234–241.

Метрики для задач теории расписаний с несколькими приборами

Александр Алексеевич Лазарев[1,2,3]⋆ jobmath@mail.ru
Лемтюжникова Дарья Владимировна[1,4] darabbt@gmail.com
Франк Вернер[5] frank.werner@mathematik.uni-magdeburg.de

[1]Москва, Институт проблем управления им. В. А. Трапезникова
[2]Москва, Московский государственный университет
[3]Москва, Высшая школа экономики
[4]Москва, Московский авиационный институт
[5]Магдебург, Германия, Университет Отто фон Герике

Один из подходов к решению задач теории расписаний с несколькими приборами — метрический подход — основан на введении метрик [1]. Он заключается в получении оценки абсолютной погрешности и нахождении приближённого решения для задач теории расписаний для нескольких приборов с критерием минимизации максимального временно́го смещения.

Вводится понятие метрики (расстояния) между примерами задачи. Идея предлагаемого подхода состоит в построении по исходному примеру задачи другого примера, для которого удаётся найти оптимальное или приближённое решение, с минимальным расстоянием до исходного примера по введённой метрике.

Метрический подход позволяет найти новые эффективные полиномиально разрешимые частные случаи NP–трудных задач. Полученные оценки могут быть использованы при решении задач с помощью таких алгоритмов, как метод ветвей и границ, ветвлений и отсчений, ветвлений и оценок, программирования в ограничениях и т.д.

В частности, были найдены полиномиально разрешимые частные случаи NP–трудных задач теории расписаний для нескольких приборов различного типа. Это задачи для двух и нескольких параллельных приборов, а также задачи конвейерного типа.

Работа поддержана грантом РНФ № 17-19-01665.

[1] *Лазарев А. А.* Теория расписаний. Методы и алгоритмы. // — М.: ИПУ РАН, 2019. — 427 с.

Metrics for scheduling problems with many machines

Alexander Lazarev[1,2,3]⋆ jobmath@mail.ru
Darya Lemtyuzhnikova[1,4] darabbt@gmail.com
Frank Werner[5] frank.werner@mathematik.uni-magdeburg.de
[1]Moscow, V. A. Trapeznikov Institute of Control Sciences of RAS
[2]Moscow, Lomonosov Moscow State University
[3]Moscow, Higher School of Economics National State University
[4]Moscow, Moscow Aviation Institute National State University
[5]Magdeburg, Germany, Otto von Guericke University

Thera are many approaches to solving scheduling problems. One of them is a metric approach [1]. It is based on the metrics. It consists of obtaining an estimate of the absolute error and finding an approximate solution for scheduling theory problems for many machines on the criterion of minimizing the maximum lateness.

The concept of metric (distance) between the examples of the problem is introduced. The idea of the proposed approach is to construct another example according to the original one, for which it is possible to find the optimal or approximate solution with a minimum distance to the original example in the metric.

The metric approach allows to find new effective polynomial solvable special cases of NP-hard problems. The obtained estimates can be used for solving the problems by such algorithms as the Branches and Bounds method, Constraint Programming, Branch and Cuts, Branch and Price, etc.

Polynomially solvable particular cases of NP – hard scheduling problems were found for many machines. These are problems for two and many parallel machines, as well as flow-shop problems.

This research is funded by RSCF, grant 17-19-01665.

[1] *Lazarev A. A.* Scheduling Theory. Methods and algorithms. // — Moskow: ICS RAS, 2019. — 427 p. (in Russian).

Двойственные и обратные задачи в теории расписаний

Лазарев Александр Алексеевич[1,2,3] jobmath@mail.ru
Правдивец Николай Александрович[1]⋆ pravdivets@ipu.ru
Вернер Франк[4] frank.werner@mathematik.uni-magdeburg.de
[1]Москва, Институт проблем управления им. В. А. Трапезникова
[2]Москва, Московский государственный университет
[2]Москва, Высшая школа экономики
[4]Магдебург, Германия, Университет Отто фон Герике

Рассматривается задача теории расписаний для одного прибора, на котором нужно обслужить требования множества $N = \{1, 2, \ldots, n\}$, $n = |N|$. Прибор не может обслуживать более одного требования одновременно. Прерывания обслуживания требований запрещены.

Несмотря на то, что в математическом программировании исходные и двойственные задачи обычно имеют одинаковый статус сложности, оказалось, что двойственные задачи теории расписаний, рассматриваемые в этой статье, имеют меньшую сложность, чем исходные.

Предложен алгоритм решения NP-трудной задачи теории расписаний $1 \mid r_j \mid \varphi_{\max}$ для произвольных неубывающих функциях штрафа, реализующий метод ветвей и границ. Нижняя оценка решений подпримеров выполняется с использованием решения "двойной задачи", которое может быть найдено за количество операций, не превышающее $O(n^2)$.

В дополнение к двойственной задаче, решена также и обратная задача. Алгоритм её решения имеет сложность $O(n^2)$ операций. Если в задаче минимизации максимального временного смещения мы "пытаемся выровнять" временное смещение, минимизируя максимальное, то в "обратной" задаче смещения "выравниваются" за счёт максимизации минимального. Принудительные простои запрещены.

Работа поддержана грантом РНФ № 17-19-01665.

[1] *Лазарев А. А.* Теория расписаний. Методы и алгоритмы., М.: ИПУ РАН, 2019. — 427 с.

On the Dual and Inverse scheduling problems

Alexander A. Lazarev[1,2,3] jobmath@mail.ru
Nikolay Pravivets[1]⋆ pravdivets@ipu.ru
Frank Werner[4] frank.werner@mathematik.uni-magdeburg.de
[1]Moscow, V. A. Trapeznikov Institute of Control Sciences of RAS
[2]Moscow, Lomonosov Moscow State University
[3]Moscow, Higher School of Economics National State University
[4]Magdeburg, Germany, Otto von Guericke University

We consider single machine scheduling problems, where a set of n jobs has to be processed on a single machine, where $N = \{1, 2, \ldots, n\}$ and $n = |N|$. The machine cannot process more than one job at a moment. Preemptions of the processing of a job are prohibited.

