Российская академия наук
Вычислительный центр РАН ФИЦ ИУ РАН

Математические методы распознавания образов

20-я Всероссийская конференция с международным участием

Москва, 2021

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 20-й Всероссийской конференции с международным участием, г. Москва 2021 г. — М.: Российская академия наук, 2021. — 508 с.

ISBN 978-5-907366-47-3

В сборнике представлены тезисы докладов 20-й Всероссийской конференции «Математические методы распознавания образов», проводимой Российской академией наук, Вычислительным центром Федерального исследовательского центра «Информатика и управление» РАН.
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 978-5-907366-47-3

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 20th Russian National Conference with International Participation, Moscow, 2021. — Moscow: Russian Academy of Sciences, 2021. — 508 p.

ISBN 978-5-907366-47-3

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.
The conference has being held biennially since 1989. It is one of the most recognizable scientific forums on data mining, machine learning, pattern recognition, image analysis, signal processing, and discrete analysis.

The conference website `http://mmro.ru/en/`.

ISBN 978-5-907366-47-3

Оргкомитет

Сопредседатели: Журавлев Юрий Иванович, *акад. РАН*, *ФИЦ ИУ РАН*
Соколов Игорь Анатольевич, *акад. РАН* *ФИЦ ИУ РАН*

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

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

Программный комитет

Сопредседатели: Стрижов Вадим Викторович, *д.ф.-м.н.* *ФИЦ ИУ РАН*
Воронцов Константин Вячеславович, *д.ф.-м.н.* *МФТИ (НИУ)*

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

Organizing Committee

Chairs: Yury Zhuravlev, *acad. of RAS*, *FRCCSC*
Igor Sokolov, *acad. of RAS* *FRCCSC*

Secretary: Yury Chekhovich, *C.Sc.*

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

Program Committee

Chair: Vadim Strijov, *D.Sc.*, *FRCCSC*
Konstantin Vorontsov, *D.Sc.* *MIPT (NRU)*

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

Интеллектуальный анализ данных 10

Нейронные сети и глубокое обучение 117

Методы оптимизации для интеллектуального анализа данных 145

Вычислительная сложность и приближенные методы 155

Обработка и анализ изображений и сигналов, компютерное зрение 187

Информационный поиск и анализ текстов 288

Анализ данных веба и социальных сетей 368

Индустриальные приложения науки о данных 374

Анализ биомедицинских данных, биоинформатика 399

Методы математического моделирования в интеллектуальном анализе данных 438

Интеллектуальный анализ геопространственных данных 442

Интеллектуальная оптимизация и эффективный менеджмент 446

Brief contents

Data mining 10

Neural networks and deep learning 117

Data mining optimization techniques 145

Algorithmic complexity and approximate methods 155

Image and signal processing, computer vision 187

Information retrieval and text analysis 288

Analysis of web and social network data 368

Industrial data science applications 374

Analysis of biomedical data, bioinformatics 399

Methods of mathematical modeling in data mining 438

Geospatial data mining 442

Intelligent optimization and effective management 446

Метрическая коррекция парных сравнений на основе прямого изменения собственных значений

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

В современных задачах машинного обучения и интеллектуальной обработки исходные экспериментальные данные часто представлены парными сравнениями между элементами множества в виде функций сходства или различия. Положительная определенность матрицы парных сравнений в случае измерения сходства или эквивалентной матрицы сходства в случае измерения различий позволяет говорить о вложенности множества в некоторое пространство, в котором элементы множества образуют корректную конфигурацию без метрических нарушений.

Нарушения конфигурации приводят к появлению отрицательных собственных значений при разложении матрицы сходства по ортонормированному базису ее собственных векторов. В этом случае обычно применяется известное дискретное разложение Карунена-Лоэва для снижения размерности матрицы сходства за счет устранения вкладов собственных векторов, соответствующих отрицательным собственным значениям. В итоге, в редуцированном пространстве собственных векторов можно вычислить новую положительно определенную матрицу сходства. При этом дисперсия нормированных данных уменьшается.

В рамках развиваемого нами подхода [1], в данной работе предложено не устранять отрицательные собственные значения, а непосредственно изменять их прямо на положительные. При таком подходе размерность матрицы сходства не изменяется, т.е. дисперсия нормированных данных также не изменяется. Также не нужно заново вычислять новую матрицу сходства в редуцированном пространстве собственных векторов. Более того, часто это и невозможно сделать из-за отсутствия собственно исходной матрицы данных, представляющей измерения признаков (характеристик объектов), которые нужно проецировать в новое ортонормированное пространство.

Но такое прямое изменение собственных значений не должно быть произвольным. Новые собственные значения следует определить из условий, которым должна соответствовать измененная матрица парных сравнений. В работе показано, как, например, выбрать новые собственные значения, чтобы матрица парных сравнений при допустимом уровне изменения значений ее элементов обладала бы лучшей обусловленностью.

Отметим, что в рамках известной теории возмущений в проблеме собственных значений [2] оценивается влияние матрицы возмущений на возмущения собственных значений и соответствующих им собственных векторов симметрической матрицы. В нашей задаче метрической коррекции предлагается действо-

вать в обратном порядке, когда оценивается влияние «возмущения» от коррекции собственных значений на свойства восстановленных матриц парных сравнений.

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

[1] *Двоенко С. Д. Пшеничный Д. О.* Исправление и коррекция матриц парных сравнений // Известия ТулГУ. Технические науки, 2021, Вып. 12. С. 133–143.

[2] *Rellich F.* Perturbation Theory of Eigenvalue Problems // NY: Gordon and Breach, 1969. 138 p.

Metric correction of paired comparisons based on direct changing of eigenvalues

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

In modern problems of machine learning and intelligent data processing, raw experimental data are often represented by paired comparisons between elements of a set in the form of similarity or dissimilarity functions. The positive definiteness of the matrix of paired comparisons in the case of similarity measurement or the equivalent similarity matrix in the case of dissimilarity measurement allows us to talk about the immersing of the set in some space in which the set elements form a correct configuration without metric violations.

Violations in the set configuration lead to the appearance of negative eigenvalues when the similarity matrix is decomposed based on the orthonormal basis of its eigenvectors. In this case, the well-known discrete Karhunen-Loeve expansion is usually used to reduce the dimensionality of the similarity matrix by eliminating contributions of eigenvectors corresponded to negative eigenvalues. As a result, a new positive definite similarity matrix can be calculated in the reduced space of eigenvectors. But the variance of normalized data decreases at the same time.

Within the framework of our approach under developing [1], in this paper it is proposed not to eliminate negative eigenvalues, but directly to change them to positive ones. Based on this approach, the dimensionality of the similarity matrix does not be changed, and therefore the variance of normalized data also does not be changed too. Additionally, it does not need to recalculate a new similarity matrix in the reduced space of eigenvectors. Moreover, it is often impossible to do this due to the absence of the original data matrix itself, which represents measurements of features (characteristics of objects) that need to be projected into a new orthonormal space.

But such a direct changing of eigenvalues should not be arbitrary. New eigenvalues should be determined according to conditions to which the modified matrix of paired comparisons should correspond. In this paper, for example, we show how to determine new eigenvalues so that the matrix of paired comparisons with an acceptable level of changing in its elements would have better conditionality.

Note, in the framework of the well-known Perturbation Theory of Eigenvalue Problems [2], the influence of the perturbation matrix on perturbations of eigenvalues and corresponding eigenvectors of the symmetric matrix is estimated. In our problem of metric correction, we propose actually the reversed approach, when the influence of the "perturbation" from the corrected eigenvalues on the properties of the restored matrices of paired comparisons is estimated.

This research is funded by RFBR, grants 20-07-00055, 19-07-01178.

[1] *Dvoenko S. Pshenichny D.* A recovering and correction of matrices of paired comparisons // Izvestiya of the TSU. Tech. Sciences, 2021. Vol 12. Pp. 133–143.

[2] *Rellich F.* Perturbation Theory of Eigenvalue Problems // NY, Gordon and Breach, 1969. 138 p.

Проверка согласованности метрик качества изображений

Двоенко Сергей Данилович[1] sergedv@yandex.ru
Курбаков Михаил Юрьевич[1]⋆ muwsik@mail.ru
[1]Тула, Тульский государственный университет

В задачах анализа изображений требуется оценить качество их обработки. В современных условиях такие задачи и сами изображения оказываются весьма различными. Поэтому разные методы оценки качества изображений, которые называются метриками качества, в целом реализуют различные представления о том, что такое «более качественное изображение». В итоге, сравнение оценок различных метрик качества является актуальной проблемой.

С одной стороны, решение данной проблемы может быть достигнуто за счёт усреднения оценок по некоторому набору метрик. В таком случае формализация проблемы сопоставления метрик приводит к задаче многокритериальной оптимизации, что в общем случае может привести к проблеме Парето.

С другой стороны, индивидуальное восприятие качества изображения обычно не претендует на точную количественную оценку. Поэтому имеет смысл перейти к измерениям в менее мощных, например, порядковых шкалах. Это означает переход к ранжированиям, которые будут представлены в виде набора изображений, упорядоченных в соответствии с оценками метрик качества. Тогда решением задачи согласования ранжирований, например, на основе медианы Кемени [1], является ранжирование, которое имеет смысл аналога среднего арифметического индивидуальных ранжирований по метрикам качества.

Наш подход на основе ранжирований также позволяет оценить чувствительность совокупности метрик к изменению качества изображения и предполагает два вида экспериментов:

– определение таких видов шума, которые в совокупности по множеству метрик значительно влияют на качество изображений заданного класса;

– определение таких классов изображений, которые в совокупности по множеству метрик значительно подвержены влиянию шума заданного типа.

В данной работе на основе экспериментального исследования выдвигается гипотеза, что при согласованности выделенной группы метрик качества между собой и согласованности медианы Кемени с ранжированием на основе индивидуального восприятия, результаты первого эксперимента должны подтверждаться результатами второго эксперимента и наоборот.

В итоге, появляется возможность для обоснованных рекомендаций по применению наиболее предпочтительных метрик для оценки качества изображений при их обработке.

После обработки данного изображения некоторым множеством алгоритмов оценка качества некоторой метрикой позволит упорядочить обработанные изображения по значениям данной метрики и указать наилучшее, т.е. указать наилучший алгоритм обработки. Разные метрики формируют свои ранжирования

изображений и в общем случае указывают свои наилучшие изображения. Согласование ранжирований изображений в виде медианы Кемени позволит утверждать, что некоторый алгоритм обеспечивает наилучшее качество обработки данного изображения по совокупности данного множества метрик качества.

Работа поддержана грантом РФФИ No. 20-07-00055.

[1] *Dvoenko S., Pshenichny D.* Rank Aggregation Based on New Types of the Kemeny's Median // Pattern Recognition and Image Analysis 31, 2021. Pp. 185–196.

Checking the consistency of image quality metrics

Dvoenko Sergey[1] sergedv@yandex.ru
Kurbakov Mikhail[1]⋆ muwsik@mail.ru
[1]Tula, Tula State University

In the field of image analysis, it is required to evaluate the processing quality. Under modern conditions, such problems and images themselves turn out to be very different. Therefore, different methods of the quality evaluating of images, which are called quality metrics, generally implement different ideas about what a "better image" is. As a result, the actual problem consists in comparing of various quality metrics.

From the one side, the solution of this problem can be achieved by the averaging of estimates over a certain set of metrics. In this case, the formalization of the metric comparing problem leads to a multi-criteria optimization problem, which can lead to a Pareto problem in general.

On the other side, an individual perception of an image quality usually does not claim to be an accurate quantitative evaluation. Therefore, it may be suitable to use measurements in less powerful, for example, ordinal scales. This means using rankings presented as a set of images ordered according to the quality metrics scores. Then the solution of the problem of matching rankings, for example, based on the Kemeny's median [1], is a ranking that appears as an analogue of the arithmetic mean of individual rankings by quality metrics.

Our ranking-based approach also allows us to evaluate the sensitivity of a set of metrics to variability in image quality and supposes two types of experiments:

– to determine types of noise, which together, based on the set of metrics given, significantly affect the quality of images of a given class;

– to determine a class (set) of images, which together, based on the set of metrics given, are significantly affected by noise of a given type.

In this paper, on the basis of experiments, a hypothesis is put forward that if the selected group of quality metrics is coordinated with each other and the Kemeny's median is coordinated with a ranking according to an individual perception, results of the first experiment should be confirmed by results of the second experiment and vice versa.

As a result, there is an opportunity for reasonable recommendations to use the most preferred metrics for evaluating the quality of images under processing.

After processing of the image by some algorithms, evaluating the processing quality by some metric allows to arrange the processed images by values of this metric and to specify the best one, i.e. to specify the best processing algorithm. Different metrics form their own image rankings and generally indicate their best images. Comparing of rankings in the form of the Kemeny's median allows to confirm that some algorithm provides the best quality of processing of this image according to the given set of quality metrics.

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

[1] *Dvoenko S., Pshenichny D.* Rank Aggregation Based on New Types of the Kemeny's Median // Pattern Recognition and Image Analysis 31, 2021. Pp. 185–196.

Оценка близости выпуклых оболочек для задач машинного обучения

Немирко Анатолий Павлович apn-bs@yandex.ru

Санкт-Петербург, Санкт-Петербургский государственный электротехнический университетет "ЛЭТИ"

Многие задачи машинного обучения (регрессия, классификация, кластеризация, сокращение размерности признакового пространства) используют представление классов в виде множеств точек в многомерном признаковом пространстве. Другим способом описания классов является использование выпуклых оболочек исходных множеств. Несмотря на потерю статистических свойств данных, которая часто является несущественной, выпуклые оболочки определяют границы изменения экспериментальных данных, компактно хранят всю основную информацию о них, что позволяет быстро отвечать на разнообразные запросы на этом множестве. Такое геометрическое описание классов отражает ориентацию классов в пространстве, их взаимное расположение друг относительно друга и конфигурацию областей их пересечений, Основным препятствием развития этого направления является вычислительная сложность, быстро возрастающая при увеличении числа признаков и экспериментальных точек.

В работе рассмотрены методы машинного обучения при описании классов именно выпуклыми оболочками в многомерном признаковом пространстве. Найдены способы приближенной оценки локализации точки по отношению к границе выпуклой оболочки (внутри, на границе или вне выпуклой оболочки) без вычисления самой выпуклой оболочки [1,2]. Это позволило приближенно оценить расстояние от точки до границы выпуклой оболочки [3] независимо от ее локализации и реализовать метод ближайшей выпуклой оболочки. Предложены также эффективные методы вычисления множества экстремальных точек выпуклой оболочки [1], которые упрощают описания классов и реализацию решающих правил.

Предложенный недавно метод нахождения расстояния от точки до границы выпуклой оболочки [3] основан на применении линейного программирования. В отличие от NP-полных задач квадратичного программирования эта задача относится к P классу сложности и может решаться за слабо полиномиальное время. Для этого на практике существуют очень эффективные алгоритмы. Вышеперечисленные методы обладают хорошей масштабируемостью и могут применяться во многих задачах машинного обучения при описании классов в многомерном пространстве. Они с успехом опробованы на задачах автоматической медицинской диагностики рака груди и распознавания опасных аритмий по ЭКГ. Общая точность классификации на тестовой выборке для рака груди составила 97,9% (для SVM с линейным ядром 92,8%) и для опасных аритмий 91,5% (для SVM с линейным ядром 93,0%)

В данной работе предложены новые методы машинного обучения, основанные на приведенных выше результатах и не требующие вычисления выпуклых оболочек или триангуляции множества точек. Предложенная процедура локализации точки по отношению к границе выпуклой оболочки [1] упрощает решение задачи визуализации точек, попавших в зону пересечения выпуклых оболочек. Это позволяет уточнить возможность увеличения эффективности решающего правила. В работе на модельных и реальных данных показаны примеры отображения на плоскости точек, попавших в зону пересечения двух выпуклых оболочек, описывающих классы в многомерном пространстве.

Предложен также простой способ оценки близости двух выпуклых оболочек друг к другу независимо от их пересекаемости, который основан на приведенных выше результатах. Ищется эвклидово расстояние между границами двух выпуклых оболочек, на прямой линии, соединяющей два центроида рассматриваемых классов. Оно вычисляется как разность двух расстояний от центроидов классов до границ выпуклых оболочек противоположных классов. Эти расстояния определяются с использованием метода линейного программирования. Этот способ применен как мера расстояния при объединения промежуточных кластеров на основе ближайших выпуклых оболочек в кластерном анализе на базе выпуклых оболочек.

Работа поддержана грантом РФФИ No. 19-29-01009.

[1] *Dulá J., Helgason R.* A new procedure for identifying the frame of the convex hull of a finite collection of points in multidimensional space // European Journal of Operational Research, 1996. Vol. 92(2). Pp. 352–367.

[2] *Nemirko A.* Image recognition algorithms based on the representation of classes by convex hulls // Lecture Notes in Computer Science, 2021. Vol. 12665. Pp. 44–50.

[3] *Nemirko A., Dulá J.* Nearest convex hull classification based on linear programming // Pattern Recognition and Image Analysis, 2021. Vol. 31(2). Pp. 205–211.

Estimating the proximity of convex hulls for machine learning problems

Nemirko Anatoly apn-bs@yandex.ru
Saint Petersburg, Saint Petersburg Electrotechnical University "LETI"

Many machine learning problems (regression, classification, clustering, feature space dimensionality reduction) use the representation of classes as sets of points in a multidimensional feature space. Another way to describe classes is to use the convex hulls of the original sets. Despite the loss of statistical properties of the data, which is often insignificant, convex hulls define the boundaries of changes in experimental data, compactly store all the basic information about them, which allows you to quickly respond to a variety of queries on this set. Such a geometric description of classes reflects the orientation of classes in space, their relative position relative to each other and the configuration of their intersection areas. The main obstacle to the development of this direction is the computational complexity, which rapidly increases with an increase in the number of features and experimental points.

The paper considers machine learning methods that use the description of classes by convex hulls in a multidimensional feature space. Methods of approximate estimation of the localization of a point with respect to the boundary of a convex hull (inside, on the boundary or outside the convex hull) without calculating the convex hull itself are found [1,2]. This made it possible to approximate the distance from a point to the boundary of a convex hull [3] regardless of its localization and implement the nearest convex hull method. Efficient methods for calculating the set of extreme points are also proposed [1], which simplify class descriptions and the implementation of decision rules.

The recently proposed method of finding the distance from a point to the boundary of a convex hull [3] is based on the use of linear programming. Unlike NP-complete quadratic programming problems, this problem belongs to the P class of complexity and can be solved in weakly polynomial time. In practice, there are very effective algorithms for this. The above methods have good scalability and can be used in many machine learning tasks when describing classes in a multidimensional space. They have been successfully tested on the tasks of automatic medical diagnosis of breast cancer and the recognition of dangerous arrhythmias by ECG. The overall classification accuracy on the test sample for breast cancer was 97.9% (for linear SVM 92.8%) and for dangerous arrhythmias 91.5% (for linear SVM 93.0

In this paper, we propose new machine learning methods based on the above results and do not require the computation of convex hulls or triangulation of a set of points. The proposed procedure for localizing a point with respect to the boundary of a convex hull [1] simplifies the solution of the problem of visualizing points that fall into the intersection zone of convex hulls. This makes it possible to clarify the possibility of increasing the effectiveness of the decisive rule. In the work on model and real data, examples of mapping on the plane of points that fall

into the intersection zone of two convex hulls describing classes in multidimensional space are shown.

A simple method is also proposed for estimating the proximity of two convex hulls to each other, regardless of their intersection, which is based on the above results. We are looking for the Euclidean distance between the boundaries of two convex hulls, on a straight line connecting two centroids of the classes under consideration. It is calculated as the difference between the two distances from the centroids of the classes to the boundaries of the convex hulls of the opposite classes. These distances are determined using the linear programming method. This method is applied as a measure of distance when combining intermediate clusters based on the nearest convex hulls in cluster analysis based on convex hulls.

This research is funded by RFBR, grant 19-29-01009.

[1] *Dulá J., Helgason R.* A new procedure for identifying the frame of the convex hull of a finite collection of points in multidimensional space // European Journal of Operational Research, 1996. Vol. 92(2). Pp. 352–367.

[2] *Nemirko A.* Image recognition algorithms based on the representation of classes by convex hulls // Lecture Notes in Computer Science, 2021. Vol. 12665. Pp. 44–50.

[3] *Nemirko A., Dulá J.* Nearest convex hull classification based on linear programming // Pattern Recognition and Image Analysis, 2021. Vol. 31(2). Pp. 205–211.

Слабо-контролируемое обучение на основе матрицы нечетких отношений

Бериков Владимир Борисович[1,2]⋆ berikov@math.nsc.ru
Литвиненко Александр[3] litvinenko@uq.rwth-aachen.de
[1]Новосибирск, Институт математики им. С. Л. Соболева СО РАН
[2]Новосибирск, Новосибирский государственный университет
[3]Рейнско-Вестфальский технический университет Ахена, Германия

В работе решается задача анализа данных в постановке слабо-контролируемого обучения. В данной задаче для некоторых наблюдений $x_i \in \mathbf{R}^d$, $i = 1, \ldots, n$ известны точные значения прогнозируемой величины y_i, для некоторых – известны неточно из-за наличия случайного шума или других причин, таких как нехватка ресурсов для тщательной разметки, а для других – неизвестны. Подобные задачи возникают во многих областях исследования, в которых имеется большие объемы данных, получаемых, например, на основе измерений сенсоров, но разметка полученных наблюдений является ресурсозатратной процедурой (требует привлечения высококвалифицированных экспертов, специальных дополнительных исследований и т.п.). Часто доля точно размеченных прецедентов мала по сравнению с общим объемом выборки. В качестве примера можно привести задачу анализа томографических изображений: при обучении нейронной сети распознаванию зоны поражения мозга при инсульте требуется провести анализ множества цифровых КТ снимков, в которых специалист-рентгенолог размечает поврежденную область головного мозга. При большом числе снимков не всегда удается провести тщательную разметку. Во многих случаях рентгенологу удобнее отметить область поражения в виде рамки или овала, охватывающего интересующую зону. В других подобных задачах требуется проводить классификацию на уровне множеств прецедентов; при этом каждый элемент множества может быть размечен неточно.

Предложенный в работе [1] метод решения данной задачи требует оптимизации функции потерь, при выводе которой используется регуляризация многообразия на основе заданной модели неточности и представление матрицы сходства прецедентов в малоранговой форме. Такое представление позволяет значительно ускорить вычисления и снизить требования по памяти.

В случае задачи регрессионного анализа ($y_i \in \mathbf{R}$), модель неточности может быть определена с помощью нормального распределения: $\mathcal{N}(a_i, \sigma_i)$, где a_i, σ_i - параметры, характеризующие ожидаемое значение и разброс возможных значений y_i. Способ задания параметров определяется конкретной задачей; например, при анализе изображений можно предположить, что чем ближе i-й пиксель находится к центру участка, ограниченного контуром разметки, тем она более точна (т.е. меньше величина σ_i).

Для нахождения матрицы сходства в малоранговой форме используется матрица отношений, полученная на основе кластерного ансамбля. Матрица опре-

деляет, к какому кластеру был отнесен тот или иной прецедент базовыми алгоритмами ансамбля. Использование кластерного анализа в случае задачи слабоконтролируемого обучения позволяет извлечь дополнительную информацию из неразмеченных и неточно размеченных прецедентов [2].