Although in mathematical programming the original and dual problems have usually the same complexity status, it turned out that the dual problems of the scheduling problems considered in this paper have a lower complexity than the original problems.

We proposed an algorithm for solving the NP-hard problem $1 \mid r_j \mid \varphi_{\max}$ for arbitrary non-decreasing penalty functions, which implements the branch and bound method. By considering the dual problem, a lower bound of sub examples solutions is carried out using a solution to the "dual" problem, which can be found in the number of operations not exceeding $O(n^2)$.

In addition to the dual problem, the inverse problem has also been solved for the lateness objective function. We proposed an algorithm with a complexity of $O(n^2)$ operations. Whereas in the problem of minimizing the maximum lateness we "try to equalize" the lateness while minimizing the maximum, in the "inverse" problem the latenesses are "equalized" due to the maximization of the minimum. Definitely, manual machine idle times are prohibited.

This research is funded by RSCF, grant 17-19-01665.

[1] *Lazarev A.A.* Scheduling Theory. Methods and Algorithms. (In print in Russian: Teoriya raspisaniy. Metody i algoritmy.) / Moscow: ICS RAS, 2019. —427 p.

Машинное обучение в задачах прогноза отказов оборудования

Некрасов Иван Васильевич[1] ivannekr@mail.ru
Правдивец Николай Александрович[1]⋆ pravdivets@ipu.ru
[1]Москва, Институт проблем управления им. В. А. Трапезникова РАН

В настоящее время во всех видах производственной деятельности производится глубокая автоматизация и интеллектуализация таких классических производственных задач, как календарное планирование, управление технологическим процессом, контроль качества, мониторинг состояния основного оборудования и т.п.

С точки зрения теории управления, современная автоматизированная система технического обслуживания и ремонта (ТОиР) представляется в виде замкнутого информационного контура, задачей которого является генерирование календарного плана вывода оборудования из работы с учетом текущей производственной ситуации на предприятии. Информация о текущей производственной ситуации формализуется, в основном, производственным планом и состоянием оборудования, и поступает в контур ТОиР в виде сигнала обратной связи.

Центральной технической задачей контура управления ТОиР является оценка и прогнозирование состояния основного оборудования (задача мониторинга). Однако, модель функционирования оборудования не всегда может быть синтезирована точными методами. Это определяет широкое применение альтернативных подходов к моделированию на основе данных, в частности методов машинного обучения. В настоящей работе рассмотрены основные методы прогнозирования технического состояния оборудования с применением машинного обучения, проанализирована их информативность с точки зрения диагностики неполадок оборудования, а также приведена общая логика автоматизированного принятия решения о корректировке плана ТОиР на основе вырабатываемых ими выходных данных.

Работа поддержана грантом РНФ № 17-19-01665.

[1] *Лазарев А.А., Некрасов И.В., Правдивец Н.А.* Evaluating Typical Algorithms of Combinatorial Optimization to Solve Continuous-Time Based Scheduling Problem // Algorithms, Bazel (Швейцария): MDPI, 2018. — С. 1–13.

On the Dual and Inverse scheduling problems

Ivan Nekrasov[1] ivannekr@mail.ru
Nikolay Pravivets[1]⋆ pravdivets@ipu.ru
[1]Moscow, V. A. Trapeznikov Institute of Control Sciences of RAS

Currently, there is an active introduction of digital technologies in all industrial spheres. Deep automation and intellectualization is being conducted for such classic industrial problems as production scheduling, technology workflow handling, quality control, assets monitoring and maintenance, etc. Optimization of maintenance schedule is one of the key points that can bring significant savings and efficiency increase to an industrial enterprise.

From the control theory standpoint the automated enterprise asset management system (referenced further as EAM) can be considered as a closed-loop information system that generates the asset's outage schedule with reference to the current production state. The current production state is mainly formalized based on production schedule and current equipment health which both are structured as a feedback information flow of the EAM-system.

The central technical problem for the closed-loop maintenance systems is how the technical state of the equipment should be calculated and estimated (the monitoring problem). Currently the general approach for equipment state estimation is associated with modeling. However, the process of asset performance is not always suitable for precise modeling techniques (for instance using the physical dependencies and equations) which gives a huge field for applying data based modeling approach including various machine learning (ML) methods. This paper gives an overview of main ML methods for the equipment technical state estimation and analyses how these methods are applied to equipment failure prediction. It also describes the workflow how the estimation algorithms' output information is utilized in the automated process of correcting the maintenance schedule.

This research is funded by RSCF, grant 17-19-01665.

[1] *Lazarev A.A., Nekrasov I.V., Pravdivets N.A.* Evaluating Typical Algorithms of Combinatorial Optimization to Solve Continuous-Time Based Scheduling Problem // Algorithms, Bazel (Switzerland): MDPI, 2018. — p. 1–13.

Мультиагентные модели и методы самоорганизации расписаний для решения сложных задач адаптивного управления ресурсами в реальном времени

Скобелев Петр Олегович[1,2]⋆ petr.skobelev@gmail.com
[1]Самара, ИПУСС РАН
[2]Москва, НАО "ГК "Генезис знаний"

Рассматриваются вызовы современной экономики и примеры сложных задач управления ресурсами в реальном времени: управление грузовыми перевозками и мобильными бригадами, машиностроительным производством, цепочками поставок и другие.

Выделяются особенности современных задач управления ресурсами в различных предметных областях, затрудняющие применение классических методов и средств, связанные с ростом сложности бизнеса, большой размерностью данных, разнообразием требований заказов и ресурсов, необходимостью применять индивидуальный подход к каждому участнику, потребностью принимать решения в реальном времени и т.д.

Формулируется задача адаптивного управления ресурсами в реальном времени.

Рассматривается мультиагентный подход к построению сложных расписаний, разработанный на основе теории сложных адаптивных систем. В этом подходе решение любой сложной задачи планирования и оптимизации ресурсов формируется как неулучшаемое «конкурентное равновесие» - консенсус, достигаемый на виртуальном рынке мультиагентной системы за счет коллективного принятия решений и взаимодействия программными агентами потребностей и возможностей, представляющих интересы, предпочтения и ограничения всех участников. Итерационный процесс построения решения любой сложной задачи и достижения такого консенсуса рассматривается как процесс самоорганизации агентов, способных самостоятельно принимать решения, но путем переговоров согласованно выявлять и разрешать конфликты, причем как с учетом частных эгоистичных интересах участников, так и объединяющей их общей системы.

Показываются и анализируются направления развития моделей и методов коллективного принятия решений в ходе самоорганизации агентов на виртуальном рынке системы: от централизованного и последовательного детерминированного принятия решений, реактивных агентов и общей целевой функции системы – к полностью распределенному и параллельному, недетерминированному принятию решений, с проактивностью агентов, их собственными функциями удовлетворенности и бонусов и штрафов агентов, возможностью самообучения на основе опыта, а также, в перспективе, направляемой самоорганизацией (guided self-organization) с гомеостатической гармонизацией значений критериев.

Приводятся результаты исследований и разработок мультиагентных систем для адаптивного управления ресурсами в реальном времени, а также примеры применений, показывающие возможность существенно (на 15-40%) повышать эффективность работы предприятий за счет "юберизации"ресурсов.

Обсуждаются преимущества предлагаемого подхода для применений в сложных задачах: простота и понятность для разработчиков, устойчивость к изменениям требований, возможность адаптивного изменения планов, распараллеливаемость, производительность, масштабируемость, живучесть и тд.

Намечаются пути дальнейшего развития подхода, связанные созданием цифровых платформ и эко-систем умных сервисов для управления ресурсами в реальном времени.

[1] *Rzevski G.*, *Skobelev P.* Managing Complexity . London-Boston: WIT Press, 2014. — С. 156.

Multi-agent models and methods of schedules self-organization for solving complex problems of adaptive resource management in real time

Skobelev Petr[1,2]⋆ petr.skobelev@gmail.com
[1]Samara, Institute of Control of Complex Problem of Russian Academy of Science
[2]Moscow, "Knowledge Genesis" Group

The challenges of modern economy and examples of complex problems of real time resource management are considered, including problem domains of cargo transportation, factories, supply chains and some others.

The number of issues are identified which restrict applications of classical combinatorial search and different heuristics methods and tools for modern resource management: growing complexity of business, high dimension of decision making space, variety of demand and resource requirements, individuality of all participants, real time constraints, etc.

The problem statement for adaptive resource management is formulated.

The multi-agent approach for solving complex problems of resource management is presented based on theory of complex adaptive systems.

In this approach the solution of any problem of scheduling and optimization is formed as a "competitive equilibrium" of demand and resource agents, representing preferences and constraints of all participants. Agents interact on virtual market of the multi-agent system until this equilibrium is reached as a consensus among agents and can't be self-improved by any one of agents. The process of consensus formation is considered as a process of agents self-organization, which all are able to take their own individual decisions, but also discover and solve conflicts with other agents by negotiations, with the view on individual egoistic interests of agents and also interests of system as a whole.

The modern view and formal definition of virtual market is presented and benefits and advantages for developers and users are discussed: easy to learn for programmers, high adaptability, high performance and scalability, resilience of plans, etc.

The key directions of development of models and methods of collective decision making are presented and analyzed starting from centralized and sequential, deterministic decision making by reactive agents with common fitness function – to fully decentralized and parallel, non-deterministic decision making by pro-active agents with individual fitness functions and bonus-penalties, advanced by guided self-organization.

The results of research and development of multi-agent systems for adaptive resource management are presented and key issues and risks of systems design and implementation are analyzed and addressed.

A few business cases are given showcasing the improve of efficiency of business up to 15-40% by "uberisation" of resources.

The next steps of future research are outlined with the focus on developing digital platform and eco-systems of smart services for resource planning and optimization.

[1] *Rzevski G.*, *Skobelev P.* Managing Complexity. London-Boston: WIT Press, 2014. — p. 156.

Расширение алгоритма FUMILI для оптимизации квадратичных функционалов со связями между параметрами

Курбатов Владимир Сергеевич[1] kurbatov@jinr.ru
Токарева Виктория Андреевна[2] victoria.tokareva@kit.edu
Цирков Дмитрий Алексеевич[1]⋆ cyrkov@jinr.ru
[1]Объединённый институт ядерных исследований, RU-141980 Дубна, Россия
[2]Технологический институт Карлсруэ, DE-76021 Карлсруэ, Германия

Для определённого круга задач вычислительной математики, в том числе, для проблемы кинематического фитирования данных в физике частиц, является полезным расширение системы уравнений за счет введения т. н. уравнений связи вида $\varphi_k(\boldsymbol{p}) = \varphi_k(p_1, \ldots, p_n) = 0$, задающих известные из предметной области соотношения между параметрами $\boldsymbol{p}$. Чаще всего они оказываются нелинейными и вычислительно сложными, и на практике оказывается невозможным либо непрактичным снизить размерность системы, непосредственно решая уравнения связи.

FUMILI — это алгоритм оптимизации и пакет программного обеспечения, предназначенный для нахождения экстремумов квадратичных функционалов вида $F(\boldsymbol{x}, \boldsymbol{p}) = \sum_i f_i^2(x_i, \boldsymbol{p})$, возникающих при решении задач оптимизации методами наименьших квадратов или максимального правдоподобия. В данной работе рассматривается расширение алгоритма оптимизации FUMILI, названное методом исключения дифференциалов, позволяющее решать задачу оптимизации для функционалов с произвольным числом связей. Программно метод был реализован как расширение существующего кода FUMILI.