В докладе обсуждается предлагаемая модификация данного метода, основанная на модели ансамбля нечетких алгоритмов кластеризации. Нечеткие методы (в данной работе используется метод Fuzzy C-Means, FCM) обладают тем преимуществом, что на границах кластеров дают более "адекватные" результаты группирования. Ансамбль применяется, чтобы повысить устойчивость решений, а также оценить и уменьшить влияние случайности в выборе инициализаций на итоговое нечеткое разбиение [3]. Матрица нечетких отношений используется для нахождения матрицы сходства прецедентов, которая также представляется в малоранговом виде. Алгоритм FCM имеет линейную трудоемкость по отношению к размерности входных данных, поэтому может применяться для достаточно больших объемов выборки.

Предложенный метод исследуется на искусственных и реальных наборах данных с использованием статистического моделирования.

Работа поддержана грантом РФФИ No. 19-29-01175.

[1] *Berikov V., Litvinenko A.* Weakly supervised regression using manifold regularization and low-rank matrix representation // Lecture Notes in Computer Science, 2021. Vol. 12755. Pp. 447–461.

[2] *Berikov V., Litvinenko A.* Semi-supervised regression using cluster ensemble and low-rank co-association matrix decomposition under uncertainties // Conf. Proceedings of 3rd International Conference on Uncertainty Quantification in Computational Sciences and Engineering, 2019. Pp 229–242.

[3] *Berikov V.* A Probabilistic Model of Fuzzy Clustering Ensemble // Pattern Recognition and Image Analysis, 2018. Vol. 28(1). Pp 1–10.

Weakly supervised learning based on a fuzzy relationship matrix

Berikov Vladimir[1,2]⋆ berikov@math.nsc.ru
Litvinenko Alexander[3] litvinenko@uq.rwth-aachen.de
[1]Novosibirsk, Sobolev Institute of mathematics SB RAS
[2]Novosibirsk, Novosibirsk State University
[3]RWTH Aachen University, Germany

The paper considers the problem of data analysis in the formulation of weakly supervised learning. In this problem, for some observations $x_i \in \mathbf{R}^d$, $i = 1, \ldots, n$, the exact values of the predicted value y_i are known; for some they are known inaccurately due to the presence of random noise or other reasons, such as a lack of resources for scrupulous labeling, while others are unknown. Similar problems arise in many areas of research, in which there is a large amount of data obtained, for example, on the basis of sensor measurements. Labeling the observations is a resource-intensive procedure (requiring highly qualified experts, additional special research, etc.). Often, the proportion of accurately labeled precedents is small compared to the total sample size. An example is the problem of analyzing tomography images. There by training a neural network to recognize a brain lesion zone affected by stroke, it is required to analyze a set of digital CT scans, in which a radiologist annotates the damaged area. With a large number of images, it is not always possible to carry out thorough labeling. It is often more convenient for the radiologist to annotate the affected area in the form of a frame or oval covering the area of interest. Other similar tasks require classification at the level of sets of precedents; moreover, each element of a set can be labeled inaccurately.

In the method proposed in [1] to the solution of the considered problem, it is required to optimize the loss function, the derivation of which is based on manifold regularization for the given model of inaccuracy and on the representation of the similarity matrix in the low-rank form. This representation can significantly speed up computations and reduce memory requirements.

In the case of a regression problem ($y_i \in \mathbf{R}$), the imprecision model can be determined using normal distribution: $\mathcal{N}(a_i, \sigma_i)$, where a_i, σ_i are parameters characterizing the expected value and the scatter of possible values of y_i. The way of setting parameters is determined by a specific task; for example, when analyzing images, it can be assumed that the closer the ith pixel is to the center of the area bounded by the labeling contour, the more accurate the labeling is (i.e., the smaller the value of σ_i).

To find the similarity matrix in the low-rank form, a matrix of relations obtained on the basis of a cluster ensemble is used. This matrix determines to which cluster a particular precedent is assigned by the base algorithms of the ensemble. The usage of cluster analysis in the case of weakly supervised learning allows one to extract additional information from unlabeled and inaccurately labeled precedents [2].

This work discusses the proposed modification of this method, based on the model of an ensemble of fuzzy clustering algorithms. Fuzzy methods (in this work, the Fuzzy C-Means, FCM is applied) have the advantage they give more "adequate" grouping results at the cluster boundaries. The ensemble is utilized to increase the stability of solutions, as well as to estimate and reduce the influence of randomness in the initializations on the resulting fuzzy partition [3]. A fuzzy relationship matrix is used to find the similarity matrix for the precedents, which is also presented in a low-rank form. FCM algorithm has a linear complexity regarding the dimension of input data. Therefore it can be efficiently implemented for sufficiently large sample sizes.

The proposed method is investigated on artificial and real data sets using Monte-Carlo modeling.

This research is funded by RFBR, grant 19-29-01175.

[1] *Berikov V., Litvinenko A.* Weakly supervised regression using manifold regularization and low-rank matrix representation // Lecture Notes in Computer Science, 2021. Vol. 12755. Pp. 447–461.

[2] *Berikov V., Litvinenko A.* Semi-supervised regression using cluster ensemble and low-rank co-association matrix decomposition under uncertainties // Conf. Proceedings of 3rd International Conference on Uncertainty Quantification in Computational Sciences and Engineering, 2019. Pp. 229–242.

[3] *Berikov V.* A Probabilistic Model of Fuzzy Clustering Ensemble // Pattern Recognition and Image Analysis, 2018. Vol. 28(1). Pp 1–10.

Поиск неприводимых пороговых ассоциативных правил в частично упорядоченных 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

Рассматривается задача поиска ассоциативных правил специального вида в данных с элементами из декартова произведения конечных частично упорядоченных множеств. Для сокращения временных затрат при анализе небинарных данных используются модификация классического бинарного FP-дерева и параллельные вычисления на основе технологии CUDA. Приводятся результаты тестирования последовательного и параллельного алгоритмов.

Поиск ассоциативных правил является одной из центральных задач интеллектуального анализа информации, имеет важное прикладное значение и обычно осуществляется на основе нахождения по имеющейся базе транзакций часто встречающихся элементов (частых событий). Ассоциативное правило устанавливает зависимость между двумя частыми событиями, согласно которой одно частое событие X с некоторой «достоверностью» влечёт другое частое событие Y. При этом события X и Y порождаются одним общим частым событием, обозначаемым далее $X \odot Y$. Наиболее информативными считаются ассоциативные правила, порождаемые «максимальными» частыми элементами $X \odot Y$ с «минимальной» посылкой X. Такие правила называются неприводимыми (Elbassioni K. M., 2014), и задача их нахождения особенно важна в случае больших данных. Вопросы поиска ассоциативных правил наиболее изучены в случае неупорядоченных бинарных данных (Agrawal R., Imielinski T., Swami A., 1993).

Один из наиболее известных способов нахождения частых элементов в бинарных данных основан на представлении информации, содержащейся в базе транзакций, в виде древовидной структуры, называемой Frequent Pattern Tree (FP-деревом). В более общем случае, как правило, осуществляется бинаризация исходных данных по некоторому числовому набору порогов, и задача сводится к построению бинарного FP-дерева. Частые элементы и ассоциативные правила, найденные по бинаризированным данным, называются пороговыми. Результат существенно зависит от выбора набора порогов. Однако перебор по всем возможным вариантам бинаризации требует существенных временных затрат. В [1] для сокращения времени анализа небинарных данных, в том числе с элементами из декартова произведения конечных частично упорядоченных множеств, предложена модель порогового FP-дерева, названная TFP-деревом.

В настоящей работе поставлена задача поиска неприводимых пороговых ассоциативных правил в данных, представленных в виде декартова произведения $P = P_1 \times \ldots \times P_n$, где P_i, $i \in \{1, .., n\}$, — конечное частично упорядоченное множество, называемое далее атрибутом. Для решения поставленной задачи разра-

ботаны и исследованы два алгоритма: последовательный алгоритм TFP-tree и его параллельная версия на основе технологии CUDA (алгоритм DPTFP-tree).

Таблица 1. Поиск неприводимых правил алгоритмом TFP-tree.

$m \times n * k$	$\lvert H_D \rvert$	$\lvert F_D \rvert$	$\lvert A_D \rvert$	t, с	t_1, с	$\triangle t$, %
$300 \times 40 * 4$	1512	27866	633	0.5	0.4	75
$300 \times 40 * 5$	10584	229951	12779	9.4	6.7	59.7
$300 \times 40 * 6$	86436	2217292	259420	460.1	378.6	78.5
$300 \times 40 * 7$	691488	22897938	3982077	2710.2	2358.7	85.1
$3000 \times 40 * 4$	1764	18795	752	4.2	3.4	76.5
$30000 \times 40 * 4$	2401	27875	842	85.4	64.2	67
$300000 \times 40 * 4$	1764	29288	698	1018.3	631.6	38.8
$900000 \times 40 * 4$	1764	28268	625	2910.9	1813.6	39.5
$1800000 \times 40 * 4$	1764	38275	540	7481.5	4815.7	44.6

Каждый шаг алгоритма TFP-tree состоит из двух этапов. На первом этапе на основе анализа базы транзакций D, в которой каждая транзакция является некоторым элементом из P, строится множество так называемых значимых наборов порогов H_D и для каждого набора из H_D ищется множество максимальных пороговых частых элементов F_D. На втором этапе для каждого элемента из F_D находится множество неприводимых пороговых ассоциативных правил A_D. В алгоритме DPTFP-tree множество H_D разбивается на непересекающиеся подмножества, каждое из которых подаётся на отдельный вычислительный блок графического процессора для поиска искомых правил.

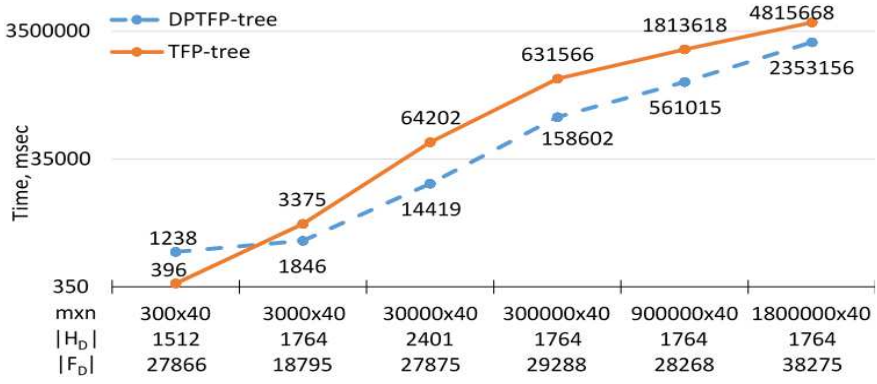


Рис. 1. Зависимость времени синтеза F_D от m.

В таблице 1 представлены результаты работы алгоритма TFP-tree на случайных модельных данных, при условии, что каждый атрибут P_i является цепью и мощность P_i не превосходит 10. Для тройки (m, n, k), где m — число транзак-

ций, k — число небинарных атрибутов, указаны мощности множеств H_D, F_D и A_D, общее время работы алгоритма t, время его работы на первом этапе t_1 и величина $\triangle t = t_1/t \cdot 100\%$. Нетрудно видеть, что время первого этапа в большинстве случаев составляет более половины от общего времени работы алгоритма и при $|H_D| > 600000$ достигает 85%. При увеличении m время, затрачиваемое алгоритмом на первом этапе, существенно уменьшается.

На рис. 1 показана зависимость времени синтеза множества F_D алгоритмами TFP-tree и DPTFP-tree от числа транзакций m (при $n = 40$, $k = 4$). Видно, что при $m > 3000$ алгоритм DPTFP-tree работает быстрее алгоритма TFP-tree в среднем в три раза.

Таким образом, показана эффективность применения предлагаемого подхода к синтезу неприводимых ассоциативных правил в произведении частичных порядков.

Работа частично финансирована грантом РФФИ No. 19-01-00430-а.

[1] *Genrikhov I., Djukova E.* Finding frequent elements for a product of partial orders and association rules // Int. conf. ITNT-2020, 2020. Pp. 1–5.

Finding irredundant threshold association rules in partially ordered data

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

The problem of finding association rules of some special type in data with elements from the Cartesian product of finite partially ordered sets is considered. To reduce the time spent in the analysis of nonbinary data, a modification of the classical binary FP-tree and parallel calculations based on CUDA technology are used. The results of testing the sequential and parallel algorithms are presented.

Finding association rules in data is one of the central problem in intelligent data analysis, has an important application value and is usually carried out on the basis of finding frequently occurring elements (frequent events) in the available transaction database. The association rule establishes a relationship between two frequent events, according to which one frequent event X with some "reliability" entails another frequent event Y. In this case, the elements X and Y are generated by one common frequent element, denoted further $X \odot Y$. The most informative are those association rules that are generated by the "maximum" frequent elements $X \odot Y$ with the "minimum" precondition X. Such rules are called irredundant (Elbassioni K. M., 2014), and the task of finding them is especially important in the case of big data. The issues of searching for association rules are most studied in the case of unordered binary data (Agrawal R., Imielinski T., Swami A., 1993).

One of the most well-known ways to find frequent elements in binary data is based on the representation of information contained in the transaction database in a compact tree structure called a Frequent Pattern Tree (FP-tree). In the case of nonbinary data, as a rule, the original data is binarized by means of a certain numerical set of thresholds, and the task is reduced to the construction of a binary FP-tree. Frequent elements and association rules found from binarized data are called thresholds. The result depends significantly on the choice of a set of thresholds. However, the search through all data binarization variants is computationally costly. In [1], to reduce the analysis time of binary data, including with elements from the Cartesian product of finite partially ordered sets, a threshold FP-tree model called a TFP-tree is proposed.

In this work, the problem of finding irredundant threshold association rules in data presented as Cartesian product is set $P = P_1 \times \ldots \times P_n$, where P_i, $i \in \{1, .., n\}$, — is a finite partially ordered set, next referred to as an attribute. Two algorithms have been developed and studied to solve this problem: the sequential TFP-tree algorithm and its parallel version based on the CUDA technology (DPTFP-tree algorithm).

Table 1. Search for irredundant rules by the TFP-tree algorithm.

$m \times n * k$	$\lvert H_D\rvert$	$\lvert F_D\rvert$	$\lvert A_D\rvert$	t, sec	t_1, sec	$\triangle t$, %
$300 \times 40 * 4$	1512	27866	633	0.5	0.4	75
$300 \times 40 * 5$	10584	229951	12779	9.4	6.7	59.7
$300 \times 40 * 6$	86436	2217292	259420	460.1	378.6	78.5
$300 \times 40 * 7$	691488	22897938	3982077	2710.2	2358.7	85.1
$3000 \times 40 * 4$	1764	18795	752	4.2	3.4	76.5
$30000 \times 40 * 4$	2401	27875	842	85.4	64.2	67
$300000 \times 40 * 4$	1764	29288	698	1018.3	631.6	38.8
$900000 \times 40 * 4$	1764	28268	625	2910.9	1813.6	39.5
$1800000 \times 40 * 4$	1764	38275	540	7481.5	4815.7	44.6

Each step of the TFP-tree algorithm consists of two stages. At the first stage, based on the analysis of the transaction database D, in which each transaction is some element of P, the set of so-called significant sets of thresholds H_D are constructed and for each set from H_D, the set of maximum threshold frequent elements F_D is searched for. At the second stage for each element from F_D, the set of irredundant threshold association rules A_D is searched for. In the parallel DPTFP-tree algorithm, the set H_D is divided into disjoint subsets, each of which is sent to a separate computing block of the graphics processor to search for the desired rules.

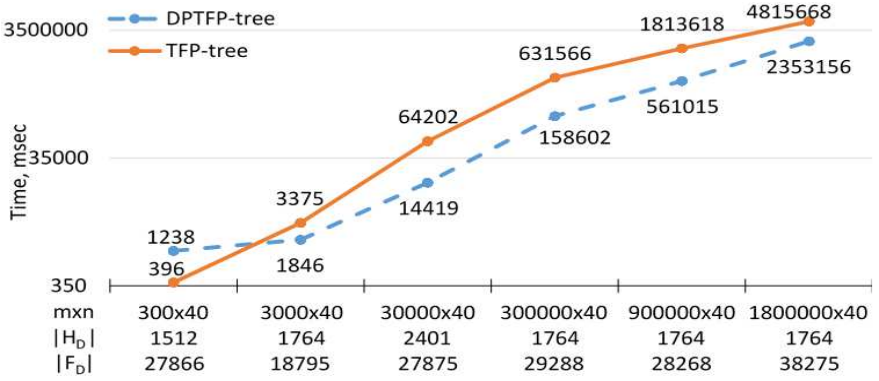


Fig. 1. The dependence of the synthesis time of the F_D on m.

Table 1 shows the results of the TFP-tree algorithm on random model data, provided that each attribute P_i is a chain and the power P_i does not exceed 10. For the triple (m, n, k), where m — number of transactions, k — number of nonbinary attributes, the powers of the sets H_D, F_D and A_D, the total running time of the algorithm t, the time of its operation at the first stage t_1 and the value $\triangle t = t_1/t \cdot$ $\cdot\, 100\%$ are indicated. It is easy to see that the time of the first stage in most cases is more than half of the total time of the algorithm and with a large $|H_D| > 600000$

is reaches to 85%. With an increase m, the time spent by the algorithm at the first stage of operation is significantly reduced.

Figure 1 shows the dependence of the synthesis time of the set F_D by the TFP-tree and DPTFP-tree algorithms on the number of transactions m (by $n = 40$, $k = 4$). It can be seen that for $m > 3000$ the DPTFP-tree algorithm is faster than TFP-tree algorithm by an average of three times.

Thus, the effectiveness of the proposed approach to the synthesis of irredundant association rules in a product of partial orders is shown.

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

[1] *Genrikhov I., Djukova E.* Finding frequent elements for a product of partial orders and association rules // Int. conf. ITNT-2020, 2020. Pp. 1–5.

Машинное обучение на логических высказываниях: меры сходства, нетривиальности и кластеризация логических формул

Викентьев Александр Александрович[1,2]⋆ vikent@math.nsc.ru
Бериков Владимир Борисович[1,2] berikov@math.nsc.ru
[1]Новосибирск, Институт математики им. С. Л. Соболева СО РАН
[2]Новосибирск, Новосибирский государственный университет

В докладе рассматривается одна из актуальных задач на стыке математической логики и машинного обучения – анализ логических высказываний, полученных из базы знаний информационной системы или от экспертов какой-либо предметной области. При анализе требуется найти близкие высказывания, выявить достоверные, найти нетривиальные утверждения и т.д. Разбиение множества высказываний на подмножества похожих элементов (кластеризация) позволяет проводить структурирование базы знаний и облегчает поиск высказываний, наиболее релевантных запросу. Для кластеризации знаний, построения решающих функций на основе формул-высказываний, требуется ввести расстояние между формулами. В данной работе высказывания записаны в виде формул n–значной логики, что позволяет учитывать их возможную неоднозначность. С привлечением многозначного класса моделей (переменная может входить в модель с различными значениями истинности) определяются расстояния между формулами: $\rho(\varphi,\psi) = \frac{1}{n^{|S(\Sigma)|}} \sum_{k=0}^{n-1} \sum_{l=0}^{n-1} \frac{|k-l|}{n-1} M\left(\frac{k}{n-1}, \frac{l}{n-1}\right)$, где $n^{|S(\Sigma)|}$ – число всех многозначных моделей, $M\left(\frac{k}{n-1}, \frac{l}{n-1}\right)$ – число тех моделей, на которых формула φ принимает значение $\frac{k}{n-1}$, а ψ – значение $\frac{l}{n-1}$.

Введены также новые меры нетривиальности, обобщающие предложенные ранее в работах [1, 2, 3] и имеющие вид: $I(\varphi) = \rho(\varphi, \mathbf{1})$, где $\mathbf{1}$ – тождественно истинная формула. Доказано, что введенные семейства мер обладают свойствами метрики. Предложены различные методы кластеризации логических знаний и методы сравнения результатов на основе индексов качества кластеризации. Рассмотрены конкретные примеры анализа высказываний для логик различной значности. Полученные результаты обобщены на случай коллективной кластеризации [4] на основе ансамбля метрик.

Все полученные выше результаты, использующие многозначные модели, справедливы и для других многозначных логик, отличающихся от логики Лукасевича определением истинностных значений логических связок. Для них будут возникать другие расстояния (метрики с неравенством треугольника), а значит и другие способы кластеризации. Применять такие логики целесообразно в случае, когда большинство экспертов дает предпочтение в пользу конкретной логики из возможного списка.

Предложенный подход дает возможность проводить машинное обучение на экспертных знаниях, а также выполнять их коллективную кластеризацию.

Работа выполнена в рамках государственного задания ИМ СО РАН (проект No.0314-2019-0015).

[1] *Vikent'ev A., Lbov G.* Setting the metric and informativeness on statements of experts // Pattern Recognition and Image Analysis, 1997. Vol. 7(2). Pp. 175–183.

[2] *Vikent'ev A., Avilov M.* New Model Distances and Uncertainty Measures for Multivalued Logic // Lecture Notes on Computer Science, 2016. Vol. 9883. Pp. 89–98.

[3] *Vikent'ev A., Serov M., Vikentiev R., Berikov V.* Collective Distances for Clustering N-Valued Logic Formulas Representing Knowledge Base of Intellectual System // 2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON), 2019. Pp. 664-669.

[4] *Berikov V.* Weighted ensemble of algorithms for complex data clustering // Pattern Recognition Letters, 2014. Vol. 38. Pp. 99–106.

Machine learning on logical statements: measures of similarity, non-trivialities and clustering of logical formulas

Vikentiev Alexander[1,2]⋆ vikent@math.nsc.ru
Berikov Vladimir[1,2] berikov@math.nsc.ru
[2]Novosibirsk, Sobolev Institute of Mathematics SB RAS
[2]Novosibirsk, Novosibirsk State University

The work considers one of the topical problems at the intersection of mathematical logic and machine learning, i.e., the analysis of logical statements, obtained from the knowledge base of an information system or from experts from an applied area of research. For the analysis, one needs to find out similar statements, identify valid propositions, find non-trivial predicates, etc. The partitioning of the set of assertions into subsets of similar elements (clustering) allows one to structure a knowledge base and facilitates the search for the statements that are most relevant to a request. For the clustering of knowledge and finding decision functions on the basis of the formulas-statements, it is required to determine a distance between formulas. In this paper, propositions are written in the form of formulas of n-valued logic, which allows one to take into account their possible ambiguity. Using a multi-valued class of models (a variable enters the model with different truth values), the distances between the formulas are determined as follows: $\rho(\varphi,\psi) = \frac{1}{n^{|S(\Sigma)|}}\sum_{k=0}^{n-1}\sum_{l=0}^{n-1}\frac{|k-l|}{n-1}M\left(\frac{k}{n-1},\frac{l}{n-1}\right)$, where $n^{|S(\Sigma)|}$ is the number of all multi-valued models, $M\left(\frac{k}{n-1},\frac{l}{n-1}\right)$ is the number of those models on which the formula φ takes the value $\frac{k}{n-1}$, and ψ takes the value $\frac{l}{n-1}$.

New measures of non-triviality have also been introduced, generalizing those proposed earlier in the works [1, 2, 3] and having the form: $I(\varphi) = \rho(\varphi, \mathbf{1})$, where $\mathbf{1}$ is the identity true formula. It was proved that the families of measures introduced have properties of the metric. Various methods of logical knowledge clustering are proposed, and methods for the comparing the results based on cluster validity indices are introduced. Specific examples of the analysis of the statements for logics of different values are given. The results obtained are generalized to the case of collective clustering [4] based on ensemble metrics.