В докладе будут рассмотрены возможные применения данного метода на примерах кинематического фитирования моделированных и экспериментальных данных исследований физики частиц. Показано, что его использование позволяет добиться значительного прироста точности при нахождении искомых параметров.

[1] *Kurbatov V., Tokareva V., and Tsirkov D.* FUMILI-based minimization with constraints using method of elimination of differentials // EPJ Web Conf. 201, EDP Sciences, 2019. — p. 07001.

Extension of the FUMILI algorithm for optimization of quadratic functionals with constraints on parameters

Vladimir Kurbatov[1] kurbatov@jinr.ru
Victoria Tokareva[2] victoria.tokareva@kit.edu
Dmitry Tsirkov[1]⋆ cyrkov@jinr.ru
[1]Joint Institute for Nuclear Research, RU-141980 Dubna, Russia
[2]Karlsruher Institut für Technologie, DE-76021 Karlsruhe, Germany

For a certain range of problems in computational mathematics, including the problem of kinematic fitting of data in particle physics, it is useful to expand the system of equations by introducing the so-called constraints in the form of equations $\varphi_k(\boldsymbol{p}) = \varphi_k(p_1, \ldots, p_n) = 0$ that determine relations between the parameters $\boldsymbol{p}$, known from the subject area. Most often, they turn out to be nonlinear and computationally complex, and it is impossible or impractical to reduce the dimensionality of the system by solving the constraint equations directly.

FUMILI is an optimization algorithm and a software package designed for finding the extrema of quadratic functionals of the form $F(\boldsymbol{x}, \boldsymbol{p}) = \sum_i f_i^2(x_i, \boldsymbol{p})$ that arise in solving optimization problems by the methods of least squares or maximum likelihood estimation. In this work we consider the extension of the FUMILI optimization algorithm, called the method of elimination of differentials, that allows to solve search for the extrema of functionals with an arbitrary number of constraints. A software realization of the method has been implemented as an extension to the existing FUMILI code.

The talk will consider possible applications of this method using examples of kinematic fitting of simulated and experimental data from particle physics studies. It is shown that its use allows to achieve a significant improvement in accuracy when finding the desired parameters.

[1] *Kurbatov V., Tokareva V., and Tsirkov D.* FUMILI-based minimization with constraints using method of elimination of differentials // EPJ Web Conf. 201, EDP Sciences, 2019. — p. 07001.

Динамические байесовские сети как инструмент тестирования веб-приложений методом фаззинга

Азарнова Татьяна Васильевна[1]⋆ ivdas92@mail.ru
Полухин Павел Валерьевич[1] alfa_force@bk.ru
[1]Воронеж, Воронежский Государственный Университет

В современных условиях развитие информационной сферы и веб-технологий занимает значительное место в процессе создания средств хранения и обработки данных в различных отраслях экономики, производства и образования. Веб-технологии создают оптимальные возможности для обеспечения доступности, функциональности и своевременности доступа к обрабатываемой информации. Наряду с явными преимуществами веб-приложения обладают недостатками, связанными проблемами безопасности. В настоящее время существует достаточно обширная классификация уязвимостей веб-приложений OWASP и MITRE, объединяющая различные по способу обнаружения и степени критичности ошибки. Для обнаружения ошибок веб-приложений используются специальные методы тестирования, особое место среди которых занимает технология фаззинга. Под фаззингом понимается процесс тестирования, заключающийся в мониторинге за внедрением случайных данных, призванных вызвать событие сбоя или ошибки функционирования веб-приложения. Метод фаззинга позволяет накапливать статистические данные по результатам тестирования, обработка которых специальными математическими методами может позволить осуществлять непрерывную оптимизацию процедуры тестирования. Современные направления процедур тестирования веб-приложений методом фаззинга во многом связаны с развитием математических методов анализа стохастических процессов, адекватно моделирующих основные этапы, подпроцессы и параметры тестирования. Процедура фаззинга использует дискретное представление времени. Одним из современных, но в то же время хорошо апробированных методов моделирования стохастических процессов с дискретным временем, являются динамические байесовские сети (ДБС). В данной работе рассмотрены основные аспекты использования и настройки оптимальных параметров моделей динамических байесовских сетей для тестирования веб-приложений методом фаззинга. ДБС могут быть рассмотрены в виде последовательности статических байесовских сетей (БС). БС представляет собой иерархическую графическую вероятностную модель, описывающуюся стохастическими условными связями между дочерними и родительскими вершинами. Семантика совместного распределения срезов ДБС может быть представлена в следующем виде

$$P(X_{i:T}) = \prod_{t=1}^{T} \prod_{i=1}^{N} P(X_i^t) Parents(X_i^t). \quad (1)$$

где X_t^i – вершины ДБС – случайные величины, представленные таблицами условных вероятностей, $Parents(X_t^i)$ - родительские вершины для X_t^i. Для построения байесовских сетей используется цепное правило, которое можно представить следующим образом

$$P(x_1,\ldots,x_n) = P(x_n|x_{n-1},\ldots,x_1)\times P(x_{n-1}|x_{n-2},\ldots,x_1)\times \\ \ldots\times P(x_2|x_1)P(x_1) = \prod_{i=1}^{n} P(x_i|x_{i-1},\ldots,x_1,\ldots,x_1). \quad (2)$$

Настройка основных параметров ДБС осуществляется с помощью процедур обучения. Процедуры обучения ДБС подразделяется на обучение структуры и параметров. Обучение структуры направлено на определение вершин и формирования топологических связей между вершинами. В данной работе для процедура обучения структуры ДБС тестирования методом фаззинга предложено использовать тесты на условную независимость с оценкой значимости по критерию Пирсона и критерию знаков (G^2 критерий). Определение направленности связей между вершинами ДБС реализуется на основе использования Ньютоновских алгоритмов, в частности алгоритма Левенберга-Марквардта (ЛМ) в сочетании с методом Бройдена, позволяющего получать приближенные значения Якобиана для каждой итерации алгоритма ЛМ

$$\Delta\omega = [J^T J + \lambda diag[H^*]]^{-1} J^T e(\omega) \quad (3)$$

$$J_{n+1} = J_n + \frac{F(X_{n+1}) - F(X_n) - J_n}{zz^T} z^T, z = X_{n+1} - X_n \quad (4)$$