All the results obtained above using multi-valued models are also valid for other multi-valued logics, which differ from Lukasiewicz logic in the definition of the truth values of logical connectives. For these logics there will be other distances (metrics with triangle inequality), and hence other ways of clustering. It is advisable to use such logic in the case when the majority of experts give preference in favor of a specific logic from the possible list.

The proposed approach makes it possible to carry out machine learning based on expert knowledge, as well as to perform collective clustering of formulas.

The study was carried out within the framework of the state contract of the Sobolev Institute of Mathematics (project no 0314-2019-0015).

[1] *Vikent'ev A., Lbov G.* Setting the metric and informativeness on statements of experts // Pattern Recognition and Image Analysis, 1997. Vol. 7(2). Pp. 175–183.

[2] *Vikent'ev A., Avilov M.* New Model Distances and Uncertainty Measures for Multivalued Logic // Lecture Notes on Computer Science, 2016. Vol. 9883. Pp. 89–98.

[3] *Vikent'ev A., Serov M., Vikentiev R., Berikov V.* Collective Distances for Clustering N-Valued Logic Formulas Representing Knowledge Base of Intellectual System // 2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON), 2019. Pp. 664-669.

[4] *Berikov V.* Weighted ensemble of algorithms for complex data clustering // Pattern Recognition Letters, 2014. Vol. 38. Pp. 99–106.

Энтропийное моделирование сетевых структур

Тырсин Александр Николаевич[1,2] author_at2001@yandex.ru

[1]Екатеринбург, Уральский федеральный университет

[2]Челябинск, Южно-Уральский государственный университет

Одной аз актуальных проблем является создание адекватного инструментария для исследования и мониторинга состояния сетевых структур. Сетевыми структурами назовем системы, каждый из элементов которой связан хотя бы с одним из других элементов системы. Они могут быть представлены в виде связных графов, в которых связь между элементами (вершинами) задается в виде тесноты корреляционной взаимосвязи.

В настоящее время достаточно распространено использование энтропии для описания сложных систем в различных областях. Доклад посвящен проблематике использования дифференциальной энтропии (далее энтропии) для сетевых структур. Представим сетевую структуру в виде непрерывного случайного вектора $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_m)$. Известно, что энтропию $H(\mathbf{Y})$ непрерывного случайного вектора $\mathbf{Y}$ может быть разложена на две составляющих – энтропию хаотичности и энтропию самоорганизации.

Для сетевых структур, наряду с оценкой самой энтропии, будут полезны и другие энтропийные характеристики, такие как энтропия взаимосвязи между несколькими подсистемами и энтропия системы в отдельной вершине.

Пусть заданы несколько подсистем $\mathbf{Y}^k$ системы $\mathbf{Y}$, таких что $\mathbf{Y}^k = (Y_{k,1}, \ldots, Y_{k,m_k}) \subset \mathbf{Y}$, $k = 1, 2, \ldots, K$, $K \in 2, 3, \ldots, m$, любая компонента Y_j может входить в состав не более чем одной подсистемы (случайного вектора) $\mathbf{Y}^k$. Определим энтропию взаимосвязи между подсистемами (случайными векторами) $\mathbf{Y}^1, \ldots, \mathbf{Y}^K$ как

$$H(\bigcap_{k=1}^{K} \mathbf{Y}^k) = \sum_{k=1}^{K} H(\mathbf{Y}^k) - H(\bigcup_{k=1}^{K} \mathbf{Y}^k).$$

Пусть сетевая структура представлена в виде случайного вектора $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_m)$. Определим энтропию в вершине Y_l как

$$H(Y_l(\mathbf{Y})) = H(\mathbf{Y}) - H(\mathbf{Y} \backslash Y_l).$$

Энтропия взаимосвязи между несколькими подсистемами и энтропия системы в отдельной вершине позволят исследовать сетевые структуры: оценивать взаимосвязанность меду собой различных участков, а также оценивать как меняется энтропия внутри таких систем. В докладе рассмотрены вопросы описания данных величин, а также примеры их применения на модельных и реальных данных.

Работа поддержана грантом РФФИ No. 20-51-00001.

[1] *Тырсин А. Н.* Энтропийное моделирование многомерных стохастических систем // Воронеж: Наука, 2016. 156. с.

Entropy modeling of network structures

Tyrsin Alexander[1,2] author_at2001@yandex.ru

[1]Yekaterinburg, Ural Federal University

[2]Chelyabinsk, South Ural State University

The creation of adequate tools for the study and monitoring of the state of network structures is one of the urgent problems. Network structures are systems, each of the elements of which is connected to at least one of the other elements of the system. They can be represented in the form of connected graphs, in which the relationship between the elements (vertices) is given in the form of the magnitude of the correlation relationship.

Currently, it is quite common to use entropy to describe complex systems in various fields. The report is devoted to the problems of using differential entropy (hereinafter entropy) for network structures. Let's imagine the network structure as a continuous random vector $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_m)$. It is known that the entropy of $H(\mathbf{Y})$ a continuous random vector $\mathbf{Y}$ can be decomposed into two components – the entropy of randomness and the entropy of self-organization.

For network structures, along with the assessment of entropy itself, other entropy characteristics will be useful, such as the entropy of the relationship between several subsystems and the entropy of the system in a separate vertex.

Let there be several subsystems $\mathbf{Y}^k$ of the system $\mathbf{Y}$, where $\mathbf{Y}^k = (Y_{k,1}, \ldots, Y_{k,m_k}) \subset \mathbf{Y}$, $k = 1, 2, \ldots, K$, $K \in 2, 3, \ldots, m$, such that Y_j can be part of no more than one subsystem (random vector) $\mathbf{Y}^k$. We define the entropy of the relationship between subsystems (random vectors) $\mathbf{Y}^1, \ldots, \mathbf{Y}^K$ as

$$H(\bigcap_{k=1}^{K} \mathbf{Y}^k) = \sum_{k=1}^{K} H(\mathbf{Y}^k) - H(\bigcup_{k=1}^{K} \mathbf{Y}^k).$$

Let the network structure be represented as a random vector $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_m)$. Let's define the entropy at the vertex Y_l how

$$H(Y_l(\mathbf{Y})) = H(\mathbf{Y}) - H(\mathbf{Y} \backslash Y_l).$$

The entropy of the relationship between several subsystems and the entropy of the system in a separate vertex will allow us to investigate network structures: to assess the interconnectedness of different sections between each other, as well as to assess how entropy changes within such systems. The report discusses the description of these quantities, as well as examples of their application on model and real data.

This research is funded by RFBR, grant 20-51-00001.

[1] *Tyrsin A. N.* Entropy modeling of multidimensional stochastic systems // Voronezh: Scientific Book, 2016. 156 p.

Асимптотически оптимальная расшифровка монотонной логической функции

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

Авторами предложен и экспериментально исследован новый подход к решению задачи расшифровки двузначной монотонной функции, определённой на элементах декартова произведения частичных порядков $\mathcal{P}$. Предложенный подход базируется на решении задачи дуализации над произведением частичных порядков, является эффективным в типичном случае и принципиально отличается от известных традиционных подходов, ориентированных на самый сложный вариант задачи.

Расшифровка двузначной монотонной функции является одной из центральных задач дискретной математики и ставится следующим образом. Пусть $\mathcal{P} = \mathcal{P}_1 \times \mathcal{P}_2 \times \ldots \times \mathcal{P}_n$ — декартово произведение конечных частично упорядоченных множеств. Элемент $p = (p_1, \ldots, p_n) \in \mathcal{P}$ предшествует элементу $q = (q_1, \ldots, q_n) \in \mathcal{P}$, если $p_1 \preceq q_1$ в $\mathcal{P}_1, \ldots, p_n \preceq q_n$ в $\mathcal{P}_n$ (q следует за p). На множестве $\mathcal{P}$ определена двузначная монотонная функция F. Функция F задана при помощи некоторого оператора B, который для любого $x \in \mathcal{P}$ возвращает значение $F(x)$. Требуется путем обращения к оператору B найти все нули и единицы функции F.

Для задачи расшифровки центральными являются понятия верхнего нуля и нижней единицы двузначной монотонной функции. Если ноль функции F таков, что любой следующий за ним элемент множества $\mathcal{P}$ является единицей, то такой ноль называется верхним нулем функции F. Если единица функции F такова, что любой предшествующий ей элемент множества $\mathcal{P}$ является нулем, то такая единица называется нижней единицей функции F. Так как множество всех верхних нулей функции F определяет все нули F, а множество нижних единиц — все единицы, то для расшифровки функции F достаточно построить множества всех ее верхних нулей и всех ее нижних единиц.

Покажем, что к задаче поиска верхних нулей и нижних единиц двузначной монотонной функции сводится поиск по базе транзакций максимальных частых и минимальных нечастых элементов декартова произведения частичных порядков $\mathcal{P}$.

Действительно, пусть D — некоторая совокупность не обязательно различных элементов из $\mathcal{P}$; $s \in [0, 1]$. Элементы множества $\mathcal{P}$, содержащиеся в D, называются транзакциями. Элемент $x \in \mathcal{P}$ называется s-частым, если $\nu(x)$ — доля транзакций в D, следующих за x, не менее s. Иначе x — s-нечастый. Если элемент частый и за ним не следует никакой другой частый элемент, то он называется максимальным частым. Если элемент нечастый и при этом ему не предшествует никакой другой нечастый элемент, то такой элемент называется

минимальным нечастым. Требуется найти множества X_{max} и Y_{min}, состоящие соответственно из всех максимальных частых и минимальных нечастых элементов множества $\mathcal{P}$.

Рассмотрим функцию F_D, определенную на множестве $\mathcal{P}$, принимающую значение 0 и 1 соответственно на частых и нечастых элементах этого множества и заданную при помощи оператора $B(F_D)$, который для любого элемента x из $\mathcal{P}$ возвращает значение $F_D(x)$ путём вычисления частоты встречаемости x в D. Нетрудно видеть, что поиск X_{max} и Y_{min} эквивалентен поиску верхних нулей и нижних единиц функции F_D.

Традиционный подход к задаче расшифровки основан на построении оптимального по Шеннону алгоритма (предложен В.К. Коробковым в 1965 г.). Согласно этому подходу, сложность алгоритма расшифровки F следует оценивать числом обращений к оператору B в «худшем случае». Задача построения оптимального алгоритма расшифровки монотонной булевой функции решена Ж. Анселем в 1968 г. В 1976 году В. Б. Алексеевым построен алгоритм расшифровки, который является оптимальным в случае, когда F — булева функция, и близок по сложности к оптимальному в более общем случае, а именно, когда F определена на декартовом произведении конечных цепей [1].

В настоящей работе представлены результаты сравнительного анализа двух методов расшифровки функции F_D для случая, когда $\mathcal{P}$ — декартово произведение конечных цепей. Первый использует предложенную в [2] идею последовательно-совместного перечисления X_{max} и Y_{min} и опирается на решение задачи дуализации над произведением частичных порядков. Второй — упомянутый выше алгоритм В. Б. Алексеева. На случайных данных для каждого тестируемого метода оценивается время работы и число обращений к оператору $B(F_D)$. Экспериментальное исследование проведено для случая, когда $\mathcal{P}$ — декартово произведение конечных цепей. Результаты исследования свидетельствуют о том, что последовательно-совместное перечисление наиболее эффективно при большом числе цепей n в декартовом произведении и при высокой значности каждой отдельной цепи $\mathcal{P}_i$. В последовательно-совместной расшифровке используется асимптотически оптимальный алгоритм дуализации над произведением цепей (предложен Е. В. Дюковой, Г. О. Масляковым и П. А. Прокофьевым в 2017 году). Асимптотически оптимальные алгоритмы дуализации имеют теоретическое обоснование эффективности в типичном случае и на сегодняшний день являются лидерами по скорости счета.

Таким образом, исследованы актуальные вопросы снижения временных затрат, возникающие при логическом анализе данных с элементами из декартова произведения конечных частично упорядоченных множеств. Для задачи расшифровки двузначной монотонной функции, принимающей значение 0 и 1 соответственно на частых и нечастых элементах декартова произведения конечных цепей, предложен оригинальный алгоритм, выявлены условия его эффектив-

ности и показана целесообразность применения асимптотически оптимальных алгоритмов дуализации.

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

[1] *Алексеев В. Б.* О расшифровке некоторых классов монотонных многозначных функций // Журн. вычисл. матем. и матем. физики, 1976. Т. 16(1). С. 189–198.

[2] *Драгунов Н. А., Дюкова Е. В.* Поиск минимальных нечастых и максимальных частых наборов в частично упорядоченных данных // Математические методы распознавания образов: Тезисы докладов 19-й Всероссийской конференции с международным участием, 2019. С. 10–12.

Asymptotically optimal decoding of a monotone logical function

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

The authors propose and experimentally investigate a new approach to solving the problem of decoding a two-valued monotone function defined on elements of a Cartesian product of partial orders $\mathcal{P}$. The proposed approach is based on solving the dualization problem over the product of partial orders, is effective in the typical case and fundamentally differs from the well-known traditional approaches focused on the most complex version of the problem.

Decoding a two-valued monotone function is one of the central problems of discrete mathematics and is posed as follows. Let $\mathcal{P} = \mathcal{P}_1 \times \mathcal{P}_2 \times \ldots \times \mathcal{P}_n$ is a Cartesian product of finite partially ordered sets. Element $p = (p_1, \ldots, p_n) \in \mathcal{P}$ precedes the element $q = (q_1, \ldots, q_n) \in \mathcal{P}$ if $p_1 \preceq q_1$ in $\mathcal{P}_1, \ldots, p_n \preceq q_n$ in $\mathcal{P}_n$ (q follows p). A two-valued monotone function F is defined on the set $\mathcal{P}$. The function F is defined via some operator B, which for any $x \in \mathcal{P}$ returns the value $F(x)$. It is required to find all zeros and units of the function F by calling to the operator B.

For the decoding task, the concepts of upper zero and lower unit of a two-valued monotone function are central. If a zero of the function F is such that any element of the set $\mathcal{P}$ following it is a unit, then such a zero is called the upper zero of the function F. If a unit of the function F is such that any element of the set $\mathcal{P}$ preceding it is a zero, then such a unit is called the lower unit of the function F. Since the set of all upper zeros of the function F defines all zeros of F, and the set of lower units — all units, constructing the sets of all its upper zeros and all its lower units is enough to decode the function F.

We show that finding through the transaction database maximal frequent and minimal infrequent elements of a Cartesian product of partial orders $\mathcal{P}$ is reduced to the problem of finding the upper zeros and lower units of a two-valued monotone function.

Indeed, let D be some collection of not necessarily distinct elements from $\mathcal{P}$; $s \in [0, 1]$. The elements of the set $\mathcal{P}$ contained in D are called transactions. The element $x \in \mathcal{P}$ is called s-frequent if $\nu(x)$ — the ratio of transactions in D following x is at least s. Otherwise, x is s-infrequent. If an element is frequent and it is not followed by any other frequent element, then it is called the maximal frequent. If an element is infrequent and at the same time it is not preceded by any other infrequent element, then such an element is called minimal infrequent. It is required to find the sets X_{max} and Y_{min}, consisting respectively of all the maximal frequent and minimal infrequent elements of the set $\mathcal{P}$.

Consider the function F_D over the set $\mathcal{P}$, taking values 0 and 1, respectively, on frequent and infrequent elements of this set and defined via the operator $B(F_D)$, which for any element x of $\mathcal{P}$ returns the value $F_D(x)$ by calculating the frequency

of x occurrence in D. It is obvious that finding X_{max} and Y_{min} is equivalent to finding the upper zeros and lower units of the function F_D.

The traditional approach to the decoding problem is based on construction of Shannon optimal algorithm (proposed by V.K. Korobkov in 1965). According to this approach, the complexity of the decoding algorithm F should be estimated by the number of calls to the operator B in the worst case. The problem of constructing an optimal algorithm for decoding a monotone Boolean function is solved by Zh. Ansel in 1968. In 1976 V. B. Alekseev constructed a decoding algorithm that is optimal in the case when F is a Boolean function, and is close in complexity to optimal in a more general case, when F is defined on a Cartesian product of finite chains [1].

This paper presents the results of a comparative analysis of two methods for decoding the function F_D for the case when $\mathcal{P}$ is a Cartesian product of finite chains. The first one uses the idea proposed in [2] of sequentially-joint enumeration of X_{max} and Y_{min} and relies on solving the dualization problem over the product of partial orders. The second one is the algorithm of V. B. Alekseev mentioned above. Based on random data for each method under test, the operating time and the number of calls to the operator $B(F_D)$ are estimated. An experimental study was carried out for the case when $\mathcal{P}$ is a Cartesian product of finite chains. The results of the study show that sequential-joint enumeration is the most effective with a large number of n chains in the Cartesian product and with a high valency of each individual chain $\mathcal{P}_i$. The sequential-joint decoding uses an asymptotically optimal dualization algorithm over the product of chains (proposed by E. V. Djukova, G. O. Maslyakov and P. A. Prokofiev in 2017). Asymptotically optimal dualization algorithms have a theoretical justification for efficiency in the typical case and are currently the leaders in counting speed.

Thus, the actual issues of reducing time costs arising from the logical analysis of data with elements from a Cartesian product of finite partially ordered sets are investigated. For the task of decoding a two-valued monotone function taking the value 0 and 1, respectively, on frequent and infrequent elements of the Cartesian product of finite chains, an original algorithm is proposed, conditions of its effectiveness are revealed and the expediency of using asymptotically optimal dualization algorithms is shown.

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

[1] *Alekseev V. B.* O rasshifrovke nekotoryh klassov monotonnyh mnogoznachnyh funkcij // Zhurn. vychisl. matem. i matem. fiziki, 1976. Vol. 16(1). Pp. 189–198.

[2] *Dragunov N. A., Djukova E. V.* Finding Minimal Infrequent and Maximal Frequent Sets in Partially Ordered Data // Mathematical Methods for Pattern Recognition: Book of abstract of the 19th Russian National Conference with International Participation, 2019. Pp. 13–14

Комбинирование рейтингов, полученных из разных источников

Фадеев Егор Павлович[1]⋆ fadeev.ep@physics.msu.ru
Ященко Михаил Андреевич[1] iashchenko.ma18@physics.msu.ru
Зубюк Андрей Владимирович[1] zubjuk@physics.msu.ru
[1]Москва, Московский Государственный Университет имени М.В. Ломоносова, физический факультет

Временами мы опираемся на рейтинги при принятии решений. Например, абитуриент может обращать внимание на рейтинги университетов, чтобы определиться, куда подавать документы, покупатель смотрит на рейтинги товаров, кинолюбитель обращает внимание на рейтинги фильмов на том или ином агрегаторе. В большинстве случаев находится несколько разных рейтингов по одной тематике от разных составителей, которые далеко не всегда согласуются между собой. Это приводит к ситуации, когда непонятно, какое решение принимать: согласно первому рейтингу университет No.1 лучше университета No.2, а согласно второму — наоборот.

В работе предлагается подход к агрегированию таких рейтингов на основе теории возможностей Ю.П. Пытьева, которая в отличие от теории вероятностей, является качественной. Пусть $\Omega = \{\omega_1, \cdots, \omega_N\}$ — множество сравниваемых объектов, тогда распределение возможности на этом множестве — функция $\pi : \Omega \to [0, 1]$, такая что $\max\{\pi(\omega)|\omega \in \Omega\} = 1$. Все значения этой функции кроме нуля могут быть использованы только для сравнений, а $\pi(\omega) = 0$ означает, что ω отсутствует в рейтинге. Таким образом, распределения π_1 и π_2 эквиваленты ($\pi_1 \sim \pi_2$), если для любых сравниваемых объектов $\omega', \omega'' \in \Omega$ соотношение $\pi_1(\omega') \leqslant \pi'(\omega'')$ выполнено тогда и только тогда, когда $\pi_2(\omega') \leqslant \pi_2(\omega'')$.

Рассмотрим произвольный рейтинг на $\Omega = \{\omega_1, \cdots, \omega_N\}$. Существуют рейтинги, которые расставляют сравниваемые объекты по позициям. Объекты на одинаковой позиции неотличимы по этому рейтингу, а объекты на разных позициях можно ранжировать согласно нему. Рейтинг такого типа может быть смоделирован с помощью такого распределения возможности π, что а) $\pi(\omega') = \pi(\omega'')$, если ω' и ω'' занимают одинаковую позицию в рейтинге; б) $\pi_1(\omega') > \pi(\omega'')$, если ω' занимает более высокую позицию в рейтинге, чем ω''. Рейтинги второго типа ставят в соответствие каждому $\omega \in \Omega$ некоторое число $R(\omega)$ (оценку). Рейтинг второго типа можно смоделировать с помощью распределения возможности $\pi(\omega) = R(\omega)/\max\limits_{\Omega} R(\omega)$.

Применение качественной теории возможностей для моделирования рейтингов первого типа оправдано тем фактом, что такие рейтинги сами по себе являются качественными. Рейтинги второго типа являются количественными, но эти оценки часто носят условный и приближенный характер. В таком случае имеет смысл отбросить из рассмотрения разницу между оценками двух сравни-

ваемых объектов, а принимать во внимание только то, оценка какого объекта выше.

Предположим, что существуют два рейтинга π_1 и π_2 на Ω. Тогда может быть интересно сопоставить их между собой и определить, в чем эти рейтинги сходятся и расходятся. В [1] были введены агрегирующие операции супремум $\pi_1 \wedge \pi_2$ и инфимум $\pi_1 \vee \pi_2$ для двух распределений возможности π_1 и π_2 на основе отношения специфичности. Распространяя эту интерпретацию на рейтинги, можно сказать, что рейтинг π_1 не менее специфичен, чем рейтинг π_2, если эти а) рейтинги не противоречат друг другу, т.е. нет сравниваемых объектов, которые ранжируются разным образом согласно рейтингам б) рейтинг π_1 может ранжировать такие объекты, которые неотличимы друг от друга согласно π_2. Инфимум $\pi_1 \wedge \pi_2$ — наименее специфичное распределение возможности из всех распределений, которые не менее специфичны чем оба распределения π_1 и π_2. Аналогично, супремум $\pi_1 \vee \pi_2$ — наиболее специфичное распределение возможности из всех распределений, которые не более специфичны, чем π_1 и π_2.

На рис. 1 приведен пример рейтингов для одних и тех же сравниваемых объектов (игр в жанре Sci-Fi) с разных ресурсов (metacritic.com и GoG.com) и их супремум и инфимум. Инфимум $\pi_1 \wedge \pi_2$ представляет собой такой рейтинг, который ранжирует как можно больше объектов, но не противоречит исходным рейтингам. Если бы присутствовали объекты, ранжируемые исходными рейтингами в противоположных порядках, то их возможность была бы равна 0 в инфимуме. Супремум $\pi_1 \vee \pi_2$ — такой рейтинг, который не ранжирует объекты, которые не удаётся однозначно отранжировать в соответствии со всеми рейтингами. Отличие между супремумом и инфимумом в данном примере наблюдается на объектах No.2, No.3, No.4 и No.7.

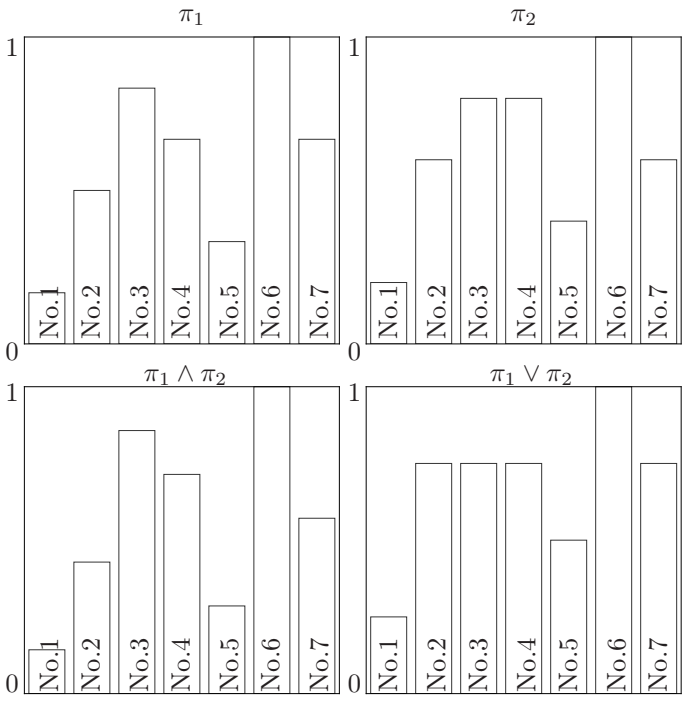


Рис. 1. Рейтинги игр в жанре Sci-Fi на ресурсах metacritic.com (сверху слева) и GoG.com (сверху справа) и их инфимум (снизу слева) и супремум (снизу справа): “Gone Home” (No.1), “Sunless Sea” (No.2), “Distant Worlds: Universe” (No.3), “NEO Scavenger” (No.4), “Halfway” (No.5), “Kerbal Space Program” (No.6), “Satellite Reign” (No.7)

Работа поддержана грантом РФФИ No. 19-29-09044.

[1] *Zubyuk A., Fadeev E.* Aggregation operators for comparative possibility distributions and their role in group decision making // Atlantis Studies in Uncertainty Modelling, 2021. Pp. 608–615.

Aggregation of ratings from various sources

Fadeev Egor[1]⋆ fadeev.ep@physics.msu.ru
Yashchenko Mikhail[1] iashchenko.ma18@physics.msu.ru
Zubyuk Andrey[1] zubjuk@physics.msu.ru
[1]Moscow, Lomonosov Moscow State University, Faculty of Physics

Decisions are made based on ratings in some cases. For example, an enrollee may take into account ratings of universities in order to decide which of them to choose, a customer looks on ratings of products, a viewer may take into account ratings of movies on some aggregater. In many cases, there are several ratings on the topic from different sources, which don't always agree. This leads to a situation, when it isn't clear, which decision is to take: according to the first rating the university No.1 is better than the university No.2, but it is the opposite according to the second rating.

An approach to aggregation of such ratings based on the Pytiev possibility theory is proposed in this paper. The theory is qualityive in contrast to probability one. Let $\Omega = \{\omega_1, \cdots, \omega_N\}$ be the set of all objects being compared. Than, a possibility distribution on ω is a function $\pi : [0,1] \to \Omega$, such that $\max\{\pi(\omega)|\omega \in \Omega\} = 1$. Only zero value of π is meaningful: $\pi(\omega) = 0$ means that ω is absent in the rating. All other values can be used only to compare them. That is, two possibility distributions π_1 and π_2 are equivalent ($\pi_1 \sim \pi_2$) iff for any two objects $\omega'; \omega'' \in \Omega; \pi_1(\omega') \leqslant \pi_1(\omega''), \pi_2(\omega') \leqslant \pi_2(\omega'')$.

Let us consider a rating on $\Omega = \{\omega_1, \cdots, \omega_N\}$. There are two options. A rating of the first type places objects in question in some positions. Objects in the same position cannot be dedistinguished based on the rating. Objects in different positions can be ordered in accordance with the rating. Such a rating can be modeled via a possibility distribution π satisfying a) $\pi(\omega') = \pi(\omega'')$, iff ω' and ω'' are placed in the same position according to the rating; b) $\pi_1(\omega') > \pi(\omega'')$, iff ω' is placed in a higher position according to the rating, than ω''. A rating of the second type maps Ω on $\mathbb{R}$, i.e. for any object ω it gives a number $R(\omega)$ (a grade). Such a rating can be modeled via the possibility distribution $\pi(\omega) = R(\omega)/\max\limits_{\Omega} R(\omega)$.

A rating of the first type is qualitative by itself, therefore it is justified to apply such a qualitative theory as the possibility one. Ratings of the second type are quantitative, but those grades are often arbitrarily to some extent. It makes sense to neglect the difference between to grades and to only take into account the order the grades are ordered.

Suppose, there are two ratings π_1 and π_2 on Ω. Than it can be interesting to compare the ratings and to spot the objects the ratings agree and disagree on. Two aggregative operations the supremum $\pi_1 \wedge \pi_2$ and the infimum $\pi_1 \vee \pi_2$ based on the specifily relation were introduced in [1]. Expanding the terminology on ratings it can be said that a rating π_1 is no less specific than a rating π_2 if a) the ratings don't contradict, i.e. $\nexists \omega', \omega'' : \pi_1(\omega') < \pi_1(\omega'')$ and $\pi_2(\omega') > \pi_2(\omega'')$; b) some

indistinguishible objects according to π_1 can be distinguished (order) according to π_2. The infimum $\pi_1 \wedge \pi_2$ is the least specific possibility distribution amongst all non contradicting distributions more specific than both π_1 and π_2. The supremum $\pi_1 \vee \pi_2$ is the most specific possibility distribution amongst all non contradicting distribution less specific than both π_1 and π_2.

An example of two ratings on the same topic (games in Sci-Fi genre) from different sources (metacritic.com and GoG.com) alongside their infimum and supremum is presented in figure 1. The infimum is the rating, which orders as many objects as possible but doesn't contradict to original ratings. In the case of contradicting on objects ω' and ω'' ratings π_1 and π_2, the possibility of these objects would be equal to 0 according to the infimum of π_1 and π_2. The supremum $\pi_1 \vee \pi_2$ is the rating which doesn't distinguish objects which cannot be distinguished unanimously by both ratings. The difference between the infimum and suprmemum can be spotted on objects No.2, No.3, No.4 and No.7 in the given example.

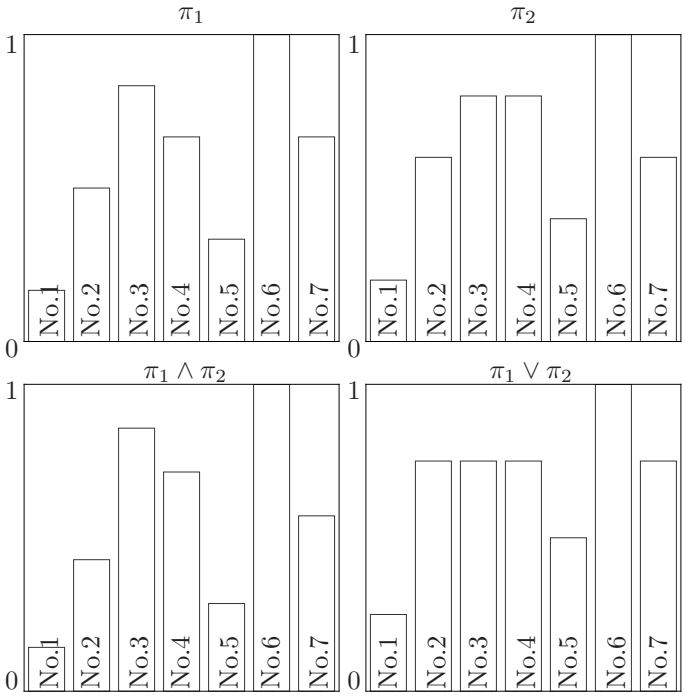


Fig. 1. Ratings of games in Sci-Fi genre from metacritic.com (top left) and GoG.com (top right) and their infimum (bottom left) and supremum (bottom right). Gone Home (No.1), Sunless Sea (No.2), Distant Worlds: Universe (No.3), NEO Scavenger (No.4), Halfway (No.5), Kerbal Space Program (No.6), Satellite Reign (No.7)

This research is funded by RFBR, grant 19-29-09044.

[1] *Zubyuk A., Fadeev E.* Aggregation operators for comparative possibility distributions and their role in group decision making // Atlantis Studies in Uncertainty Modelling, 2021. Pp. 608–615.

Теоретико-информационный подход к построению нижних границ вероятности ошибки в задачах кодирования дискретного источника и классификации данных

Ланге Михаил Михайлович[1]⋆ lange_mm@ccas.ru
Ланге Андрей Михайлович[1] lange_am@mail.ru
[1]Москва, Федеральный исследовательский центр "Информатика и управление"РАН

Исследуется вероятностная модель кодирования зашумленных дискретных сообщений и вероятностная модель классификации данных, для которых найдены зависимости наименьшего количества обрабатываемой информации от допустимой вероятности ошибки. Полученные соотношения являются двухфакторными критериями качества в задачах кодирования и классификации, и аналогичны известной в теории информации функции «скорость-погрешность» (rate distortion function).

Модель кодирования источника. Рассматривается модель кодирования блоков дискретных сообщений, переданных по каналу с шумом. Модель включает последовательные стохастические преобразования $U^n \xrightarrow{P_{V^n|U^n}} V^n \xrightarrow{Q_{\hat{U}^n|V^n}} \hat{U}^n$, которые соответствуют каналу наблюдения и тест-каналу. Здесь U и V – алфавиты сообщений размера $m \geqslant 2$ соответственно на выходе источника и канала наблюдения; U^n и V^n – множества блоков $u^n = (u_1, \ldots, u_n)$, $u_k \in U$, $k = 1, \ldots, n$ и $v^n = (v_1, \ldots, v_n)$, $v_k \in V$, $k = 1, \ldots, n$ размера n; $\hat{U}^n$ – множество блоков $\hat{u}^n = (\hat{u}_1, \ldots, \hat{u}_n)$ на выходе тест-канала, которые воспроизводят блоки $u^n = (u_1, \ldots, u_n)$ с побуквенной погрешностью в метрике Хемминга. На множестве U^n задано безусловное распределение $P_{U^n} = \{P(u^n),\ u^n \in U^n\}$; на множестве V^n – условные распределения $P_{V^n|U^n} = \{P(v^n|u^n),\ v^n \in V,\ u^n \in U^n\}$. Для оптимизации модели на множестве $\hat{U}^n$ вводятся свободные условные распределения $Q_{\hat{U}^n|V^n} = \{Q(\hat{u}^n|v^n),\ \hat{u}^n \in \hat{U}^n,\ v^n \in V^n\}$.

Введенные распределения дают среднюю взаимную информацию на сообщение $n^{-1} I_{Q_{\hat{U}^n|V^n}}(V^n; \hat{U}^n)$ и среднюю вероятность ошибки на сообщение $n^{-1} E_{Q_{\hat{U}^n|V^n}}(V^n; \hat{U}^n)$, зависящие от распределений $Q_{\hat{U}^n|V^n}$. Для рассматриваемой модели в теории информации введена обобщенная функция «скорость-погрешность»

$$R(\varepsilon) = \min_n \min_{Q_{\hat{U}^n|V^n}: n^{-1} E_{Q_{\hat{U}^n|V^n}}(V^n;\hat{U}^n) \leqslant \varepsilon} n^{-1} I_{Q_{\hat{U}^n|V^n}}(V^n; \hat{U}^n),$$

которая ограничивает снизу скорость любого кода с допустимой вероятностью ошибки $\varepsilon > 0$. В данной работе для источника независимых и одинаково распределенных сообщений и канала наблюдения без памяти получена нижняя

граница функции $R(\varepsilon)$:

$$R(\varepsilon) \geqslant \underline{R}(\varepsilon) = I(V;U) - h(\varepsilon - \varepsilon_{\min}) - (\varepsilon - \varepsilon_{\min})\ln(m-1), \quad \varepsilon_{\min} \leqslant \varepsilon \leqslant \varepsilon_{\max},$$

где $0 \leqslant I(V;U) \leqslant H(U)$ – средняя взаимная информация между V и U, $H(U)$ – энтропия множества сообщений источника, $h(z) = -z\ln z - (1-z)\ln(1-z)$, $\underline{R}(\varepsilon_{\min}) = I(V;U)$ и $\underline{R}(\varepsilon_{\max}) = 0$. В общем случае $\varepsilon_{\min}$ зависит от условной энтропии $H(U|V) = H(U) - I(V;U)$, а $\varepsilon_{\max} = (m-1)\min\limits_{u\in U} P(u)$. В случае равновероятных сообщений имеем $\varepsilon_{\max} = (m-1)/m$ и асимптотику $\varepsilon_{\min} = (2H(U|V)/(m-1))^{1/2}(m-1)/m$ при малых значениях $H(U|V)$. Для бесшумного канала наблюдения имеем $H(U|V) = 0$ и $\underline{R}(\varepsilon)$ совпадает с границей Шеннона.

Модель классификации данных. Модель классификации групповых объектов задается парой стохастических преобразований $\boldsymbol{\Omega} \xrightarrow{P_{\mathbf{X}^n|\boldsymbol{\Omega}}} \mathbf{X}^n \xrightarrow{Q_{\hat{\boldsymbol{\Omega}}|\mathbf{X}^n}} \hat{\boldsymbol{\Omega}}$. Здесь $\boldsymbol{\Omega} = \{\omega_i,\ i = 1,\ldots,c\}$ и $\hat{\boldsymbol{\Omega}} = \{\omega_j,\ j = 1,\ldots,c\}$, $c \geqslant 2$ – множества классов и решений о классах по блокам $\mathbf{x}^n = (\mathbf{x}_1,\ldots,\mathbf{x}_n) \in \mathbf{X}^n$ из n объектов $\mathbf{x}_k \in \mathbf{X}$, $k = 1,\ldots,n$, одного класса. На множестве $\boldsymbol{\Omega}$ задано априорное распределение $P_{\boldsymbol{\Omega}} = \{P(\omega_i),\ i = 1,\ldots,c\}$; на множестве X^n – условные по классам распределения $P_{\mathbf{X}^n|\boldsymbol{\Omega}} = \{P(\mathbf{x}^n|\omega_i), \mathbf{x}^n \in \mathbf{X}^n,\ i = 1,\ldots,c\}$. Для оптимизации модели на множестве $\hat{\boldsymbol{\Omega}}$ используются свободные условные распределения $Q_{\hat{\boldsymbol{\Omega}}|\mathbf{X}^n} = \{Q(w_j|\mathbf{x}^n),\ j = 1,\ldots,c, \mathbf{x}^n \in \mathbf{X}^n\}$. Полагая, что погрешность между значениями $\omega_i \in \Omega$ и $\omega_j \in \hat{\Omega}$ измеряется в метрике Хемминга $[\omega_i \neq \omega_j]$, введенные распределения дают среднюю взаимную информацию $I_{Q_{\hat{\boldsymbol{\Omega}}|X^n}}(\mathbf{X}^n;\hat{\boldsymbol{\Omega}})$ и среднюю вероятность ошибки $E_{Q_{\hat{\boldsymbol{\Omega}}|X^n}}(\mathbf{X}^n;\hat{\boldsymbol{\Omega}})$ +ионалов, зависящих от $Q_{\hat{\boldsymbol{\Omega}}|X^n}$.

При заданных значениях $\varepsilon > 0$ эти функционалы позволяют ввести функцию

$$\tilde{R}(\varepsilon) = \min_n \min_{Q_{\hat{\boldsymbol{\Omega}}|\mathbf{X}^n}: E_{Q_{\hat{\boldsymbol{\Omega}}|\mathbf{X}^n}}(\mathbf{X}^n,\hat{\boldsymbol{\Omega}}) \leqslant \varepsilon} I_{Q_{\hat{\boldsymbol{\Omega}}|\mathbf{X}^n}}(\mathbf{X}^n;\hat{\boldsymbol{\Omega}}),$$

для которой в работе [1] получена нижняя граница

$$\tilde{R}(\varepsilon) \geqslant \underline{\tilde{R}}(\varepsilon) = I(\mathbf{X};\Omega) - h(\varepsilon - \tilde{\varepsilon}_{\min}) - (\varepsilon - \tilde{\varepsilon}_{\min})\ln(c-1), \quad \tilde{\varepsilon}_{\min} \leqslant \varepsilon \leqslant \tilde{\varepsilon}_{\max}.$$

Здесь $\underline{\tilde{R}}(\tilde{\varepsilon}_{\min}) = I(\mathbf{X},\boldsymbol{\Omega}) = H(\Omega) - H(\Omega|\mathbf{X})$, $\underline{\tilde{R}}(\tilde{\varepsilon}_{\max}) = 0$; $\tilde{\varepsilon}_{\min}$ зависит от $H(\Omega|\mathbf{X})$ и $\tilde{\varepsilon}_{\max} = (c-1)\min_i P(\omega_i)$. В случае равновероятных классов для $\tilde{\varepsilon}_{\min}$ и $\tilde{\varepsilon}_{\max}$ справедливы формулы, приведенные в модели кодирования, с заменами $H(U|V) = H(\boldsymbol{\Omega}|\mathbf{X})$ и $m = c$.

Примеры вычисления границ $\underline{R}(\varepsilon)$ и $\underline{\tilde{R}}(\varepsilon)$. В силу монотонного убывания полученных границ, их обращения $\underline{R}^{-1}(I)$ и $\underline{\tilde{R}}^{-1}(I)$ дают нижние границы

вероятностей ошибки кодирования и классификации при фиксированных значениях количества обрабатываемой информации I. Ниже приведены примеры вычисления функций $\underline{R}(\varepsilon)$ и $\underline{\tilde{R}}(\varepsilon)$ для равновероятных сообщений источника ($m \geqslant 2$) и равновероятных классов ($c \geqslant 2$).

Характеристики $I(V;U)$ и $\varepsilon_{\min}$ вычислены для симметричного канала наблюдения без памяти с фиксированной вероятностью ошибки δ передачи любого символа источника. В этом случае $I(V;U) = \ln m - h(\delta) - \delta \ln(m-1)$ и при малых значениях δ условная энтропия $H(U|V) = h(\delta) + \delta \ln(m-1)$ дает асимптотику $\varepsilon_{\min}$. Характеристики $I(\mathbf{X};\mathbf{\Omega})$ и $\tilde{\varepsilon}_{\min}$ вычислены для канала наблюдения без памяти с условными по классам вероятностями $P(\mathbf{x}|\omega_i)$, $\mathbf{x} \in \mathbf{X}$, $i = 1, \ldots, c$, которые заданы убывающими экспонентами от квадратов расстояний между объектом $\mathbf{x}$ и «центральными» представителями классов. Численные реализации функции $\underline{\tilde{R}}(\varepsilon)$ на множествах изображений лиц и подписей даны в работе [1].

Планируется исследовать избыточность вероятности ошибки классификации относительно нижней границы для различных решающих алгоритмов.

[1] *Ланге А. М., Ланге М. М., Парамонов С. В.* О соотношении взаимной информации и вероятности ошибки в задаче классификации данных // ЖВМ МФ, 2021. Т. 61(7). С. 1192–1205.

Information-theoretical approach to construct lower bounds for error probability in tasks of discrete source coding and data classification

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

Probabilistic models for discrete noisy source coding and data classification are studied. For these models, the dependences of a minimal information on a given admissible error probability are found. The obtained relations yield the two-factor fidelity criterions in source coding and data classification tasks and these relations are similar to the rate distortion function known in the information theory.

Source coding model. The model of coding discrete letters transmitted via a noisy channel is considered. This model includes a pair of the probabilistic transformations $U^n \xrightarrow{P_{V^n|U^n}} V^n \xrightarrow{Q_{\hat{U}^n|V^n}} \hat{U}^n$ that correspond to the observation channel and the test-channel, respectively. Here, U and V are the alphabets of size $m \geqslant 2$ for the output letters of the source and the observation channel; U^n and V^n are the appropriate sets of n-letter blocks $u^n = (u_1, \ldots, u_n)$, $u_k \in U$, $v^n = (v_1, \ldots, v_n)$, $v_k \in V$, $k = 1, \ldots, n$; $\hat{U}^n$ is a set of the output blocks $\hat{u}^n = (\hat{u}_1, \ldots, \hat{u}_n)$ that reproduce the source blocks $u^n = (u_1, \ldots, u_n)$ with the single-letter Hamming distortion. A probability distribution $P_{U^n} = \{P(u^n),\ u^n \in U^n\}$ and the conditional distributions $P_{V^n|U^n} = \{P(v^n|u^n),\ v^n \in V,\ u^n \in U^n\}$ are given. To optimize the model, the free conditional distributions $Q_{\hat{U}^n|V^n} = \{Q(\hat{u}^n|v^n),\ \hat{u}^n \in \hat{U}^n,\ v^n \in V^n\}$ are introduced in the set $\hat{U}^n$.

The above distributions provide the average mutual information per one letter $n^{-1} I_{Q_{\hat{U}^n|V^n}}(V^n; \hat{U}^n)$ and the corresponding average error probability $n^{-1} E_{Q_{\hat{U}^n|V^n}}(V^n; \hat{U}^n)$ so that these functionals depend on the distributions $Q_{\hat{U}^n|V^n}$. Given $\varepsilon > 0$, in the information theory, the general rate distortion function is defined by the following conditional minimum

$$R(\varepsilon) = \min_n \min_{Q_{\hat{U}^n|V^n}: n^{-1} E_{Q_{\hat{U}^n|V^n}}(V^n;\hat{U}^n) \leqslant \varepsilon} n^{-1} I_{Q_{\hat{U}^n|V^n}}(V^n; \hat{U}^n)$$

that constrains a code rate from below subject to an admissible error probability $\varepsilon > 0$. For the independent and identically distributed source letters as well as for the memoryless observation channel, we give a lower bound to the function $R(\varepsilon)$ as follows

$$R(\varepsilon) \geqslant \underline{R}(\varepsilon) = I(V;U) - h(\varepsilon - \varepsilon_{\min}) - (\varepsilon - \varepsilon_{\min}) \ln(m-1), \qquad \varepsilon_{\min} \leqslant \varepsilon \leqslant \varepsilon_{\max}.$$

Here, $0 \leqslant I(V;U) \leqslant H(U)$ is the average mutual information between V and U, $H(U)$ is the entropy of the source letters, $h(z) = -z\ln z - (1-z)\ln(1-z)$, $\underline{R}(\varepsilon_{\min}) = I(V;U)$, and $\underline{R}(\varepsilon_{\max}) = 0$. Generally, the value $\varepsilon_{\min}$ depends on the conditional entropy $H(U|V) = H(U) - I(V;U)$ whereas $\varepsilon_{\max} = (m-1)\min\limits_{u\in U} P(u)$. For the equiprobable source letters, we have $\varepsilon_{\max} = (m-1)/m$ and the asymptotical value $\varepsilon_{\min} = (2H(U|V)/(m-1))^{1/2}(m-1)/m$ when the entropy $H(U|V)$ is sufficiently small. In the case of the noiseless observation channel, $H(U|V) = 0$ and $\underline{R}(\varepsilon)$ coincides with the Shannon bound.