Процедура обучения параметров ДБС для формирования начального распределения условных вероятностей $P(X_0)$, модели перехода $P(X_i|X_{i-1})$,модели восприятия $P(E_i|X_i)$ осуществляется методом максимального правдоподобия. Для генерации тестовых выборок методом фаззинга и прогнозирования появления программных ошибок в приложениях с типовым набором компонентов, используются алгоритмы вероятностного вывода. Проведенные в рамках исследования эксперименты показали, что наиболее применимыми являются алгоритмы на основе взвешивания с учётом правдоподобия, в частности алгоритм многочастичного фильтра (МЧФ). МЧФ обладает высокой точностью, однако требует генерирования достаточно большого числа выборок. Для оптимизации классического МЧФ и снижения общего числа выборок, используется теорема Рао — Блеквелла — Колмогорова (РБК), которая в рамках МЧФ, сводится к сравнению весов выборок, полученных на каждом шаге генерации МЧФ фильтра.Каждый из весов выборки будет характеризоваться соответствующей достаточной статистикой T(X)

$$W(X_{t+1}^{'}, X_{t+1}^{''}|E_{1:t+1}) = P(E_{t+1}|X_{t+1}^{''}, X_{t+1}^{'})N'(X_{t+1}^{''}, X_{t+1}^{'}|E_{1:t}) \quad (5)$$

$$\mathbb{D}(W(X_{t+1}^{'}|E_{t+1})) \leqslant \mathbb{D}(W(X_{t+1}^{'}, X_{t+1}^{''}|E_{t+1})) \quad (6)$$

В процессе выполнения исследования для каждого типа программных ошибок веб-приложений происходит построение ДБС с учётом специфики тестирования, а также имеющихся механизмов сканирования и детектирования программных ошибок. Использование такого подхода позволяет повысить точность восстановления распределений вероятностей $P(X_0)$, $P(X_i|X_{i-1})$, $P(E_i|X_i)$ за счет решения задач фильтрации, прогнозирования и сглаживания, а также исключить узлы, не вносящие вклад в искомое распределения вероятностей. В процессе реализации процедуры вероятностного вывода, модель на основе ДБС позволяет реализовать структурированное накопление данных относительно веб-приложений с типовым набором компонентов, а также осуществлять формирование прогнозов относительно наличия ошибок.

[1] *Azarnova T. V., Polukhin P. V.* Advanced hybrid stochastic dynamic Bayesian network inference algorithm development in the context of the web applications test execution // IOP Conf. Ser., Bristol: IOP Publishing, 2019. — С. 052028–052035.

Dynamic Bayesian networks as a framework for web-applications fuzzing testing

Azarnova Tatyana Vasilyevna[1]⋆ ivdas92@mail.ru
Polukhin Pavel Valerievich[1] alfa_force@bk.ru
[1]Voronezh, The Voronezh State University

The modern conditions in the information sphere elaboration and web technologies, occupies a significant place in the process of creating data storage and processing facilities in various sectors of the economy, production and education. Web technologies create optimal opportunities for ensuring availability, functionality and timeliness of access to the processed information. Along with the obvious advantages, web applications have drawbacks related to security issues. Currently, there is a fairly extensive classification of web application vulnerabilities OWASP and MITRE, combining different methods of detection and severity of errors. To detect errors in web applications, special testing methods are used, a special place among in which occupies fuzzing technology. Fuzzing is closely related to the testing process, which represents a monitoring procedure of random data generation algorithm, designed to cause a failure or error event under the functioning of a web application. Fuzzing method allows us to accumulate statistical data of the testing results, processing of which by special mathematical methods can allow continuous optimization of the testing procedure. Modern directions of web application fuzzing testing procedures are tightly bound with the development of mathematical methods, especially for the stochastic processes analysis that properly simulate the main stages, subprocesses and testing parameters. The fuzzing procedure uses a discrete representation of time. One of the modern, but at the same time well-tested methods of stochastic processes modeling with discrete time are dynamic Bayesian networks (DBN). In this paper, we consider the main aspects of using and configuring the optimal parameters of dynamic Bayesian network models for web applications fuzzing testing. This approach is due to the necessity for building a time distributed model, as the testing process can be distributed over a wide time interval. DBN can be considered as a sequence of static Bayesian networks (BN). BN is a hierarchical graphical probabilistic model described by stochastic conditional relationships between child and parent vertices. Semantics of joint distribution of DBN slices can be represented in the following form

$$P(X_{i:T}) = \prod_{t=1}^{T} \prod_{i=1}^{N} P(X_i^t) Parents(X_i^t). \quad (1)$$

where X_t^i — DBN vertex — random variables, represented by tables of conditional probabilities, $Parents(X_t^i)$ is the set of the parent elements for X_t^i. The common way for building Bayesian network is based on a chain rule, which can be represented

as follows

$$P(x_1,\ldots,x_n) = P(x_n|x_{n-1},\ldots,x_1) \times P(x_{n-1}|x_{n-2},\ldots,x_1)\times \\ \ldots \times P(x_2|x_1)P(x_1) = \prod_{i=1}^{n} P(x_i|x_{i-1},\ldots,x_1,\ldots,x_1). \tag{2}$$

Setting the basic parameters of the DBN is carried out by learning procedures. The DBN learning procedure is divided into the structure and parameters learning. Structure learning be intended to defining vertices and forming topological relationships between vertices. In this research, for the structure learning procedure of fuzzing testing, DBN model set forward to use conditional independence tests with evaluation of significance by Pearson's criterion and the criterion of signs (G^2 criterion). Determination of relationships orientation between the vertices of DBN is implemented through the use of Newtonian algorithms, especially Levenberg — Marquardt (LM) algorithm combined with the Broyden method

$$\Delta\omega = [J^T J + \lambda diag[H^*]]^{-1} J^T e(\omega) \tag{3}$$

$$J_{n+1} = J_n + \frac{F(X_{n+1}) - F(X_n) - J_n}{zz^T} z^T, z = X_{n+1} - X_n \tag{4}$$

It is important to note that the Broyden method is the most adapted optimization approach for the LM algorithm, as it allows to obtain approximate values of the Jacobi matrix for each iteration of the algorithm. In particular, this allows to parallelize individual blocks, during the execution of the LM algorithm (Jacobian calculation), the total number of rows of the matrix will be processed $N_j/n_p, N_j$ — the total number of rows of the Jacobian matrix that require calculation, n_p — the number of available parallel processes within the computing system.

The procedure of DBN parameters learning for initial conditional probability distribution $P(X_0)$, transition model $P(X_i|X_{i-1})$, perception model $P(E_i|X_i)$ formation is carried out by the maximum likelihood method. Probabilistic inference algorithms are used to generate fuzzing test samples and predict the software errors appearance of the applications with a same typical set of components. Performed experiments within a framework of the research, showed that the most applicable algorithms are based on the likelihood weighing approach, in particular the multiparticle filter algorithm (MPF). The MPF is highly accurate, but requires the generation of a sufficiently large number of samples. To optimize the classical MPF and reduce the total number of samples, the Rao — Blackwell — Kolmogorov (RBK) theorem is used, which in the framework of the MPF reduces to comparing the samples weights, obtained at the each step of the MPF filter generation. Each of several sample weights will be characterized by a corresponding sufficient statistic $T(X)$

$$W(X_{t+1}^{'}, X_{t+1}^{''}|E_{1:t+1}) = P(E_{t+1}|X_{t+1}^{''}, X_{t+1}^{'})N'(X_{t+1}^{''}, X_{t+1}^{'}|E_{1:t}) \tag{5}$$

$$\mathbb{D}(W(X_{t+1}^{'}|E_{t+1})) \leqslant \mathbb{D}(W(X_{t+1}^{'}, X_{t+1}^{''}|E_{t+1})) \tag{6}$$

During research process for each type of the web applications software errors, the DBN is built taking into account the specifics of testing, as well as the available scanning mechanisms and detecting software errors. Such approach employment allows to increase the accuracy of the probability distributions $P(X_0)$, $P(X_i|X_{i-1})$, $P(E_i|X_i)$ by solving the filtering, predicting and smoothing problems, as well as to exclude nodes that do not contribute to the desired probability distribution. In the process of implementing the probabilistic inference procedure, the DBN model allows to perform a structured data accumulation of web applications with a typical set of components, as well as to make predictions about the errors existence.

[1] *Azarnova T. V., Polukhin P. V.* Advanced hybrid stochastic dynamic Bayesian network inference algorithm development in the context of the web applications test execution // IOP Conf. Ser., Bristol: IOP Publishing, 2019. — P. 052028–052035.

Применение методов интеллектуального анализа данных в оценке функциональной эффективности команд менеджеров

Азарнова Татьяна Васильевна[1]⋆ ivdas92@mail.ru
Аснина Наталия Георгиевна[2] andrey050569@yandex.ru
Бондаренко Юлия Валентиновна[1] bond.julia@mail.ru
[1]Воронеж, ФГБОУ ВО «Воронежский государственный университет»
[2]Воронеж, ФГБОУ ВО «Воронежский государственный технический университет»

Применение команд менеджеров в управлении проектами можно рассматривать как устоявшуюся тенденцию современного бизнеса. Результативность проектов напрямую зависит от эффективности команд менеджеров. При формировании команды необходимо наряду с профессиональными знаниями учитывать ролевые особенности всех входящих в неё членов, анализировать возможность их плодотворной, результативной совместной работы. В данной работе предложена система алгоритмов формирования функционально-эффективных команд менеджеров, базирующихся на методах поиска ассоциативных правил, методике Р. М. Белбина [1] для оценки ролевых профилей менеджеров и нейросетевой технологии распознавания функциональной эффективности команд менеджеров по ролевому составу [2, 3]. Для решения задачи оценки ролевых профилей менеджеров команды была выбрана методика британского доктора психологических наук Р. М. Белбина. Сущность данной методики заключается в специальном тестировании всех членов команды, на основании которого для каждого члена команды строится его ролевой профиль, представляющий собой вектор, каждая координата которого, отражает бальную оценку выраженности у данного члена команды таких ролей как: генератор идей, аналитик, исследователь ресурсов, координатор, контролер, реализатор, мотиватор, вдохновитель команды. На основании ролевых профилей членов команды в работе предлагается строить оценку ролевого состава команды, представленную в виде матрицы близости ролевых профилей. В качестве данной матрицы используется либо матрица расстояний, либо матрица коэффициентов ранговой корреляции Спирмена. Эффективность работы команды менеджеров оценивается по функциональной модели оценки менеджмента (ФМОМ или MFAM). В основе MFAM лежат шесть основных функций менеджмента: планирование, организация, мотивация, контроль, координация и коммуникация. Данные функции формируют структуру управления — характер взаимосвязей команды, ее коммуникационный профиль, и оцениваются экспертным тестированием. В рамках исследования делается попытка построить интеллектуальные инструменты прогнозирования функциональной эффективности команды менеджеров по ее ролевому составу. В качестве инструментов выбраны нейронные сети. На вход нейросети подается матрица ролевого состава команды, а на выход: оценка функциональной эффективности команды — вектор, компоненты которого отражают конкретные значения по каждому критерию функциональной моде-

ли оценки менеджмента, коммуникационный профиль и уровень состояния менеджмента в команде. В качестве алгоритма обучения нейронной сети в работе используется алгоритм обратного распространения ошибок (back propagation).