Data classification model. For the group objects, the classification model is given by a pair of the probabilistic transformations $\mathbf{\Omega} \xrightarrow{P_{\mathbf{X}^n|\mathbf{\Omega}}} \mathbf{X}^n \xrightarrow{Q_{\hat{\mathbf{\Omega}}|\mathbf{X}^n}} \hat{\mathbf{\Omega}}$, where $\mathbf{\Omega} = \{\omega_i,\ i = 1,\dots,c\}$ and $\hat{\mathbf{\Omega}} = \{\omega_j,\ j = 1,\dots,c\}$, $c \geqslant 2$, are the sets of the class labels and the class label decisions about the group objects given by the blocks $\mathbf{x}^n = (\mathbf{x}_1,\dots,\mathbf{x}_n) \in \mathbf{X}^n$ of n objects $\mathbf{x}_k \in \mathbf{X}$, $k = 1,\dots,n$ of the same class. Let $P_{\mathbf{\Omega}} = \{P(\omega_i),\ i = 1,\dots,c\}$ be a given prior distribution in the set $\mathbf{\Omega}$ and $P_{\mathbf{X}^n|\mathbf{\Omega}} = \{P(\mathbf{x}^n|\omega_i), \mathbf{x}^n \in \mathbf{X}^n,\ i = 1,\dots,c\}$ be the given class-conditional distributions in the set X^n. To optimize the model, we use the free conditional distributions $Q_{\hat{\mathbf{\Omega}}|\mathbf{X}^n} = \{Q(w_j|\mathbf{x}^n),\ j = 1,\dots,c, \mathbf{x}^n \in \mathbf{X}^n\}$ in the set $\hat{\mathbf{\Omega}}$.

If the distortion between $\omega_i \in \Omega$ and $\omega_j \in \hat{\Omega}$ is defined by the Hamming metric $[\omega_i \neq \omega_j]$, the above distributions provide the average mutual information $I_{Q_{\hat{\mathbf{\Omega}}|X^n}}(\mathbf{X}^n;\hat{\mathbf{\Omega}})$ and the average error probability $E_{Q_{\hat{\mathbf{\Omega}}|X^n}}(\mathbf{X}^n;\hat{\mathbf{\Omega}})$ as the functionals of the distributions $Q_{\hat{\mathbf{\Omega}}|X^n}$. Given $\varepsilon > 0$, these functionals yield the function

$$\tilde{R}(\varepsilon) = \min_n \min_{Q_{\hat{\mathbf{\Omega}}|\mathbf{X}^n}: E_{Q_{\hat{\mathbf{\Omega}}|\mathbf{X}^n}}(\mathbf{X}^n,\hat{\mathbf{\Omega}}) \leqslant \varepsilon} I_{Q_{\hat{\mathbf{\Omega}}|\mathbf{X}^n}}(\mathbf{X}^n;\hat{\mathbf{\Omega}}),$$

that is similar to the general rate distortion function $R(\varepsilon)$. In the paper [1], the lower bound to the function $\tilde{R}(\varepsilon)$ has been obtained as follows

$$\tilde{R}(\varepsilon) \geqslant \underline{\tilde{R}}(\varepsilon) = I(\mathbf{X};\Omega) - h(\varepsilon - \tilde{\varepsilon}_{\min}) - (\varepsilon - \tilde{\varepsilon}_{\min})\ln(c-1), \qquad \tilde{\varepsilon}_{\min} \leqslant \varepsilon \leqslant \tilde{\varepsilon}_{\max},$$

where $\underline{\tilde{R}}(\tilde{\varepsilon}_{\min}) = I(\mathbf{X},\mathbf{\Omega}) = H(\Omega) - H(\Omega|\mathbf{X})$ and $\underline{\tilde{R}}(\tilde{\varepsilon}_{\max}) = 0$. Similarly, the value $\tilde{\varepsilon}_{\min}$ depends on the conditional entropy $H(\Omega|\mathbf{X})$ whereas $\tilde{\varepsilon}_{\max} = (c-1)\min_i P(\omega_i)$. For the same prior class probabilities, $\tilde{\varepsilon}_{\min}$ and $\tilde{\varepsilon}_{\max}$ are defined by the formulas shown in the source coding model, where $H(U|V) = H(\mathbf{\Omega}|\mathbf{X})$ and $m = c$.

Calculation of the bounds $\underline{R}(\varepsilon)$ and $\underline{\tilde{R}}(\varepsilon)$. Since the obtained bounds decrease monotonically, the appropriate inversions $\underline{R}^{-1}(I)$ and $\underline{\tilde{R}}^{-1}(I)$ yield the lower bounds to error probabilities subject to a given information quantity I. The calculated bounds $\underline{R}(\varepsilon)$ and $\underline{\tilde{R}}(\varepsilon)$ have been obtained for the equiprobable source letters ($m \geqslant 2$) and classes ($c \geqslant 2$).

The characteristics $I(V;U)$ and $\varepsilon_{\min}$ have been calculated for the symmetrical memoryless observation channel with a fixed letter transmission error probability δ.

In this case, $I(V;U) = \ln m - h(\delta) - \delta \ln(m-1)$ and, for the small value δ, the conditional entropy $H(U|V) = h(\delta) + \delta \ln(m-1)$ yields the asymptotical value $\varepsilon_{\min}$. The characteristics $I(\mathbf{X};\mathbf{\Omega})$ and $\tilde{\varepsilon}_{\min}$ have been calculated for the memoryless observation channel, whose class-conditional probabilities $P(\mathbf{x}|\omega_i)$, $\mathbf{x} \in \mathbf{X}$, $i = 1, \ldots, c$ are given by the decreasing exponential functions of the squared distances between any object $\mathbf{x}$ and the central representatives of the classes. The calculations have been performed in the datasets of face and signature images. The plots of the calculated bounds are given in the paper [1].

Also, we plan to study a redundancy of the classification error probability relative to the lower bound for different decision algorithms.

[1] *Lange M. M., Lange A. M., Paramonov S V.* Tradeoff Relation between Mutual Information and Error Probability in Data Classification Problems // Computational Mathematics and Mathematical Physics, 2021. Vol. 61(7). Pp. 1181–1193.

Алгоритм распознавания, основанный на иерархической кластеризации с метрикой специального вида

Сенько Олег Валентинович[1,2] senkoov@mail.ru
Салманов Махир Юсиф оглы[1]⋆ sy.mahir@gmail.com
[1]Москва, МГУ им. М.В.Ломоносова
[2]Москва, ФИЦ ИУ РАН

Задача классификации является одним из центральных задач методов машинного обучения. Во многих алгоритмах классификации используются принцип компактности [1], предполагающий, что каждому из распознаваемых классов соответствуют одна или несколько областей признакового пространства, в котором этот класс преобладает. При этом также предполагается, что области имеют относительно простую и компактную геометрическую форму. Естественным способом обучения в рамках гипотезы компактности является выделение для каждого класса соответствующих ему областей. Предполагается также существование простой и надёжной процедуры установление принадлежности распознаваемого объекта каждой из областей. Преимуществом указанного подхода является его прозрачность, информативность и интерпретируемость.

Возможным способом его реализации является использование методов кластерного анализа, позволяющий выделять группы объектов, являющихся близкими друг к другу в смысле некоторой выбранной метрики. Однако традиционный кластерный анализ не позволяет учесть, каким классам принадлежат группируемые объекты, что может приводить к выделению кластеров без существенного преобладания какого-либо из классов. Поэтому нами предполагается новый подход, являющийся модификацией известного метода агломеративной иерархической кластеризации. Для обеспечения преобладания в кластере объектов одного из классов предлагается использовать индексы неоднородности, применяемые при обучении решающих деревьев [2]. В настоящей работе используется энтропийный индекс.

Идея метода заключается в кластеризации объектов с использованием модифицированной метрики, которая состоит из суммы некоторой метрики расстояния и энтропийного индекса. Соответственно метрика, используемая в описанном алгоритме выглядит следующим образом:

$$\rho\_dist(cluster_i, cluster_j) \; + \; \alpha * entropy\_index(cluster_i, cluster_j),$$

где ρ_dist - метрика, оценивающая расстояние между кластерами $cluster_i$ и $cluster_j$,
$entropy\_index$ - индекс неоднородности при объединении кластеров $cluster_i$ и $cluster_j$,
α - гиперпараметр, который позволяет управлять влиянием энтропии при кластеризации. Два кластера объединяются, когда модифицированное расстояние

оказывается минимальным. Можно заметить, что при больших значениях α кластера формируются, в основном, из объектов обучающей выборки, имеющих одинаковое целевое значение. Подобное взвешивание метрик приведет к негативному эффекту, так как никакого обучения не будет - кластера "скопируют"группы из обучающей выборки. Поэтому при обучении необходимо избегать завышения значене параметра α. Эффективнось метода во многм зависит от правильного выбора шага, на котором происходит прерывание процесса слияния кластеров. В настоящей работе процесс прерывался, когда общее число классов оказывалось равным числу распознаваемых классов.

Предложенный алгоритм применяется к задаче предсказания типа кристаллической структуры галогенидов по их химической формуле. В рассматриваемой выборке данных содержится 551 описаний галогенидов с 11 типами кристаллической структуры. При обучении данные поделены в соотношении 70/30 на обучающую и тестовую подвыборки. В качестве расстояния между кластерами ρ_dist используется метрика

$$max\_dist(cluster_i, cluster_j) \; + \; mean\_dist(cluster_i, cluster_j),$$

где $max\_dist(cluster_i, cluster_j)$ - максимальное расстояние между объектами кластеров $cluster_i$ и $cluster_j$,
$mean\_dist(cluster_i, cluster_j)$ - среднее расстояние между объектами кластеров $cluster_i$ и $cluster_j$.

В итоге применение описанного алгоритма получено 11 кластеров. Поиск оптимального соответствия между выделенными кластерами и типами кристаллической структуры сводился к решению задачи максимизации общее число верно классифицированных галогенидов. Данная задача эквивалентна известной оптимизационной задачи о назначениях. Для её решения был использован венгерский алгоритм.

В результате оптимальным значением гиперпараметра α стал 20. При этом процент верно классифицированных объектов равен 93% на обучающей выборке и 70% на тестовой выборке. Данный результат примерно соответствует результатам, полученным с помощью альтернативных методов. Однако разработанная технология позволяет получить дополнительную важную информацию о распределении объектов по кластерам с известной локализацией в пространстве признаков. Полученные результаты в целом свидетельствуют о потенциальной перспективности подхода. Однако они указывают на необходимость даьнейших исследований, связанных, в частности, с установлением оптимального числа кластеров.

Работа поддержана грантом РФФИ No. 21-51-53019.

[1] *Загоруйко Н. Г.* Гипотезы компактности и λ-компактности в методах анализа данных // Сиб. журн. индустр. матем., 1998. Т. 1(1). С. 114–126.

[2] *Hastie T., Tibshirani R., Friedman J.* The elements of statistical learning : Data mining, inference, and prediction. Second Edition // New York:: Springer Verlag, 2009.

Recognition algorithm based on hierarchical clustering with a metric of a special kind

Senko Oleg[1,2] senkoov@mail.ru
Salmanov Mahir[1]⋆ sy.mahir@gmail.com
[1]Moscow, M.V.Lomonosov Moscow State University
[2]Moscow, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences

The classification task is one of the central tasks of machine learning methods. Many classification algorithms use the principle of compactness [1], which assumes that each of the recognized classes corresponds to one or more regions of the feature space where that class predominates. It is also assumed that the regions have a relatively simple and compact geometric shape. A natural way of learning under the compactness hypothesis is to select the appropriate regions for each class. It is also assumed that there is a simple and reliable procedure to determine the membership of the detected object to each of the regions. The advantage of this approach is its transparency, information content and interpretability.

One possible way of implementation is the use of cluster analysis methods, which allows filtering out groups of objects that are close in terms of a selected metric. However, traditional cluster analysis cannot take into account to which classes the grouped objects belong, which can lead to the selection of clusters without significant dominance of one of the classes. Therefore, we assume a new approach, which is a modification of the well-known method of agglomerative hierarchical clustering [2]. To ensure the predominance of objects of one of the classes in a cluster, it is proposed to use heterogeneity indices, which are used in training decision trees. In this paper, the entropy index is used.

The idea behind the method is as follows: clustering with the help a modified metric consisting of the sum of a certain distance metric and the entropy index. Accordingly, the metric used in the described algorithm is as follows:

$$\rho\_dist(cluster_i, cluster_j) \; + \; \alpha * entropy\_index(cluster_i, cluster_j),$$

where ρ_dist is a metric that estimates the distance between $cluster_i$ and $cluster_j$, $entropy\_index$ is an index of heterogeneity when combining $cluster_i$ and $cluster_j$, α is a hyperparameter to control the influence of entropy in clustering. Two clusters are merged when modified metric is minimal. It can be seen that for large values of α, the cluster is formed mainly from objects in the training sample that have the same target value. Such weighting of the metrics leads to a negative effect, as no training takes place - the clusters are "copied" from groups in the training sample. Therefore, when training, it is necessary not to overestimate the value of the α parameter. The effectiveness of the method in many ways depends on the

correct choice of the step at which the process of cluster merging is interrupted. In this work the process was interrupted when the total number of classes was equal to the number of recognized classes.

The developed algorithm was applied to the problem of predicting the nature of the crystal structure of halides based on their chemical formula. The studied data sample contains 551 descriptions of halides. During training, the data is divided into training and test subsamples in a ratio of 70/30. The distance between clusters ρ_dist was used as the metric,

$$max\_dist(cluster_i, cluster_j) \; + \; mean\_dist(cluster_i, cluster_j),$$

where $max\_dist(cluster_i, cluster_j)$ - maximum distance between objects of $cluster_i$ and $cluster_j$,
$mean\_dist(cluster_i, cluster_j)$ - average distance between objects in $cluster_i$ and $cluster_j$.

As a result 11 clusters were received, using the described algorithm. The search for the optimal correspondence between the selected clusters and types of crystal structure was reduced to solving the problem of maximizing the total number of correctly classified halides. This problem is equivalent to the well-known optimization assignment problem. The Hungarian algorithm was used to solve it.

The Hungarian algorithm was applied to the obtained clusters, which finds the optimal correspondance between the selected clusters and the original classes of the objects. The result was that the optimal value of the hyperparameter was 20. At the same time, the percentage of correctly classified objects is 93% in the training set and 70% in the test set.This result roughly corresponds to the results obtained using alternative methods. However, the developed technique makes it possible to obtain additional important information about the distribution of objects across clusters with known localization in the feature space. On the whole, the results obtained indicate that the approach is potentially promising. However, they point to the need for further research related, in particular, to the evaluating the optimal number of clusters

This work was supported by a grant from the RFBR No. 21-51-53019.

[1] *Zagoruiko N.* Compactness and *lambda* -compactness Hypotheses in Data Analysis Methods (in russian) // Siberian Journal of Industrial Mathematics , 1998. Vol. 1(1). Pp. 114–126.

[2] *Hastie T., Tibshirani R., Friedman J.* The elements of statistical learning : Data mining, inference, and prediction. Second Edition // New York:: Springer Verlag, 2009.

Корректная классификация над произведением частичных порядков

Дюкова Елена Всеволодовна[1] edjukova@mail.ru
Масляков Глеб Олегович[1]⋆ gleb-mas@mail.ru
[1]Москва, ВЦ ФИЦ ИУ РАН

В задаче классификации под прецедентной (обучающей) информацией понимается совокупность примеров изучаемых объектов, в которой каждый объект представлен в виде числового вектора, полученного на основе измерения или наблюдения ряда его параметров или характеристик, называемых признаками. Каждый пример (обучающий объект или прецедент) приписан к определённому классу объектов. Требуется по признаковому описанию предъявленного объекта, о котором заранее неизвестно, какому классу он принадлежит, определить (распознать) этот класс.

Основное достоинство логического подхода к задаче классификации на основе прецедентов — возможность получения результата при отсутствии дополнительных предположений вероятностного характера и при небольшом числе прецедентов. Предполагается, что каждый признак принимает ограниченное число допустимых значений, и для каждого признака задаётся бинарная функция близости между его значениями, что позволяет проводить сравнение описания распознаваемого объекта с описаниями прецедентов. Анализ прецедентной информации сводится к поиску в исходных данных определенных закономерностей, позволяющих различать объекты из разных классов. Найденные закономерности имеют содержательное описание в терминах той прикладной области, в которой решается задача. По их наличию или, наоборот, отсутствию в описании распознаваемого объекта, решается вопрос о его классификации. При этом большое внимание уделяется вопросам синтеза алгоритмов, безошибочно классифицирующих материал обучения. Такие алгоритмы называются корректными. К наиболее известным логическим классификаторам относятся алгоритмы корректного голосования, предложенные впервые в отечественных работах (М.М. Бонгард, М.Н. Вайнцвайг, Ю.И. Журавлёв), а также методы Logical Analysis of Data (П. Хаммер, 1986 г.) и Formal Concept Analysis (Р. Вилле, 1981 г.). Последние два направления логической классификации в основном нацелены на обработку бинарных данных.

Алгоритмы корректного голосования базируются на поиске закономерностей, которые представляют собой специальные наборы из допустимых значений признаков, называемые корректными элементарными классификаторами. Для описания таких классификаторов наравне с комбинаторным аппаратом применяется аппарат логических функций. Тогда элементарный классификатор (эл.кл.) — это элементарная конъюнкция, определённая на признаковых описаниях объектов. Если на описании некоторого объекта элементарная конъюнкция принимает значение 1, то говорят, что этот объект содержит данный

эл.кл. Эл.кл. корректен для некоторого класса K, если нельзя указать пару прецедентов, одновременно содержащих этот эл.кл., причём один из прецедентов принадлежит классу K, а другой ему не принадлежит. Как правило, ищутся эл.кл. с небольшим рангом, в частности, тупиковые корректные эл.кл., среди которых отбираются наиболее информативные. Поиск тупиковых корректных эл.кл. сводится к задаче монотонной дуализации (поиску неприводимых покрытий булевой матрицы), которая относится к классу труднорешаемых. Мировыми лидерами по скорости счёта являются асимптотически оптимальные алгоритмы монотонной дуализации.

Каждый найденный корректный эл.кл. класса K участвует в процедуре голосования с целью вычисления общей оценки принадлежности распознаваемого объекта классу K. Распознаваемый объект относится к классу, получившему наибольшую оценку, или алгоритм отказывается от распознавания, если классов с наибольшей оценкой несколько. Классификаторы отличаются используемыми множествами корректных эл.кл. и способами вычисления оценок принадлежности распознаваемого объекта классам. Хорошие результаты показывает голосование по (тупиковым) представительным эл.кл. классов, которые могут порождаться, например, специальными наборами признаков — (тупиковыми) тестами, а также могут порождаться жадным способом на основе построения корректных решающих деревьев.

Стандартные постановки логической классификации не всегда позволяют решать прикладные задачи со сложными отношениями на множествах допустимых значений признаков. В ряде работ авторов на базе обобщения классических понятий логической классификации предложены корректные процедуры классификации при условии, что признаковые описания объектов являются элементами декартова произведения конечных частично упорядоченных множеств, и рассмотрены вопросы применения асимптотически оптимальных алгоритмов перечисления элементарных классификаторов общего вида.

На практике порядки на множествах значений признаков могут быть не заданы. Актуальными являются вопросы упорядочения значений признаков на этапе предварительного анализа обучающей выборки.

В настоящей работе представлены исследования, касающиеся возможности построения частичных порядков на множествах допустимых значений признаков, обеспечивающих корректную классификацию обучающей выборки. Сформулирован критерий корректности логической классификации над произведением частичных порядков, согласно которому для корректной классификации описание каждого прецедента из каждого класса должно быть «независимым» от множества описаний прецедентов из других классов. Показано, что выбор «корректных» частичных порядков на множествах допустимых значений признаков может быть сведён к построению и анализу неприводимых покрытий специальной булевой матрицы. Предложены два подхода к рассматриваемой задаче. Первый подход заключается в построении частичных порядков, обеспе-

чивающих корректную классификацию всех прецедентов, второй в построении для каждого класса частичных порядков, гарантирующих корректную классификацию только прецедентов данного класса. Проведено экспериментальное сравнение предложенных подходов с быстрым методом линейного упорядочения значений признаков [1], который существенно повышает качество классификации, но не гарантирует корректность классификации.

Работа частично поддержана грантом РФФИ No. 19-01-00430.

[1] *Бакланова А. О., Дюкова Е. В, Масляков Г. О.* Исследование зависимости качества классификации от выбора частичных порядков на множествах значений признаков // 9-я международная конференция. «Интеллектуализация обработки информации», 2020. С. 21–25.

Correct classification over a product of partial orders

Djukova Elena[1] edjukova@mail.ru
Masliakov Gleb[1]⋆ gleb-mas@mail.ru
[1]Moscow, CC FRC CSC RAS

In supervised classification problem, the training data is a set of examples of objects under examination in which each object is represented by a numerical vector obtained by measuring or observing its parameters or characteristics called features. Each example (training object or precedent) belongs to a certain class of objects. Given a description in terms of features of a presented object, about which it is unknown to which class it belongs, it is required to find out (recognize) the class it belongs to.

The main advantage of the logical approach to the supervised classification problem is the possibility to obtain results without additional probabilistic assumptions and using a small number of precedents. Each feature is supposed to take a limited number of acceptable values that and for each feature a binary function of proximity between its values is defined that allows to compare the description of the recognized object with the descriptions of precedents. The analysis of training data is reduced to finding certain dependences that differentiate objects belonging to different classes. The found dependences have a meaningful description in terms of the applied field in which the problem is being solved. By their presence or, conversely, absence in the description of the recognized object, the question of its classification is solved. Special attention is given to synthesizing algorithms that unmistakably classify the training objects. Such algorithms are known as correct. The most well-known algorithms in this field are the correct voting algorithms proposed for the first time in Russian works (M.M. Bongard, M. N. Weinzweig, Yu.I. Zhuravlev), and also the methods of Logical Analysis of Data (P. Hammer, 1986) and Formal Concept Analysis (R. Ville, 1981). The last two logical classification areas are aimed mainly at analyzing binary data.

The algorithms of correct voting are based on the search for patterns, which are special sets of acceptable feature values called correct elementary classifiers. The apparatus of logical functions along with the combinatorial apparatus is used to describe such classifiers. Then the elementary classifier (el.kl.) is an elementary conjunction defined on the feature descriptions of objects. If an elementary conjunction takes the value 1 on the description of some object, then it is said that this object contains this el.cl. El.cl. is correct for some class K, if it is impossible to specify a pair of precedents that simultaneously contain this el.cl., and one of the precedents belongs to the class K, and the other does not belong to it. As a rule, el.cl. with a small rank, in particular irredundant correct el.kl, are searched for, among which the most informative are selected. The search for irredundant correct el.kl. is reduced to the monotone dualization problem (search for irreducible coverings of a Boolean matrix), which belongs to the class of intractable problems. The

world best algorithms in computational speed are asymptotically optimal monotone dualization algorithms.

Each found correct el.cl. of the class K participates in the voting procedure in order to calculate the overall estimate of the recognized object membership to the class K. The recognized object belongs to the class that received the highest score, or the algorithm refuses recognition if there are several classes with the highest score. Classifiers differ in the used sets of correct el.cl. and in the methods of calculating the estimates of the recognized object membership to classes. Good results are shown by voting on (irredundant) representative el.cl. that can be generated, for example, by special sets of features called irredundant tests. They also can be generated in a greedy way based on the construction of correct decision trees.

Standard statements of logical classification do not always allow us to solve applied problems with complex relations on sets of acceptable feature values. In a number of our works on the basis of generalization of classical concepts of logical classification classical concepts correct classification procedures are proposed, provided that the feature descriptions of objects are elements of the Cartesian product of finite partially ordered sets. The questions of the application of asymptotically optimal algorithms for the enumeration of elementary classifiers of the general type are considered.

In practice, the partial orders on the sets of feature values can be not specified. In order to improve the quality of classification, the problems of ordering feature values at the stage of preliminary analysis of the training set are considered.

This paper presents the studies of the possibility for constructing partial orders on the sets of acceptable feature values that provide the correct classification of the training set. A criterion for correct logical classification over a product of partial orders is formulated, according to which, for correct classification the description of each precedent from each class must be "independent" of the set of descriptions of precedents from other classes. It is shown that the problem of choosing "correct" partial orders on the sets of acceptable feature values can be reduced to constructing and analyzing irreducible coverings of a special Boolean matrix. Two approaches to the problem under consideration are proposed. The first approach consists in constructing partial orders that ensure the correct classification of all precedents, the second consists in constructing partial orders for each class that guarantee the correct classification of only the precedents of this class. An experimental comparison of the proposed approaches with the fast method of linear ordering of feature values [1], which significantly improves the quality of classification, but does not guarantee the correctness of classification, is carried out.

This research is partially funded by RFBR, grant 19-01-00430.

[1] *Baklanova A.., Djukova E., Masliakov G.* Investigation of the dependence of the supervised classification quality on the choice of partial orders on feature values sets // Intellegent Information Processing IIP 13, 2020. Pp. 21–25.

О корреляции риска с оценкой скользящего экзамена

Неделько Виктор Михайлович nedelko@math.nsc.ru
Институт математики им. С. Л. Соболева

Метод скользящего экзамена является основным инструментов для оценки качества построенного решения. Вместе с тем точность этого метода в общем случае неизвестна.

В данной работе исследуется погрешность скользящего экзамена как оценки риска, т.е. ожидаемых средних потерь. Установлено, что оценка скользящего экзамена во многих случаях имеет отрицательную корреляцию с оцениваемой величиной (риском). Отрицательная корреляция является дополнительным (помимо высокой дисперсии) фактором, обуславливающим меньшую точность скользящего экзамена относительно оценок по (отложенной) контрольной выборке такого же размера.

Обозначим риск как

$$\mathcal{K}_Q(V_N) = \mathsf{E}_{X,Y} L(\lambda_{Q,V_N}(x), y).$$

Здесь Q – метод построения решающих функций, V_N – выборка объёма N, L – функция потерь. Математическое ожидание берётся по пространству переменных, характеризующих объекты.

Пусть $\widetilde{\mathcal{K}}_{Q,K}(V_N)$– оценка риска методом скользящего экзамена (K-fold cross-validation).

В качестве меры погрешности оценки используем средний квадрат отклонения

$$err_{CV} = \sqrt{\mathsf{E}_{W_N}(\Delta(V_N))^2},$$

где

$$\Delta(V_N) = \widetilde{\mathcal{K}}_{Q,K}(V_N) - \mathcal{K}_Q(V_N).$$

Математическое ожидание берётся по всем выборкам размера N.

Разложим погрешность

$$\mathsf{E}_{W_N}(\Delta(V_N))^2 = \tilde{\sigma}^2 + \sigma^2 + bias_{CV}^2 - 2\tilde{\sigma}\sigma\varkappa,$$

где

$$\tilde{\sigma}^2 = \mathsf{D}_{W_N}\widetilde{\mathcal{K}}_{Q,K}(V_N), \quad \sigma^2 = \mathsf{D}_{W_N}\mathcal{K}_Q(V_N),$$

$$bias_{CV} = \mathsf{E}_{W_N}\Delta(V_N), \quad \varkappa = \mathrm{corr}_{W_N}(\widetilde{\mathcal{K}}_{Q,K}(V_N), \mathcal{K}_Q(V_N)).$$

Здесь corr_{W_N} – коэффициент корреляции, а $bias_{CV}$ – смещение.

В таблице 1 приведена часть результатов проведённых экспериментов на реальных данных (репозиторий UCI, задача Adult) и на синтетической модели. Использовались различные методы классификации, в частности градиентный

Таблица 1. Результаты экспериментов

	$bias_{CV}$	σ	$\tilde{\sigma}$	$\varkappa$
Boosting, UCI	0,109	0,071	0,133	-0,460
Boosting, model	0,013	0,016	0,054	-0,194
QDA, model	0,010	0,018	0,054	-0,735

бустинг на деревьях и квадратичный дискриминант. Приведены результаты для логарифмической функции потерь.

Как можно заметить, смещение оценки скользящего экзамена пренебрежимо по сравнению со стандартным отклонением, которое, в свою очередь, значительно больше стандартного отклонения риска (фактических потерь). Этот факт согласуется с известными результатами.

При этом, следует обратить внимание на неожиданный факт, а именно, отрицательную корреляцию между оценкой и оцениваемой величиной. Абсолютное значение коэффициента корреляции в экспериментах было различным, но знак всегда отрицательный. Данный факт пока не удалось объяснить, тем не менее, его важно учитывать. Оказывается, что скользящий экзамен имеет относительно высокую погрешность не только в силу большой дисперсии, но и из-за отрицательной корреляции.

Работа выполнена в рамках госзадания Института математики им С.Л. Соболева (проект No.0314-2019-0015)

[1] *Nedel'ko V.* On Decompositions of Decision Function Quality Measure // he Bulletin of Irkutsk State University. Series Mathematics, 2020. Vol. 33. Pp. 64–79.

On the correlation of risk with the cross-validation estimate

Nedel'ko Victor nedelko@math.nsc.ru
Sobolev Institute of Mathematics

The cross-validation method is the main tool for evaluating the quality of the constructed solution in machine learning. However, the accuracy of this method is generally unknown.

In this paper we investigate the accuracy of cross-validation as an estimate of risk. By risk we call expected average loss.

The cross-validation estimate was investigated on real data (UCI, dataset Adult) and on a synthetic data.

We used various classification methods, in particular gradient boosting on trees and quadratic discriminant.

In the experiments, the bias of the cross-validation turned out to be negligible compared to the standard deviation, which, in turn, is much greater than the standard deviation of risk (actual losses). This fact is consistent with the known results.

At the same time, we revealed an unexpected fact, namely, a negative correlation between the CV-estimate and the true error value. The absolute value of the correlation coefficient in the experiments was different, but the sign is always negative. This fact has not yet been explained, however, it is important to take it into account. It turns out that the cross-validation has a relatively high inaccuracy, not only because of the large variance, but also because of the negative correlation.

The study was carried out within the framework of the state contract of the Sobolev Institute of Mathematics (project no 0314-2019-0015).

[1] *Nedel'ko V.* On Decompositions of Decision Function Quality Measure // he Bulletin of Irkutsk State University. Series Mathematics, 2020. Vol. 33. Pp. 64–79.

О нейросетевом подходе к решению классов дифференциальных уравнений

Карандашев Яков Михайлович [1] karandashev@niisi.ras.ru
Шамин Александр Юрьевич[1]⋆ shamin_ay@mail.ru
[1]Москва, ФГУ ФНЦ НИИСИ РАН

Огромное множество процессов в физике, химии, биологии, механике моделируются системами дифференциальных уравнений как обыкновенных, так и в частных производных с соответствующими граничными и начальными условиями. В частности, системы обыкновенных дифференциальных уравнений используются в теоретической механике, химической кинетике, а уравнения в частных производных в задачах механики сплошных сред (теория упругости, гидро- аэромеханика), квантовой физике. В связи с этим разработка эффективных методов решения таких задач является актуальной для науки и техники и на сегодняшний день.

Далеко не всегда возможно аналитической исследование таких систем, в связи с чем развиты методы численного интегрирования, в том числе для так называемых жестких систем (Рунге-Кутта, Розенброка, методы конечных элементов, методы граничных элементов). В настоящее время численные методы также развиваются и совершенствуются.

Помимо классических численных подходов на сегодняшний появились новые методы моделирования физико-химических и механических процессов, основанные на нейронных сетях. Например, в [1, 2] обучается сеть для решения дифференциальных уравнений, но с использованием дополнительных численных методов в процессе ее обучения. Для механического моделирования в работе [3] успешно используются графовые нейросети. В [4, 5, 6] предлагаются принципы нейросетевого моделирования, основанного на обучении сети с использованием дифференцирования выхода сети по ее входам.

В настоящей работе развивается подход, предложенный в [7, 8], основанный на аппроксимации нейросетью решения конкретной задачи Коши, причем отличие от указанных работ заключается в том, что сеть обучается не одной конкретной начально-краевой задаче, а параметрическому классу задач. Это позволяет решать задачу с различными начальными и краевыми условиями, а также решать параметрически заданное уравнение, подавая на вход нейросети параметры, которые задают эти условия, из некоторого диапазона.

Пусть имеется уравнение $F_\beta(f^{(n)}(x), ..., f(x)) = 0, \quad x \in (a, b)$ с начальными условиями $q_i(a) = f^{(n)}(a),\ i = 0...n$. Будем искать аппроксимацию $\widetilde{f}$ решения этой задачи в виде нейросети $\widetilde{f}(x, q_0(a), ..., q_n(a), \beta)$, на вход которой подаются точка x, начальные условия $q_i(a)$, а также параметр β, отвечающий за параметризацию левой части уравнения F_β.

В качестве функции невязки предлагается использовать сумму двух (трех – для уравнений в частных производных) функций ошибки $L = L_1 + L_2$ Первая

функция – это сумма значений левой части самого дифференциального уравнения

$$L_1 = \sum_{i=1}^{N_{int}} \left(F_\beta(\widetilde{f}^{(n)}(x), ..., \widetilde{f}'(x), \widetilde{f}(x)) \right)^2$$

в конечном числе точек области. Вторая – сумма значений функции, задающих начальное условие $L_2 = \sum_{i=1}^{n} \left(\widetilde{f}^{(i)}(a) - q_i(a) \right)^2$

Третья – сумма значений функции, задающей граничное условие в некотором числе точек этой границы (в случае уравнений в частных производных). Простота применения такого подхода обусловлена тем, что взятие необходимых производных происходит посредством автоматического дифференцирования [9] и вычисляется точно, а аппроксимация решения происходит в процессе обучения сети.

Предложенная сеть реализована с использованием фреймворка Pytorch. Модуль Autograd позволяет легко работать с градиентами и, в частности, получать производные сети в том числе по ее входам.

Приведены результаты работы нейросети на некоторых задачах. Для примера приведена задача о линейном осцилляторе с диссипацией.

$$\begin{cases} y''(x) + 2\beta_1 y'(x) + \beta_2 y(x) = 0 & x \in [0;\, 4], \quad \beta_1 \in [1;\, 5] \\ y(0) = \alpha, y'(0) = \gamma & \alpha, \gamma \in [-5;\, 5], \quad \beta_2 \in [\beta_1^2 + \varepsilon;\, 30] \end{cases}$$

В результате сеть была обучена решениям задачи, абсолютные ошибки указаны в таблице.

$(\alpha,\, \gamma)$	$(1,\, 1)$	$(4,\, -2)$	$(-3,\, -1)$	$(0,\, 4)$
$\beta_1 = 1, \beta_2 = 5$	0.02	0.06	0.06	0.04
$\beta_1 = 2, \beta_2 = 9$	0.004	0.008	0.003	0.002
$\beta_1 = 3, \beta_2 = 10$	0.006	0.001	0.005	0.01

Таким образом, предложено обобщение нейросетевого подхода к решению ДУ, с использованием автоматического дифференцирования на классы задач с параметрически заданными уравнениями и начальными и граничными условиями, что позволяет решать каждую новую задачу, не обучая сеть заново.

Работа выполнена в рамках государственного задания по проведению фундаментальных научных исследований по теме "Исследование нейроморфных систем обработки больших данных и технологии их изготовления"№ 0065-2019-0003 (АААА-А19-119011590090-2).

[1] *Ricky T., Chen Y., Bettencourt J., Duvenaud D.* Neural Ordinary Differential Equations // arXiv:1806.07366, 2018.

[2] *Holl P., Koltun V., Thuerey N.* Learning to Control PDEs with Dif-ferentiable Physics // arXiv:2001.07457, 2020.

[3] *Pfaff T., Fortunato M., Sanchez-Gonzalez A., Battaglia P.* Learning Mesh-Based Simulation with Graph Networks // arXiv:2010.03409, 2020.

[4] *Васильев А. Н., Тархов Д. А.* Принципы и техника нейросетевого моделирования // СПб.: Нестор-История, 2014. 217 с.

[5] *Tarkov D., Vasiliev A.* Semi-empirical Neural Network Modeling and Digital Twins Development // Academic Press, Elsevier, 2019

[6] *Брусов В. С., Тюменцев Ю. В.* Нейросетевое моделирование движения летательных аппаратов // Москва: Изд-во МАИ, 2016, 192 с.

[7] *Maziar R., Paris P., George K.* Physics Informed Deep Learning (Part I): Data-driven Solutions of Nonlinear Partial Differential Equations // arXiv:1711.10561, 2017.

[8] *Maziar R.* Deep Hidden Physics Models: Deep Learning of Nonlinear Partial Differential Equations // arXiv:1801.06637, 2018.

[9] *Baydin A., Barak P., Radul A., Jeffrey S.* Automatic differentiation in machine learning: A survey // Journal of Machine Learning Research, 2019. Vol. 18. Pp. 1–43.

On a neural network approach to solving classes of differential equations

Karandashev Iakov[1], karandashev@niisi.ras.ru
Shamin Alexander[1]⋆ shamin_ay@mail.ru
[1]Moscow, Federal Scientific Center Scientific Research Institute for System Research of the Russian Academy of Sciences

A lot of processes in physics, chemistry, biology, mechanics are modeled by systems of differential equations, both ordinary and partial derivatives with appropriate boundary and initial conditions. In particular, systems of ordinary differential equations are used in theoretical mechanics, chemical kinetics, and partial differential equations in problems of continuum mechanics (theory of elasticity, hydro-aeromechanics), quantum physics. In this regard, the development of effective methods for solving such problems is relevant for science and technology today.

An analytical study of such systems is far from always possible, and therefore methods of numerical integration have been developed, including for the so-called stiff systems (Runge-Kutta, Rosenbrock, finite element methods, boundary element methods). At present, numerical methods are also being developed and improved.

In addition to classical numerical approaches, new methods for modeling physicochemical and mechanical processes based on neural networks have appeared today. For example, in [1, 2] a network is trained to solve differential equations, but using additional numerical methods in the process of training it. For mechanical modeling in work [3] graph neural networks are successfully used. In [4, 5, 6], the principles of neural network modeling based on network training using the differentiation of the network output by its inputs are proposed.

In this paper, we develop the approach proposed in [7, 8], based on the neural network approximation of the solution of a specific Cauchy problem, and the difference from these works is that the network is trained not for one specific initial-boundary value problem, but for a parametric class of problems. This makes it possible to solve a problem with different initial and boundary conditions, as well as to solve a parametrically given equation by supplying to the input of the neural network the parameters that specify these conditions from a certain range.

Let there be an equation $F_\beta(f^{(n)}(x), ..., f(x)) = 0, \quad x \in (a, b)$ with the initial conditions $q_i(a) = f^{(n)}(a),\ i = 0...n$. We will look for an approximation $\widetilde{f}$ of the solution to this problem in the form of a neural network $\widetilde{f}(x, q_0(a), ..., q_n(a), \beta)$, the input of which is the point x, the initial conditions $q_i(a)$, as well as the parameter β, which is responsible for the parametrization of the left-hand side of the equation F_β.

As a residual function, it is proposed to use the sum of two (three - for partial differential equations) error functions $L = L_1 + L_2$ The first function is the sum of

the values of the left side of the differential equation

$$L_1 = \sum_{i=1}^{N_{int}} \left(F_\beta(\widetilde{f}^{(n)}(x), ..., \widetilde{f}'(x), \widetilde{f}(x)) \right)^2$$

in a finite number of points of the region. The second is the sum of the values of the function that specify the initial condition $L_2 = \sum_{i=1}^{n} \left(\widetilde{f}^{(i)}(a) - q_i(a) \right)^2$

The third is the sum of the values of the function specifying the boundary condition at a certain number of points of this boundary (in the case of partial differential equations). The simplicity of this approach is due to the fact that the necessary derivatives are taken by means of automatic differentiation [9] and are calculated exactly, and the solution is approximated in the process of training the network.

The proposed network is implemented using the Pytorch framework. The Autograd module makes it easy to work with gradients and, in particular, to get derived networks, including those from its inputs.

The results of the neural network operation on some tasks are presented. As an example, the problem of a linear oscillator with dissipation is given

$$\begin{cases} y''(x) + 2\beta_1 y'(x) + \beta_2 y(x) = 0 & x \in [0;\, 4], \quad \beta_1 \in [1;\, 5] \\ y(0) = \alpha, y'(0) = \gamma & \alpha, \gamma \in [-5;\, 5], \quad \beta_2 \in [\beta_1^2 + \varepsilon;\, 30] \end{cases}$$

As a result, the network was trained to solve the problem, the absolute errors are indicated in the table.

$(\alpha,\, \gamma)$	$(1,\, 1)$	$(4,\, -2)$	$(-3,\, -1)$	$(0,\, 4)$
$\beta_1 = 1, \beta_2 = 5$	0.02	0.06	0.06	0.04
$\beta_1 = 2, \beta_2 = 9$	0.004	0.008	0.003	0.002
$\beta_1 = 3, \beta_2 = 10$	0.006	0.001	0.005	0.01

Thus, a generalization of the neural network approach to solving differential equations is proposed, using automatic differentiation into classes of problems with parametrically specified equations and initial and boundary conditions, which makes it possible to solve each new problem without retraining the network.

Funding. The work financially supported by State Program of SRISA RAS No. 0065-2019-0003 (AAAA-A19-119011590090-2).

[1] *Ricky T., Chen Y., Bettencourt J., Duvenaud D.* Neural Ordinary Differential Equations // arXiv:1806.07366, 2018.

[2] *Holl P., Koltun V., Thuerey N.* Learning to Control PDEs with Dif-ferentiable Physics // arXiv:2001.07457, 2020.

[3] *Pfaff T., Fortunato M., Sanchez-Gonzalez A., Battaglia P.* Learning Mesh-Based Simulation with Graph Networks // arXiv:2010.03409, 2020.

[4] *Vasiliev A., Tarkhov D.* Principles and techniques of neural network modeling // SPb .: Nestor-History, 2014. 217 p.

[5] *Tarkov D., Vasilyev A.* Semi-empirical Neural Network Modeling and Digital Twins Development // Academic Press, Elsevier, 2019

[6] *Brusov V., Tyumentsev Yu.* Neural network modeling of flight vehicles motion // Moscow: MAI Publishing House, 2016. 192 p.

[7] *Maziar R., Paris P., George K.* Physics Informed Deep Learning (Part I): Data-driven Solutions of Nonlinear Partial Differential Equations // arXiv:1711.10561, 2017.

[8] *Maziar R.* Deep Hidden Physics Models: Deep Learning of Nonlinear Partial Differential Equations // arXiv:1801.06637, 2018.

[9] *Baydin A., Barak P., Radul A., Jeffrey S.* Automatic differentiation in machine learning: A survey // Journal of Machine Learning Research, 2019. Vol. 18. Pp. 1–43.

Метод повышения эффективности обучения градиентного бустинга, основанный на модифицированных функциях потерь

Королев Николай Сергеевич[1]⋆ korolev.nikolay.s@gmail.com
Сенько Олег Валентинович[1,2] senkoov@mail.ru
[1]Москва, МГУ им. М.В. Ломоносова
[2]Москва, ВЦ РАН

Леса, состоящие из деревьев решений, хорошо себя зарекомендовали при решении прикладных задач. Градиентный бустинг [1] смог повысить предсказательную способность лесов. Данные алгоритмы достаточно распространены и уже решают различные задачи на практике и применяются во многих сферах человеческой деятельности.

Обозначим, $x_1, x_2, \ldots, x_N$ – точки в некотором многомерном пространстве, соответствующие известным и легко-измеряемым признакам реальных объектов; $y_1, y_2, \ldots, y_N$ – значения некоторых трудно-измеряемых признаков объектов. Встаёт задача поиска некоторой функции $f(x)$ такой, что $y_i = f(x_i) + \varepsilon_i$, где ε_i – ошибка предсказания на i-том объекте, т.е. функция $f(x)$ должна приближать реальную зависимость между искомыми значениями y_i и известными признаками x_i. Для построения функции $f(x)$ используется информация лишь о некоторых $T < N$ объектов, а качество приближения проверяется по оставшимся $N - T$ объектам.

Градиентный бустинг [1] основан на итеративном построении функции $f(x)$ за счёт использования большого количества деревьев решений, каждое из которых исправляет ошибки предыдущих. При этом для решения задачи задаётся оптимизируемый функционал ошибки. Одним из стандартных оптимизируемых функционалов является среднеквадратичная ошибка $L(f(x), X, Y) = \frac{1}{T}\sum_{i=1}^{T}(f(x_i) - y_i)^2$.

Предлагается использовать модифицированные функционалы ошибки.

Будем использовать в качестве $L(f(x), X, Y) = \frac{1}{T}\sum_{i=1}^{T}(\alpha f(x_i) - y_i)^2$, где $\alpha \in \mathbb{R}_{++}$. Основная идея данного подхода заключается в том, чтобы добавить шума в обучаемые деревья решений, для того чтобы уменьшить корреляцию между выходами различных деревьев решений в итоговом лесе, что позволяет увеличить обобщающую способность обучаемой модели [2].

Как отмечалось ранее, для повышения обобщающей способности леса деревьев решений необходимо уменьшать корреляцию между деревьями решений, поэтому разумно на каждом шаге оптимизировать функцию $\frac{1}{T}\sum_{i=1}^{T}\left[(h(x_i) + f(x_i) - y_i)^2 - \gamma(h(x_i) - f(x_i))^2\right]$. В данной функции есть дополнительная добавка $-\gamma(h(x_i) - f(x_i))^2$, позволяющая добиться дополнительной

регуляризации за счёт различия между новым обучаемым деревом и уже обученным лесом решающих деревьев.

Отметим, что в такой ситуации оказывается, что использование данной функции потерь абсолютно эквивалентно использованию смещённой среднеквадратичной ошибки с параметром $\alpha = 1 + \gamma$ (при дополнительном шкалировании learning-rate'а градиентного бустинга). Кроме того, логично использовать только лишь $0 \leqslant \gamma \leqslant 1$, т.к. в случае $\gamma < 0$ будет поощряться похожесть откликов нового дерева решений на отклики всего ансамбля, в то время как мы хотели уменьшить корреляцию между ними, а в случае $\gamma > 1$ функция $\frac{1}{T}\sum\limits_{i=1}^{T}\left[(h(x_i) + f(x_i) - y_i)^2 - \gamma(h(x_i) - f(x_i))^2\right]$ будет иметь минимум в точках $h(x_i) = \pm\infty$. В соответствии с границами изменения γ, а также выведенной зависимостью $\alpha = 1+\gamma$ получаем, что имеет смысл рассматривать лишь $\alpha \in [1; 2]$.

Для проверки качества работы представленного метода будем решать различные задачи классификации и регрессии используя обычный градиентный бустинг, сравнивая результаты работы с градиентным бустингом с использованием смещённой квадратичной ошибки. Кроме того, проводились вычислительные эксперименты с использованием среднеквадратичной ошибки с удалением от обученного ансамбля, но их результаты полностью совпадают с использованием обычной смещённой квадратичной ошибки, что соответствует теории.

Наиболее хороших результатов на различных задачах удалось достичь для $\alpha = 1.1$. Использование модифицированной функции потерь позволило улучшить предсказательную способность градиентного бустинга, как на задачах классификации, так и на задачах регрессии.

Итоговые результаты экспериментов представлены в таблице 1.