В исследовании также представлены алгоритмы обучения на основе ассоциативных правил (Associations rules learning — ARL). Ассоциативные правила используются для поиска ролевого состава команд, которые наиболее часто выступают как эффективные, и принятия решений об изменении ролевого состава команды для повышения ее потенциальной эффективности. Пусть $I = \{i_1, i_2, \ldots, i_R\}$ — множество ролей по Белбину, $D = \{T_1, T_2, \ldots, T_K\}$ — множество эффективных команд менеджеров (база данные для обучения правил). Каждая команда в базе данных D имеет уникальный идентификатор ID и состоит из подмножества объектов (ролей) из множества I, т. е. $T_k \subset I$. Команда T_k представляется в виде бинарного вектора, где

$$t_r = \begin{cases} 1, & \text{если соответствующая роль} \quad i_r \subset T_k, \\ 0, & \text{если соответствующая роль} \quad i_r \not\subset T_k. \end{cases}$$

Ассоциативные правила для рассматриваемой задачи представляют собой импликации вида $X \to Y$, где $X \subset I$, $Y \subset I$ и $X \cap Y = \varnothing$, X, Y — составы команд менеджеров. Правила $X \to Y$ имеют поддержку $S_{X \to Y}$, представляющую собой процент эффективных команд менеджеров из базы данных D, которые содержат рассматриваемый набор ролей по Белбину $X \cup Y$. Правила характеризуются также достоверностью — показатель, характеризующий насколько часто ассоциативное правило оказывается верным. Целью анализа функционально эффективных команд менеджеров является выявление следующих зависимостей: если в команде менеджеров имеется множество ролей по Белбину X, то на основании этого можно сделать вывод, что множество ролей Y также в ней должно быть.

Задачу определения ассоциативных правил можно разделить на две.

1. Нахождение всех наборов ролей по Белбину, которые удовлетворяют минимальному порогу поддержки. Такие роли называются *часто встречающимися*.

2. Генерация ассоциативных правил из множества часто встречающихся ролей из п. 1 с достоверностью, удовлетворяющей минимальному порогу достоверности.

Для генерации ассоциативных правил в работе используется алгоритм Ракеш Агравала (Rakesh Agrawal), имеющий название *Apriori*.

[1] *Белбин М.* Команды менеджеров. Секреты успехов и причины неудач, Москва: HIPPO, 2003. — 315 с.

[2] *Азарнова Т. В.* Информационные аналитические приложения для оценки человеческого потенциала организационных систем // Современная экономика: проблемы и решения. — 2017. — No 10 (94). — С. 14–32.

[3] *Азарнова Т. В.* Нейросетевой алгоритм оценки функциональной эффективности команд менеджеров на основе их ролевого состава // Вопросы науки. — 2017. — No 2. — С. 4–18.

Application of data mining methods in assessing the functional effectiveness of manager teams

Tatiana Azarnova[1]⋆ ivdas92@mail.ru
Natalia Asnina[2] andrey050569@yandex.ru
Yulia Bondarenko[1] bond.julia@mail.ru
[1]Voronezh, Federal State Budgetary Educational Institution of Higher Education "Voronezh State University"
[2]"Voronezh, Federal State Budgetary Educational Institution of Higher Education "Voronezh State Technical University

The use of managerial teams in project management can be considered as an established trend in modern business. The effectiveness of projects directly depends on the effectiveness of management teams. When forming a team, it is necessary, along with professional knowledge, to take into account the role features of all its members, to analyze the possibility of their fruitful, productive collaboration. In this paper, we propose a system of algorithms for the formation of functionally effective managerial teams based on the methods of searching for associative rules, the method of R. M. Belbin [1] for assessing role profiles of managers and neural network technology for recognizing the functional effectiveness of role management teams [2, 3]. To solve the problem of assessing the role profiles of team managers, the methodology of the British doctor of psychological sciences R. M. Belbin was chosen. The essence of this technique consists in special testing of all team members, on the basis of which for each team member is built his role profile, which is a vector, each coordinate of which reflects a ball rating for the severity of this team member of such roles as: idea generator, analyst, resource researcher, coordinator, controller, implementer, motivator, team inspirer. Based on the role profiles of team members, it is proposed to build an assessment of the role composition of the team, presented in the form of a proximity matrix of role profiles. As this matrix, either the distance matrix or the Spearman rank correlation coefficient matrix are used. The effectiveness of the team of managers is assessed by the functional model of management assessment (MFAM). MFAM is based on six core management functions: planning, organization, motivation, control, coordination and communication. These functions form the management structure — the nature of the team's relationships, its communication profile, and are evaluated by expert testing. As part of the study, an attempt is made to build intelligent tools for predicting the functional effectiveness of a team of managers by its role composition. Neural networks were selected as tools. The matrix of the team's role structure is fed to the input of the neural network, and the output: the assessment of the team's functional effectiveness is a vector whose components reflect specific values for each criterion of the functional model of management assessment, the communication profile and the level of management status in the team. As an algorithm for training a neural network, the work uses the back propagation algorithm.

The study also presents learning algorithms based on associative rules (Associations rules learning — ARL). Associative rules are used to search for the role composition of teams that most often act as effective, and make decisions on changing the role composition of a team to increase its potential effectiveness. Let $I = \{i_1, i_2, \ldots, i_R\}$ — many roles according to Belbin, $D = \{T_1, T_2, \ldots, T_K\}$ — many effective management teams (database for training rules). Each command in database D has a unique identifier ID and consists of a subset of objects (roles) from set I, i.e. $T_k \subset I$. The command T_k is represented as a binary vector, where

$$t_r = \begin{cases} 1, & \text{if the relevant role} \quad i_r \subset T_k, \\ 0, & \text{if the relevant role} \quad i_r \not\subset T_k. \end{cases}$$

The associative rules for the problem under consideration are implications of the form $X \to Y$, where $X \subset I$, $Y \subset I$ and $X \cap Y = \varnothing$, X, Y — are the teams of managers. The rules $X \to Y$ have support $S_{X \to Y}$, which is the percentage of effective managerial teams from the D database that contain the considered set of roles according to Belbin $X \cup Y$. The rules are also characterized by certainty — an indicator that characterizes how often an associative rule is true. The purpose of the analysis of functionally effective managerial teams is to identify the following dependencies: if the managerial team has many roles according to Belbin X, then based on this we can conclude that the set of roles Y should also be in it. The task of defining associative rules can be divided into two.