Набор данных	Ср-кв. ошибка	Смещ. ср-кв. ошибка	α
Аритмия	0.89	**0.90**	1.7
Ледники	0.72	**0.75**	1.1
Продажи	0.21	**0.26**	1.1
Сист. давл.	0.41	**0.46**	1.1

Таблица 1. Целевая метрика на тестовой выборке для лесов, обученных стандартной процедурой градиентного бустинга (столбец «Ср-кв. ошибка») и с использованием смещённой среднеквадратичной ошибки (столбец «Смещ. ср-кв. ошибка»)

[1] *Jerome H.* Stochastic Gradient Boosting // Computational Statistics & Data Analysis, 2002. — Pp. 367–378.

[2] *Докукин А., Сенько О.* Оптимальные выпуклые корректирующие процедуры в задачах высокой размерности // Ж. вычисл. матем. и матем. физ. 2011. С. 1751–1760.

Method for improving generalization performance of gradient boosting

Korolev Nikolai[1,2] korolev.nikolay.s@gmail.com
Senko Oleg[2]⋆ senkoov@mail.ru
[1]Moscow, Lomonosov Moscow State University
[2]Moscow, CC RAS

Decision Tree Forests have worked well for applied problems. Gradient boosting [1] has been able to increase the predictive power of forests. These algorithms are quite widespread and already used in many areas of human activity to solve various problems in practice.

Denote $x_1, x_2, \ldots, x_N$ – points in some multidimensional space corresponding to known and easily measurable features of real objects; $y_1, y_2, \ldots, y_N$ – values of some difficult-to-measure features of objects. The task is to find function $f(x)$ such as $y_i = f(x_i) + \varepsilon_i$, where ε_i – prediction error on ith object, i.e. the function $f(x)$ should approximate the real relationship between the required values of y_i and the known features of x_i. To construct the function $f(x)$ only information about some $T < N$ objects is used, and the quality of the approximation is checked by the remaining $N - T$ objects.

Gradient boosting [1] is based on iterative construction of the $f(x)$ by using a large number of decision trees, each of which fixes the mistakes of the previous ones. In this case, to solve the problem, an optimized loss function is set. One of the standard optimized loss functions is mean squared error $L(f(x), X, Y) = \frac{1}{T}\sum_{i=1}^{T}(f(x_i) - y_i)^2$.

It is proposed to use modified loss functions.

We'll use as loss function $L(f(x), X, Y) = \frac{1}{T}\sum_{i=1}^{T}(\alpha f(x_i) - y_i)^2$, where $\alpha \in \mathbb{R}_{++}$. The main idea of this approach is to add noise to the trained decision trees in order to reduce the correlation between the outputs of different decision trees in the final forest, which allows to increase the generalization performance of the trained model. [2].

As noted earlier, in order to increase the generalizing ability of a forest of decision trees, it is necessary to reduce the correlation between decision trees, therefore it is reasonable at each step to optimize the function $\frac{1}{T}\sum_{i=1}^{T}\left[(h(x_i) + f(x_i) - y_i)^2 - \gamma(h(x_i) - f(x_i))^2\right]$. This function has an additional additive $-\gamma(h(x_i) - f(x_i))^2$, allowing for additional regularization by increasing diversity between the new trainable tree and the already trained forest of decision trees.

Note that in such a situation it turns out that the use of this loss function is absolutely equivalent to the use of biased mean square error with the parameter $\alpha = 1 + \gamma$ (with additional scaling of the gradient boosting learning rate). Besides, it is logical to use only $0 \leqslant \gamma \leqslant 1$, because in the case of $\gamma < 0$, the similarity of responses of new decision tree to the responses of the entire ensemble will be encouraged, while we wanted to reduce the correlation between them, and in the case of $\gamma > 1$ function $\frac{1}{T}\sum_{i=1}^{T}\left[(h(x_i) + f(x_i) - y_i)^2 - \gamma(h(x_i) - f(x_i))^2\right]$ will have a minimum in points $h(x_i) = \pm\infty$. In accordance with the limits of γ variation, as well as the derived dependence $\alpha = 1 + \gamma$, we obtain that it makes sense to consider only $\alpha \in [1; 2]$.

To check the quality of the presented method, we will solve various classification and regression problems using conventional gradient boosting, comparing the results of work with gradient boosting using a biased mean squared error. In addition, computational experiments were carried out using the mean squared error with distance from the trained ensemble, but the results completely coincide using the biased mean squared error, which is in line with a theory.

The best results on various problems were achieved for $\alpha = 1.1$. The use of the modified loss function made it possible to improve the predictive ability of gradient boosting, both for classification problems and for regression problems.

The final results of the experiments are presented in the table 1.

Dataset	MSE	Biased MSE	α
Arrhythmia	0.89	**0.90**	1.7
Glaciers	0.72	**0.75**	1.1
Sales	0.21	**0.26**	1.1
Syst. pressure	0.41	**0.46**	1.1

Table 1. Target metric on a test set for forests trained by the standard gradient boosting procedure (column "MSE") and with usage of biased mean squared error (column "Biased MSE")

[1] *Jerome H.* Stochastic Gradient Boosting // Computational Statistics & Data Analysis, 2002. Pp. 367–378.

[2] *Dokukin A., Senko O.* Optimal convex correcting procedures in high dimensional problems // J. Comp. Math. and Math. Phys., 2011. Pp. 1751–1760.

Визуализация многомерных данных на основе построения кратчайшего незамкнутого пути

Сурков Егор Эдуардович[1]⋆ eg-su@mail.ru
Середин Олег Сергеевич[1] oseredin@yandex.ru
Копылов Андрей Валериевич[1] and.kopulov@gmail.com
Двоенко Сергей Данилович[1] sergedv@yandex.ru
[1]Тула, Тульский государственный университет

В простейшем случае визуализация данных предполагает изображение зависимости определенной функции от одного или нескольких параметров. Такая задача тривиальна для двухмерного и трехмерного случаев, в которых достаточно построить график соответствующей размерности. Однако в реальных задачах объект описывается не одной парой признаков – для описания объектов окружающего мира используются десятки и сотни влияющих на них факторов. Визуализировать зависимость при большом количестве описательных признаков довольно-таки проблематично. Таким образом, при выборе количества характеристик объекта учитываются как «простота» и наглядность визуализации, так и достоверность и точность исследования [1]. Основной идеей визуализации многомерных данных в работе является визуализация на основе поиска кратчайшего незамкнутого пути между объектами исследуемой выборки и его отображением на двумерную плоскость в виде незамкнутого графа (цепочки), столбчатой диаграммы распределения объектов вдоль найденного пути или проекции на путь. В работе предлагается к рассмотрению несколько критериев поиска кратчайшего незамкнутого пути.

1. Первый критерий заключается в поиске кратчайшего незамкнутого пути и математически выражается как минимизация следующего функционала:

$$J_1 = min \sum_{i=2}^{N} d_{i,i-1}$$

где $d_{i,j}$ - расстояние между i-й и j-й точкой, $i = 1...N$ – порядок обхода элементов в пути, N - количество элементов.

2. Второй критерий эвристический и используется в том случае, когда необходимо найти путь с наименьшей длиной и с наибольшим расстоянием между терминальными точками, таким образом «выпрямить» КНП. Второй критерий обеспечивает минимум разницы между длиной КНП и расстоянием между терминальными точками пути:

$$J_2 = min\Big(\sum_{i=2}^{N} d_{i,i-1} - d_{1,N} \Big)$$

где $d_{i,j}$ - расстояние между i-й и j-й точкой, $i = 1...N$ – порядок обхода элементов в пути, N - количество элементов.

Алгоритм A0 – представляет собой полный перебор всех вариантов соединений точек и выбора кратчайшего пути, то есть такого, у которого сумма расстояний между точками при обходе является минимальной (далее алгоритм A0). Сложность данного алгоритма равна $((N-1)\cdot N!)/2$.

Алгоритм A1 – его идея заключается в том, что выбираются две точки, расстояние между которыми минимальное, далее из них строится граф из ребер так, что для каждой точки находится наиболее приближенная к ней, и так до тех пор, пока не будут соединены все N точек. Сложность данного алгоритма $(3N^2+N)/2$.

Алгоритм A2 – модификация A1, которая заключается в том, что для каждой, так называемой, затравочной точки находится ближайшая и эти две точки представляют стартовый отрезок. Далее действия повторяются по алгоритму A1. Сложность алгоритма $N\cdot(N^2+N)=N^3+N^2$.

Алгоритм A3 – модификация A1, которая заключается в том, что граф КНП строится из каждой пары точек. Сложность алгоритма $((N^2-N)/2)(N^2+N)=$ $=(N^4-N^2)/2$.

Очевидно, что решение, найденное при использовании A1, будет среди решений A2 и A3, а решение, найденное при помощи A2, входит в A3. Алгоритм A4 – модификация алгоритма A3, которая заключается в том, что изначально выбирается одна точка, затем слева и справа от нее перебираются все пары, которые становятся терминальными для этой тройки. Далее выполняется поиск, как в алгоритме A1. Сложность такого алгоритма оценивается, как $((N^3-$ $-N^2)/2)(N^2+N)=(N^5-N^3)/2$.

Также предлагаются рекурсивные варианты этих алгоритмов A1R, A2R, A3R, A4R, которые заключаются в том, что если при поиске ближайшей точки (из списка свободных) к терминальной точке будет обнаружено несколько равноудаленных точек, то рассматривается каждый вариант такого пути.

В заключительной части работы приведены результаты экспериментов на модельных данных и на реальных данных, таких как Iris Data Set [2], Abalone Data Set [3]. Также проведен эксперимент на данных из исследования задачи детектирования падений [4]. Выполнено экспериментальное сравнение алгоритмов поиска кратчайшего незамкнутого пути, а также приведены сравнительные таблицы временных затрат на их вычисление.

Работа выполнена при финансовой поддержке Министерства науки и высшего образования РФ в рамках государственного задания FEWG-2021-0012 и частично гранта РФФИ No. 20-07-00055 (С.Д. Двоенко).

[1] *Яковлев С. С., Середин О. С.* Использование деревьев решений при визуализации многомерных данных // Известия Тульского государственного университета. Технические науки, 2018. No. 10. С. 137–145.

[2] *Fisher R.* The use of multiple measurements in taxonomic problems // Annual Eugenics, 1936.

[3] *Sam W.* Extending and benchmarking Cascade-Correlation // PhD thesis, Computer Science Department, University of Tasmania, 1995.

[4] *Сурков Е.* Исследование базисной совокупности скелетных представлений в задаче детектирования падений // Ломоносов-2021: Сборник тезисов XXVIII Международной научной конференции студентов, аспирантов и молодых ученых, 2021. С. 82–83.

Multidimensional data visualization based on the shortest unclosed path search

Surkov Egor[1]⋆ eg-su@mail.ru
Seredin Oleg[1] oseredin@yandex.ru
Kopylov Andrey[1] and.kopulov@gmail.com
Dvoenko Sergey[1] sergedv@yandex.ru
[1]Tula, Tula state university

In the simplest case data visualization assumes an image of the certain dependency function of one or more parameters. This problem is trivial for two- and three-dimensional cases where it is enough to get an appropriate dimension chart. However, in real tasks an object of the surrounding world is described by more than one pair of features to describe them, it is necessary to use tens and hundreds of factors influencing them. Visualize the dependency with a large number of descriptive features is quite problematic. Thus, when selecting the number of characteristics of an object, the following are taken into account, both the "simpleness" and clearness of visualization as well as the reliability and accuracy of the study [1].

The visualization of multidimensional data core idea in the work is visualization based on search for the shortest unclosed path between (SUP) the sample objects and its mapping to a two-dimensional plane as an open graph (chain), a column chart of the distribution of objects along the found path or a projection onto the path. In this paper consider several criteria for finding the shortest open path is proposed. The first criteria is to find the shortest unclosed path and is mathematically expressed as the minimization of the following functional:

$$J_1 = min \sum_{i=2}^{N} d_{i,i-1}$$

where $d_(i,j)$ - distances between i-th and j-th point, $i = 1...N$ – the order of elements in the path passage, N - number of elements.

2. The second criteria is heuristic and used when it is necessary to find the path with the shortest length and with the greatest distance between the terminal points, thus "straighten" the shortest unclosed path. The second criteria provides a minimum of the difference between the length of the SUP and the distance between terminal points of the path:

$$J_2 = min\Big(\sum_{i=2}^{N} d_{i,i-1} - d_{1,N} \Big)$$

where $d_(i,j)$ - distances between i-th and j-th point, $i = 1...N$ – the order of elements in the path passage, N - number of elements.

In addition, the work implements the following algorithms for finding the shortest unclosed path: Algorithm A0 - is a full search of all the options for connecting points and choosing the shortest path, that is, one in which the sum of the distances between the points after bypassing is minimal (further algorithm A0). The complexity of this algorithm is equal to $((N-1)\cdot N!)/2$.

Algorithm A1 - his idea is that two points with distance between them is minimal are selected, then an edges graph is constructed from them so for each point there is a closest to it, and so on until all points are connected. The complexity of this algorithm is equal to $(3N^2+N)/2$.

Algorithm A2 – modification A1, is that for each beginning point there is the nearest one and these two points represent the starting segment. Then the actions are repeated according to the algorithm A1. The complexity of this algorithm is equal to $N\cdot(N^2+N)=N^3+N^2$.

Algorithm A3 – modification A1, is that the SUP graph is constructed from each pair of points The complexity is equal to $((N^2-N)/2)(N^2+N)=(N^4-N^2)/2$.

Obviously, A2 and A3 will include the solution found when using A1, and the solution found with A2 is included in A3.

Algorithm A4 – modification of the algorithm A3, initially one point is selected as starting point. Then all pairs are attached to the left and right of starting point, which become terminal for this triple. Next, the search is performed as in the algorithm A1. The complexity of such an algorithm is estimated as $((N^3 - N^2)/2)(N^2+N)=(N^5-N^3)/2$.

Recursive variants of these algorithms A1R, A2R, A3R, A4R are also proposed, which consist in the fact that if several equidistant points are found when searching for the nearest point (from the free list) to the terminal point, then each option of such a path is considered.

The final part of the paper presents the results of experiments on model and real data, such as iris Data Set [2], Abalone Data Set [3]. Also experiment is conducted on data from the study of the fall detection task[4]. An experimental comparison of algorithms for finding the shortest unclosed path is performed, and comparative tables of time spent on their calculation are also presented.

This research was financially supported by Ministry of Science and Higher Education of the Russian Federation within the framework of the state task FEWG-2021-0012 and partially supported by RFBR No. 20-07-00055 (Dvoenko S.).

[1] *Yakovlev S., Seredin O.* Multidimensional data visualization based on decision trees // Tidings of the Tula State University. Technical Science, 2018. No. 10. Pp. 137–145.

[2] *Fisher R.* The use of multiple measurements in taxonomic problems // Annual Eugenics, 1936.

[3] *Sam W.* Extending and benchmarking Cascade-Correlation // PhD thesis, Computer Science Department, University of Tasmania, 1995.

[4] *Surkov E.* The study about basic assembly of skeletal representations in the fall detection task // Lomonosov-2021: Collection of abstracts XXVIII International scientific conference of students, postgraduates and young scientists, 2021. Pp. 82–83.

Применение методов Монте-Карло в задачах анализа временных рядов с мультиколлинеарностью

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

Настоящая работа является продолжением исследований, опубликованных в [1], где была предложена методология верификации закономерностей, а также сравнения применимости моделей для описания наборов многомерных временных рядов методами Монте-Карло. Методология является обобщением подхода, используемого в ряде работ, посвященных перестановочным тестам на задачи регрессии и кластеризации для временных рядов. Она предполагает сравнение определенных характеристик эмпирических данных (например, коэффициента детерминации R^2) с квантилями соответствующих характеристик наборов искусственно сгенерированных с использованием датчика случайных чисел временных рядов (стационарных, или нестационарных), для которых выполнена нулевая гипотеза об отсутствии проверяемой закономерности. При данном подходе, p-значения, используемые для верификации закономерностей — это доля превышений значений характеристик для искусственных выборок над значением для конкретного набора данных. Например, если эта доля — 5%, говорят о 95%-ном квантиле с $p = 0.05$.

Рассматриваемый подход позволяет учесть возможность эффекта ложной регрессии при анализе временных рядов, которые являются нестационарными в смысле наличия «единичного корня». Нестационарность такого типа порождается процессами типа случайного блуждания и не связана с существованием временных трендов или сезонных колебаний. Некорректность применения стандартных методов верификации для временных рядов, нестационарных в указанном смысле, отмечалась многими исследователями. Она связана с тем, что такие временные ряды нельзя рассматривать как совокупность независимых наблюдений. Нестационарность с единичными корнями вследствие этого часто приводит к появлению ложных регрессий. Регрессию принято называть ложной, если она, будучи формально значимой при использовании стандартных средств верификации, по сути является бессмысленной и разрушается при выходе из временного интервала, по которому она была найдена. Для корректной верификации регрессионных моделей для нестационарных временных рядов ранее был предложен метод коинтеграции, который основан на требовании не только статистической значимости моделей, но и стационарности остатков моделей. В настоящее время теория коинтеграции является стандартным подходом. Однако, реально она может быть применена только для временных рядов достаточно большой длины, не менее 60 наблюдений. Связано это с низкой мощностью теста Дики-Фуллера. Цель нашей работы — создание и применение методов для

верификации закономерностей для относительно коротких временных рядов, в которых менее 20 наблюдений.

Если временные ряды порождаются нестационарными случайными процессами, это может приводить не только к появлению ложной регрессионной зависимости, но также к ложной значимости других эмпирических закономерностей и эффектов. В зависимости от задачи можно говорить о ложной кластеризации, ложном выходе параметров модели за пределы ранее установленных ограничений, ложном различии между моделями и т. п. Разрабатываемая нами методология позволяет исследовать подобные эффекты в рамках одного общего подхода.

Подробнее рассмотрим явление корреляции между регрессорами, обуславливающее эффект мультиколлинеарности. Она существенно выражена, например, между экономическими показателями, характеризующими производство и социально- экономические институты в макроэкономических и мезоэкономических системах. В нашей работе используются варианты нулевых гипотез, в которых регрессоры не только независимы друг от друга, но и коррелируют друг с другом. Обозначим через $H_0(\rho)$ нулевую гипотезу, предполагающую независимость целевой переменной от пары регрессоров при условии, что коэффициент корреляции между регрессорами равен ρ. Интуитивно может показаться, распределение R^2 для регрессионной модели при справедливости $H_0(\rho)$ должно плавно зависеть от величины ρ при изменении последнего от 0 до 1. При $\rho = 0$ нулевая гипотеза $H_0(\rho)$ очевидно предполагает, что регрессоры являются взаимно независимыми, а при $\rho = 1$ нулевая гипотеза предполагает переход от двухфакторной модели к однофакторной. Однако, наши численные эксперименты показали, что для регрессоров, коррелирующих друг с другом в диапазоне $(0; 0.9999...)$ распределение и соответственно квантили множественных R^2 остаются такими же, как и для случая полностью ортогональных по отношению друг к другу регрессоров. Получено теоретическое обоснование данного факта. Независимость множественного R^2 от ρ избавляет от необходимости специально учитывать мультиколлинеарность в симулированных наборах данных, если предметом исследования является только множественный R^2 зависимости целевой переменной от регрессоров. При этом, корреляция между регрессорами влияет на точность оценки вклада каждого конкретного регрессора в модель.

[1] *Кирилюк И. Л., Сенько О. В.* Выбор моделей оптимальной сложности методами Монте-Карло (на примере моделей производственных функций регионов Российской Федерации) // Информатика и ее применения, 2020. Т. 14(2). С. 111–118.

Application of Monte Carlo methods in the tasks of analysis of time series with multicollinearity

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

This work is a continuation of the research published in [1], where a methodology for the verification of patterns was proposed, as well as for a comparison of the applicability of models for describing sets of multivariate time series by Monte Carlo methods. The methodology is a generalization of the approach used in a number of works devoted to permutation tests for regression and clustering problems for time series. It involves comparing certain characteristics of empirical data (for example, the coefficient of determination R^2) with the quantiles of the corresponding characteristics of the sets of time series (stationary or non-stationary) artificially generated using a random number generator, for which the null hypothesis of the absence of a testable pattern is fulfilled. With this approach, the p-value used to verify the patterns is the fraction of the excess of the characteristic values for artificial samples over the value for a particular dataset. For example, if this fraction is 5%, one speaks of a 95% quantile with $p = 0.05$.

The approach under consideration allows us to take into account the possibility of a spurious regression effect when analyzing time series that are nonstationary in the sense of having a "unit root". Nonstationarity of this type is generated by processes such as a random walk and is not associated with the existence of time trends or seasonal fluctuations. The incorrect application of standard verification methods for time series that are nonstationary in this sense has been noted by many researchers. It is connected with the fact that such time series cannot be considered as a set of independent observations. As a consequence, nonstationarity with unit roots often leads to the appearance of spurious regressions. It is customary to call a regression spurious if, being formally significant when using standard verification tools, it is essentially meaningless and is destroyed when it leaves the time interval for which it was found. For the correct verification of regression models for nonstationary time series, a cointegration method was previously proposed, which is based on the requirement not only of the statistical significance of the models, but also of the stationarity of the residuals of the models. Cointegration theory is now the standard approach. However, in reality it can be applied only for time series of sufficiently long length, at least 60 observations. This is due to the low power of the Dickey-Fuller test. The purpose of our work is to create and apply methods for verifying patterns for relatively short time series, in which there are less than 20 observations.

If time series are generated by non-stationary random processes, this can lead not only to the appearance of a spurious regression dependence, but also to the

spurious significance of other empirical patterns and effects. Depending on the task, one can talk about spurious clustering, spurious departure of model parameters beyond the previously established limits, spurious differences between models, etc. The methodology we develop allows us to study such effects within the framework of one general approach.

Let us consider in more detail the phenomenon of correlation between regressors, which causes the effect of multicollinearity. It is significantly expressed, for example, between economic indicators characterizing production and socio-economic institutions in macroeconomic and mesoeconomic systems. In our work, we use variants of null hypotheses, in which the regressors are not only independent of each other, but also correlate with each other. Let us denote by $H_0(\rho)$ the null hypothesis, which assumes the independence of the target variable from a pair of regressors, provided that the correlation coefficient between the regressors is ρ. It might seem intuitively that the distribution of R^2 for the regression model, if $H_0(\rho)$ is valid, should smoothly depend on the value of ρ when the latter changes from 0 to 1. For $\rho = 0$, the null hypothesis $H_0(\rho)$ obviously assumes that the regressors are mutually independent, and for $\rho = 1$ the null hypothesis assumes a transition from a two-factor model to a one-factor model. However, our numerical experiments have shown that for regressors correlating with each other in the range (0; 0.9999 ...), the distribution and, accordingly, the quantiles of multiple R^2 remain the same as for the case of completely orthogonal with respect to each other regressors. The theoretical substantiation of this fact has been obtained. The independence of the multiple R^2 from ρ eliminates the need to specifically take into account multicollinearity in simulated datasets if the subject of research is only the multiple R^2 dependence of the target variable on regressors. At the same time, the correlation between the regressors affects the accuracy of assessing the contribution of each specific regressor to the model.

[1] *Kirilyuk I., Sen'ko O.* Selection of optimal complexity models by methods of nonparametric statistics (on the example of production function models of the regions of the Russian Federation) // Informatics and applications, 2020. Vol. 14(2). Pp. 111–118.