1. Finding all sets of Belbin roles that satisfy the minimum support threshold. Such roles are called *frequent*.

2. Generation of associative rules from the set of frequently occurring roles from clause 1 with confidence satisfying the minimum threshold of reliability.

To generate associative rules, the work uses the Rakesh Agrawal algorithm, called *Apriori*.

[1] *Belbin M.* Management teams. Secrets of success and reasons for failure, Moscow: HIPPO, 2003. — 315 p.

[2] *Azarnova T. V.* Information analytical applications for the evaluation of the human potential of organizational systems // Modern economy: problems and solutions, 2017. — No. 10(94). p. 14–32.

[3] *Azarnova T. V.* Neural algorithm for assessing the functional efficiency of management teams based on their role structure // Science issues, 2017. — No. 2. — p. 4–18.

Метод градиентного спуска на основе многомерных воксельных образов

Толок Алексей Вячеславович[1,2] author_tolok_61@mail.ru
Толок Наталия Борисовна[2]⋆ nat_tolok@mail.ru
[1]Москва, МГТУ "Станкин"
[2]Москва, ИПУ им. В.А. Трапезникова РАН

Метод градиентного спуска (МГС) как правило используется при моделировании решений в экстремумах целевых функций оптимизационных задач. Многомерные постановки таких задач приводят к поиску компьютерных решений МГС, позволяющих проводить многопараметрическую оценку выбираемого направления движения. Существует множество подходов к реализации МГС на компьютере из которых можно выделить основные: метод наискорейшего спуска, метод координатного спуска Гаусса – Зейделя, метод сопряжённых градиентов и т.п. Все эти методы базируются на последовательном вычислении дифференциальных характеристик в текущем положении точки поверхности, что накладывает на алгоритм соблюдение различных дополнительных условий для корректной работы алгоритма расчёта в зависимости от компьютерного представления поверхности исследуемой функции. Значительно возрастает сложность таких подходов при увеличении размерности поставленной задачи.

В предложенной статье разбирается один из подходов к автоматизации МГС, базирующий свою работу на использовании предварительно рассчитанных локальных геометрических характеристиках функциональной области, представленных конечным набором воксельных образов, соразмерных с этой областью. Информационной базой для работы алгоритма являются компьютерные геометрические модели, полученные на основе метода функционально-воксельного моделирования [1]. Демонстрируются принципы управления градиентным движением, а также особенности перехода к увеличению размерности задач. Определены преимущества работы алгоритма, позволяющие расширить использование такого подхода к задачам локальной оптимизации на примере прокладки трасс максимального скоростного спуска. Приведён пример реализации решения эконометрической задачи с применением предложенного подхода.

[1] *Толок А. В.* Функционально-воксельный метод в компьютерном моделировании // Монография, Москва. ФИЗМАТЛИТ, 2016. – 112 с., ISBN: 978-5-9221-1680-0.

Gradient descent method based on multidimensional voxel images

Alexey Tolok[1,2] author_tolok_611@mail.ru
Natalya Tolok[2]⋆ nat_tolok@mai.ru
[1]Moscow, MSTU "STANKIN"
[2]Moscow, ICS V.A. Trapeznikov RAS

The gradient descent method (MGS) is usually used in modeling solutions at extremum points of objective functions of optimization problems. Multidimensional formulations of such problems lead to the search for computer solutions of the MGS, allowing multi-parameter estimation of the chosen direction of movement. There are many approaches to the implementation of MGS on a computer from which the main ones can be distinguished: the steepest descent method, the Gauss - Seidel coordinate descent method, the conjugate gradient method, etc. All these methods are based on sequential calculation of differential characteristics in the current position of a surface point, which imposes on the algorithm the observance of various additional conditions for the correct operation of the calculation algorithm depending on the computer representation of the surface of the function. The complexity of such approaches increases significantly with an increase in the multi-dimension of the task. In the proposed article, one of the approaches to the automation of MGS is analyzed, basing its work on the use of pre-calculated local geometric characteristics of the functional area, represented by a finite set of voxel images commensurate with this area. The information base for the algorithm is computer geometric models obtained on the basis of the functional voxel modeling method [1]. The principles of gradient motion control, as well as the features of the transition to increasing the dimension of tasks are demonstrated. The advantages of the algorithm, which allow to expand the use of this approach to the problems of local optimization by the example of laying tracks of maximum speed descent, are determined. An example of the implementation of the solution of the econometric problem using the proposed approach.

[1] *Tolok A.* Functional voxel method in computer simulation // monograph, Moscow. FIZMATLIT, 2016 .– 112 pp., ISBN: 978-5-9221-1680-0.

Содержание

Contents

Авторский указатель

Х

Ц

Ч

Ш

Я

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.

Журнал «Машинное обучение и анализ данных»

http://jmlda.org

Журнал Машинное обучение и анализ данных публикует новые теоретические и обзорные статьи с результатами научных исследований в области искусственного интеллекта, теоретической информатики и приложений. Цель журнала — развитие теории машинного обучения, интеллектуального анализа данных и методов проведения вычислительных экспериментов. Принимаются статьи на русском и английском языках.

Научное издание

Математические методы распознавания образов

Тезисы докладов
19-й Всероссийской конференции
с международным участием

Подписано в печать 19.11.2019
Формат 60×84 1/8
Усл.-печ. л. 20,1. Уч.-изд. л. 21,17
Тираж 200 экз

Издатель — Российская Академия Наук

Печать — УНИД РАН

Отпечатано в экспериментальной цифровой типографии РАН

Издается по распоряжению президиума РАН
и распространяется бесплатно