Об одной робастной схеме градиентного бустинга

Шибзухов Заур Мухадинович[1,2]⋆ intellimath@mail.ru
[1]Москва, Московский педагогический государственный университет
[2]Москва, Московский физико-технический институт

В классической схеме градиентного бустинга решается задача минимизации эмпирического риска:

$$H^* = \arg \min_{H \in L(\mathcal{H})} \mathcal{Q}(H),$$

где $L(\mathcal{H})$ — класс линейных комбинаций базовых функций из класса $\mathcal{H}$:

$$H(x) = \sum_{j=1}^{p} \alpha_j h_j(x),$$

где $\alpha_j \in \mathbb{R}$, $h_j(x) \in \mathcal{H}$, $x \in \mathbb{R}^n$,

$$\mathcal{Q}(H) = \frac{1}{N} \sum_{k=1}^{N} \ell(H(\tilde{x}_k), \tilde{y}_k), \tag{1}$$

$\{\tilde{x}_1, \ldots, \tilde{x}_N\}$ — конечный набор входов; $\{\tilde{y}_1, \ldots, \tilde{y}_N\}$ — значения, ожидаемые на выходе; $\ell(y, \tilde{y})$ — неотрицательная дифференцируемая функция потерь.

На каждом шаге процедуры градиентного бустинга решается задача минимизации:

$$h^*, \alpha^* = \arg \min_{h,\alpha} \mathcal{Q}(h, \alpha),$$

где

$$\mathcal{Q}\big(h, \alpha\big) = \frac{1}{N} \sum_{k=1}^{N} \ell(\tilde{H}_k + \alpha h_k, \tilde{y}_k), \tag{2}$$

$\alpha \in \mathbb{R}$ и $h \in \mathcal{H}$, $\tilde{H}_k = H(\tilde{x}_k)$, $\tilde{h}_k = h(\tilde{x}_k)$.

Один из методов поиска минимума $\mathcal{Q}(h, \alpha)$ основан на применении итеративного метода *поочередной минимизации* (*alternating minimization*):

$$\begin{aligned} h^{p+1} &= \arg \min_{h \in \mathcal{H}} \mathcal{Q}(h, \alpha^p) \\ \alpha^{p+1} &= \arg \min_{\alpha} \mathcal{Q}(h^{p+1}, \alpha), \end{aligned} \tag{3}$$

вначале $\alpha^0 = 1$.

Однако, эмпирическое распределение значений

$$\{z_k = z_k(h, \alpha) = \ell(\tilde{H}_k + \alpha h_k, \tilde{y}_k) \colon k = 1, \ldots, N\}$$

может содержать выбросы из-за искажений в данных или неадекватности части данных по отношению к выбранной модели зависимости. При этом, среднее арифметическое чувствительно к выбросам.

Для преодоления влияния выбросов используется более робастная постановка задачи:

$$\mathcal{Q}_{\mathsf{M}}(h,\alpha) = \mathsf{M}\{z_1(h,\alpha),\ldots,z_N(h,\alpha)\},$$

где $\mathsf{M}\{z_1,\ldots,z_N\}$ — дифференцируемая усредняющая агрегирующая функция, устойчивая к выбросам в данных [1, 2, 3].

По построению, $\partial\mathsf{M}/\partial z_k \geqslant 0$ для всех $k = 1,\ldots,N$ и

$$\partial\mathsf{M}/\partial z_1 + \cdots + \partial\mathsf{M}/\partial z_N = 1.$$

Например, цензурированное среднее арифметическое

$$\mathsf{WM}_{\rho,\delta,\varepsilon}\{z_1,\ldots,z_N\} = \frac{1}{N}\sum_{k=1}^{N}\min(z_k,\bar{z}_{\rho,\delta,\varepsilon}),$$

где $0 < \delta < 1$, $\varepsilon = 0.001$, $\bar{z}_{\rho,\delta,\varepsilon}$ – пороговое значение, которое вычисляется с помощью «сглаженного» варианта δ-квантиля:

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

$$\rho_{\delta,\varepsilon}(r) = \begin{cases}(1-\delta)\rho_\varepsilon(r), & r<0\\ \delta\rho_\varepsilon(r), & r\leqslant 0,\end{cases} \qquad \rho_\varepsilon(r) = \sqrt{r^2+\varepsilon^2}-\varepsilon.$$

Необходимое условие экстремума дает систему уравнений:

$$\begin{cases} v_k = \gamma_k(h,\alpha), \quad k=1,\ldots,N\\ \sum\limits_{k=1}^{N} v_k\ell'(\tilde{H}_k+\alpha h(\tilde{x}_k),\tilde{y}_k)\nabla h(\tilde{x}_k) = 0,\\ \sum\limits_{k=1}^{N} v_k\ell'(\tilde{H}_k+\alpha h(\tilde{x}_k),\tilde{y}_k)h(\tilde{x}_k) = 0,\end{cases}$$

где

$$\gamma_k(h,\alpha) = \frac{\partial\mathsf{M}\{z_1(h,\alpha),\ldots,z_N(h,\alpha)\}}{\partial z_k}.$$

Процедура (3) для поиска h^* и α^* принимает вид процедуры *итеративного перевзвешивания* (*iterative reweighting*):

$$\begin{aligned} h^{p+1} &= \arg\min_{h\in\mathcal{H}}\sum_{k=1}^{N}\gamma_k(h^p,\alpha^p)\ell(\tilde{H}_k+\alpha^p h(\tilde{x}_k),\tilde{y}_k)\\ \alpha^{p+1} &= \arg\min_{\alpha}\sum_{k=1}^{N}\gamma_k(h^{p+1},\alpha^p)\ell(\tilde{H}_k+\alpha h_k,\tilde{y}_k).\end{aligned} \tag{4}$$

В этой итеративной схеме на каждом шаге осуществляется минимизация взвешенной суммы потерь.

На наглядных примерах демонстрируется устойчивость процедуры (4) относительно большого количества выбросов в обучающих данных.

[1] *Calvo T., Beliakov G.* Aggregation functions based on penalties // Fuzzy Sets and Systems, 2010. Vol. 161(10). Pp. 1420–1436.

[2] *Shibzukhov Z.* Resistant neural network learning via empirical risk minimization // Advances in Neural Networks – ISNN 2019, 2019. Vol. 11554. Pp. 340–350.

[3] *Shibzukhov Z., Semenov T.* Machine learning based on minimizing robust mean estimates // In Pattern Recognition. ICPR International Workshops and Challenges, 2021. Pp. 112–119.

One Robust Scheme of Gradient Boosting

Shibzukhov Zaur[1,2]⋆ intellimath@mail.ru
[1]Moscow, Moscow Pedagigical State University
[2]Moscow, Moscow Institute Physics Technologies

In the classical scheme of gradient boosting, the following problem of minimizing empirical risk is solved:

$$H^* = \arg \min_{H \in L(\mathcal{H})} \mathcal{Q}(H),$$

where $L(\mathcal{H})$ is a class of linear combinations of the basic functions from the class $\mathcal{H}$:

$$H(x) = \sum_{j=1}^{p} \alpha_j h_j(x),$$

where $\alpha_j \in \mathbb{R}$, $h_j(x) \in \mathcal{H}$, $x \in \mathbb{R}^n$,

$$\mathcal{Q}(H) = \frac{1}{N} \sum_{k=1}^{N} \ell(H(\tilde{x}_k), \tilde{y}_k), \tag{1}$$

$\{\tilde{x}_1, \ldots, \tilde{x}_N\}$ are inputs; $\{\tilde{y}_1, \ldots, \tilde{y}_N\}$ are expected outputs; $\ell(y, \tilde{y})$ is nonnegative differentiabale loss function.

At each step of the gradient boosting procedure the following minimization problem is solved:

$$h^*, \alpha^* = \arg \min_{h,\alpha} \mathcal{Q}(h, \alpha),$$

where

$$\mathcal{Q}\big(h, \alpha\big) = \frac{1}{N} \sum_{k=1}^{N} \ell(\tilde{H}_k + \alpha h_k, \tilde{y}_k), \tag{2}$$

$\alpha \in \mathbb{R}$ and $h \in \mathcal{H}$, $\tilde{H}_k = H(\tilde{x}_k)$, $\tilde{h}_k = h(\tilde{x}_k)$.

One method for finding minimum of $\mathcal{Q}(h, \alpha)$ bases on iterative method of *alternating minimization*:

$$\begin{aligned} h^{p+1} &= \arg \min_{h \in \mathcal{H}} \mathcal{Q}(h, \alpha^p) \\ \alpha^{p+1} &= \arg \min_{\alpha} \mathcal{Q}(h^{p+1}, \alpha), \end{aligned} \tag{3}$$

at the begining $\alpha^0 = 1$.

However, empirical distribution of the values

$$\{z_k = z_k(h, \alpha) = \ell(\tilde{H}_k + \alpha h_k, \tilde{y}_k) \colon k = 1, \ldots, N\}$$

may contain outliers due to distortions in the data or the inadequacy of part of the data in relation to the selected dependency model. At the same time, the arithmetic mean is sensitive to outliers.

To overcome the impact of outliers, a more robust formulation of the problem is used:

$$\mathcal{Q}_{\mathsf{M}}(h,\alpha) = \mathsf{M}\{z_1(h,\alpha),\ldots,z_N(h,\alpha)\},$$

where $\mathsf{M}\{z_1,\ldots,z_N\}$ is differentiable averaging aggregating function, which is resistant to outliers in data [1, 2, 3].

By construction, $\partial\mathsf{M}/\partial z_k \geqslant 0$ for all $k = 1,\ldots,N$ and

$$\partial\mathsf{M}/\partial z_1 + \cdots + \partial\mathsf{M}/\partial z_N = 1.$$

For example, censored arithmetic mean

$$\mathsf{WM}_{\rho,\delta,\varepsilon}\{z_1,\ldots,z_N\} = \frac{1}{N}\sum_{k=1}^{N}\min(z_k,\bar{z}_{\rho,\delta,\varepsilon}),$$

where $0 < \delta < 1$, $\varepsilon = 0.001$, $\bar{z}_{\rho,\delta,\varepsilon}$ is the threshold value that is calculated using the «moothed» version of the δ-quantile:

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

$$\rho_{\delta,\varepsilon}(r) = \begin{cases}(1-\delta)\rho_\varepsilon(r), & r<0\\ \delta\rho_\varepsilon(r), & r\leqslant 0,\end{cases} \qquad \rho_\varepsilon(r) = \sqrt{r^2+\varepsilon^2}-\varepsilon.$$

The necessary condition of the extremum gives a system of equations:

$$\begin{cases} v_k = \gamma_k(h,\alpha), \quad k=1,\ldots,N\\ \sum\limits_{k=1}^{N} v_k\ell'(\tilde{H}_k+\alpha h(\tilde{x}_k),\tilde{y}_k)\nabla h(\tilde{x}_k) = 0,\\ \sum\limits_{k=1}^{N} v_k\ell'(\tilde{H}_k+\alpha h(\tilde{x}_k),\tilde{y}_k)h(\tilde{x}_k) = 0,\end{cases}$$

where

$$\gamma_k(h,\alpha) = \frac{\partial\mathsf{M}\{z_1(h,\alpha),\ldots,z_N(h,\alpha)\}}{\partial z_k}.$$

Procedure (3) for finding h^* and α^* takes the form of an *iterative reweighting* procedure:

$$\begin{aligned} h^{p+1} &= \arg\min_{h\in\mathcal{H}}\sum_{k=1}^{N}\gamma_k(h^p,\alpha^p)\ell(\tilde{H}_k+\alpha^p h(\tilde{x}_k),\tilde{y}_k)\\ \alpha^{p+1} &= \arg\min_{\alpha}\sum_{k=1}^{N}\gamma_k(h^{p+1},\alpha^p)\ell(\tilde{H}_k+\alpha h_k,\tilde{y}_k).\end{aligned} \tag{4}$$

In this iterative scheme, the weighted sum of losses is minimized at each step.

The stability of the procedure is demonstrated by illustrative examples (4) relatively large number of outliers in the training data.

[1] *Calvo T., Beliakov G.* Aggregation functions based on penalties // Fuzzy Sets and Systems, 2010. Vol. 161(10). Pp. 1420–1436.

[2] *Shibzukhov Z.* Resistant neural network learning via empirical risk minimization // Advances in Neural Networks – ISNN 2019, 2019. Vol. 11554. Pp. 340–350.

[3] *Shibzukhov Z., Semenov T.* Machine learning based on minimizing robust mean estimates // In Pattern Recognition. ICPR International Workshops and Challenges, 2021. Pp. 112–119.

Морфологическая теория простоты

Визильтер Юрий Валентинович[1]⋆ viz@gosniias.ru
[1]Москва, ФГУП ГосНИИАС

В рамках морфологии Пытьева [1] мозаичные изображения имеют вид

$$f(x,y)=\sum_{i=1}^{n} f_{F_i}\chi_{F_i}(x,y), \tag{1}$$

где n – число областей мозаичного разбиения $\mathbf{F}$ кадра Ω площади S на непересекающиеся области $\mathbf{F}=\{F1,\ldots,Fn\}$. $\mathbf{F}$ также называют формой F для f.

Сравнение форм по сложности задается отношением частичного порядка «не сложнее по форме». Для любых форм F и G можно указать более сложную форму $F\wedge G$ и менее сложную $F\vee G$. Известна мера сложности форм $\mu_H(F)$, согласованная с этим частичным порядком (см. [2]).

Форму F также описывают реляционной моделью (отношением сходства):

$$\begin{aligned}\eta_F(x,y,u,v)&=\sum_{i=1,\ldots,n}\chi_{F_i}(x,y)\chi_{F_i}(u,v)=\\&=\begin{cases}1,\text{если }\forall i:\chi_{F_i}(x,y)=\chi_{F_i}(u,v);\\0,\text{в противном случае.}\end{cases}\end{aligned} \tag{2}$$

Рассмотрим задачу построения вероятностной меры на множестве форм.

Определим простоту $q(F)$ формы F, как L^1-норму модели η_F при $S=1$:

$$q(F)=1-\mu_H(F)=\|\eta_F(x,y,u,v)\|_1=\sum_{(x,y,u,v)\in\Omega\times\Omega}\eta_F(x,y,u,v). \tag{3}$$

Рассмотрим множество мозаичных форм как вложение в 2^{Θ}, где $\Theta=\Omega\times\Omega$.

Определим операции объединения ($\cup$) и пересечения ($\cap$) двух форм $F,G\in\Theta$:

$$\eta_{F\cap G}(x,y,u,v)=min(\eta_F(x,y,u,v),\eta_G(x,y,u,v))$$

$$\eta_{F\cup G}(x,y,u,v)=max(\eta_F(x,y,u,v),\eta_G(x,y,u,v))$$

Простота $q(F)\in[0,1]$ есть вероятностная мера на $(\Theta,\cup,\cap)$, поскольку:

$q(O)=1$ для простейшей формы $O(\eta_O(x,y,u,v)\equiv 1)$;

$q(\varnothing)=0$ для сложнейшей формы $\varnothing(\eta_\varnothing(x,y,u,v)\equiv 0)$;

$q(F\cup G)=q(F)+q(G)-q(F\cap G)$; (правило сложения простоты)

$$q(F \cup G) = q(F) + q(G) \text{ для несовместных форм } (F \cap G = \varnothing);$$

$$q(F \cap G) = q(F)q(G) \text{ для независимых форм } (q(F \cap G)/q(F) = q(G)).$$

По аналогии с теорией вероятностей, связанной с множеством событий, операциями (И, ИЛИ) и мерой «вероятность события», мы предлагаем такую конструкцию на множестве Θ реляционных форм с операциями $(\cup, \cap)$ и мерой «простота формы» называть морфологической теорией простоты.

Простота имеет смысл геометрической вероятности для случайной пары точек попасть в одну область покрытия (разбиения). Это морфологическая (внутренняя геометрическая) вероятность, связанная с морфологией (внутренней геометрией) данной конкретной формы, а не со статистическим ансамблем форм или классом изображений данной формы.

Легко показать, что все инструменты морфологического анализа мозаичных форм можно заново ввести на основе теории простоты. В частности, морфологический коэффициент корреляции форм (МККФ) имеет вид:

$$K_M^2(G, F) = \sum_{i=1,\dots,n} p_i \sum_{j=1,\dots,m} q(G_j|F_i) = \sum_{i=1,\dots,n} \sum_{j=1,\dots,m} p_{ij}^2/p_i, \tag{4}$$

где $q(G_j|F_i) = q(G_j \cap F_i)/q(F_i)$ – условная простота G_j относительно F_i. Иными словами, МККФ (4) это средняя по кадру локальная условная простота G относительно F.

Исходя из этого, введем глобальные (по Θ) аналоги МККФ – условную и апостериорную простоту для гипотезы F и наблюдения G:

$$q(G|F) = q(G \cap F)/q(F) = \sum_{i=1,\dots,n} \sum_{j=1,\dots,m} p_{ij}^2 / \sum_{k=1,\dots,n} p_k^2, \tag{5}$$

$$q(F|G) = q(F \cap G)/q(G) = \sum_{i=1,\dots,n} \sum_{j=1,\dots,m} p_{ij}^2 / \sum_{l=1,\dots,m} p_l^2, \tag{6}$$

которые связаны формулой Байеса для простоты:

$$q(F|G) = q(G|F)q(F)/q(G). \tag{7}$$

Таким образом, в данной работе предложена теория простоты для мозаичных форм, дана интерпретация МККФ как средней локальной оценки условной простоты, предложено сравнение форм на основе условных и апостериорных оценок простоты.

[1] *Пытьев Ю., Чуличков А.* Методы морфологического анализа изображений // М: Физматлит, 2010. 336 с.

[2] *Визильтер Ю. В., Желтов С. Ю., Бусурин В. И.* Современный морфологический анализ и его применение в авиационных системах технического зрения // М: Изд-во МАИ, 2020. 176 с.

The Morphological Simplicity Theory

Vizilter Yuri[1]⋆ viz@gosniias.ru

[1]Moscow, GosNIIAS

In the framework of Pyt'ev morphology [1] we consider images as piecewise-constant 2D functions

$$f(x,y) = \sum_{i=1}^{n} f_{F_i} \chi_{F_i}(x,y), \tag{1}$$

where n is a number of tessellation $\mathbf{F}$ on frame Ω with area S into connected regions of constnt intensity, $\mathbf{F} = \{F1, \ldots, Fn\}; \mathbf{f} = (f_{F1}, \ldots, f_{Fn})^T$ is an intensity value vector. Such images we call the mosaic images.

Morphological comparison of mosaic shapes by complexity is traditionally implemented in terms of a partial order relation "not more complex by shape". The set of mosaic shapes has an algebraic lattice structure: for any shapes F and G we can find the more complex shape $F \wedge G$ and less complex shape $F \vee G$. More complex shapes are obtained by region splitting, and less complex shapes are obtained by regions merging. The generalized full-order relation based on the definition of the shape complexity measure $\mu_H(F)$ is introduced in [2].

The shape F of the mosaic image corresponds to the relational shape:

$$\begin{aligned} \eta_F(x,y,u,v) &= \sum_{i=1,\ldots,n} \chi_{F_i}(x,y)\chi_{F_i}(u,v) = \\ &= \begin{cases} 1, \text{if } \forall i : \chi_{F_i}(x,y) = \chi_{F_i}(u,v); \\ 0, \text{otherwise.} \end{cases} \end{aligned} \tag{2}$$

Consider the problem of constructing a probabilistic measure on a set of shapes. Define simplicity $q(F)$ of F as L^1-norm, of η_F model with $S = 1$:

$$q(F) = 1 - \mu_H(F) = \|\eta_F(x,y,u,v)\|_1 = \sum_{(x,y,u,v)\in\Omega\times\Omega} \eta_F(x,y,u,v). \tag{3}$$

Consider the set of mosaic shapes as an embedding to 2^{Θ}, where $\Theta = \Omega \times \Omega$. We define the union ($\cup$) and intersection ($\cap$) of the two shapes $F, G \in \Theta$:

$$\eta_{F\cap G}(x,y,u,v) = min(\eta_F(x,y,u,v), \eta_G(x,y,u,v))$$

$$\eta_{F\cup G}(x,y,u,v) = max(\eta_F(x,y,u,v), \eta_G(x,y,u,v))$$

Simplicity $q(F) \in [0,1]$ is a probabilistic measure on $(\Theta, \cup, \cap)$:

$$q(O) = 1 \text{ for the simplest shape } O(\eta_O(x,y,u,v) \equiv 1);$$

$$q(\varnothing) = 0 \text{ for the most complex shape } \varnothing(\eta_{\varnothing}(x, y, u, v) \equiv 0);$$
$$q(F \cup G) = q(F) + q(G) - q(F \cap G); \text{ (simplicity addition rule)}$$
$$q(F \cup G) = q(F) + q(G) \text{ for incompatible shapes } (F \cap G = \varnothing);$$
$$q(F \cap G) = q(F)q(G) \text{ for independent shapes } (q(F \cap G)/q(F) = q(G)).$$

We well know the probability theory based on some set of "events" with operations (AND, OR) and the measure for "event probability". We propose here the analogous construction with a set Θ of relational shapes with operations $(\cup, \cap)$ and the measure for "shape simplicity" to be called the morphological theory of simplicity.

Simplicity has the meaning of the geometric probability for a random pair of points to fall into the same coverage area (partition). This is a morphological (internal geometric) probability associated with the morphology (internal geometry) of a given specific shape, and not with a statistical ensemble of shapes or a class of images of a given shape.

It is easy to demonstrate that all tools for morphological analysis of mosaic shapes can be reintroduced based on the theory of simplicity. Morphological shape correlation coefficient (MSCC) for coverings:

$$K_M^2(G, F) = \sum_{i=1,\ldots,n} p_i \sum_{j=1,\ldots,m} q(G_j|F_i) = \sum_{i=1,\ldots,n} \sum_{j=1,\ldots,m} p_{ij}^2/p_i, \tag{4}$$

where $q(G_j|F_i) = q(G_j \cap F_i)/q(F_i)$ – conditional simplicity of G_j with respect to F_i. In other words, MSCC (4) is the average over frame local conditional simplicity of G with respect to F.

Based on this, we introduce the global (by Θ) analogues of the MSCC – conditional and posterior simplicity for hypothesis Φ and observation Γ:

$$q(G|F) = q(G \cap F)/q(F) = \sum_{i=1,\ldots,n} \sum_{j=1,\ldots,m} p_{ij}^2 / \sum_{k=1,\ldots,n} p_k^2, \tag{5}$$

$$q(F|G) = q(F \cap G)/q(G) = \sum_{i=1,\ldots,n} \sum_{j=1,\ldots,m} p_{ij}^2 / \sum_{l=1,\ldots,m} p_l^2, \tag{6}$$

which are connected by the Bayes formula for simplicity:

$$q(F|G) = q(G|F)q(F)/q(G). \tag{7}$$

Thus, this work proposes the theory of simplicity for mosaic shapes. We provide the interpretation of the MSCC as the average local estimate of conditional simplicity and propose the new tool for shape comparison based on posterior and conditional estimates of simplicity.

[1] *Pyt'ev Yu., Chulichkov A.* Morphological methods for image analysis // Moscow: Fizmatlit Publisher, 2010. 336 p.

[2] *Vizilter Yu., Zheltov S., Busurin V.* Modern morphological analysis and its application in aviation technical vision systems // Moscow: MAI, 2020. 176 p.

Теория простоты и мозаичные покрытия

Визильтер Юрий Валентинович[1]⋆ viz@gosniias.ru
[1]Москва, ФГУП ГосНИИАС

В морфологии Пытьева [1] для любых двух форм (разбиений кадра Ω) F и G можно указать более сложную форму $F \wedge G$ и менее сложную $F \vee G$. В работе [2] предложена «теория простоты» форм на основе вероятностной меры $q(F)$, определяемой как $L1$-норма реляционной модели $\eta_F(x, y, u, v) \in 0, 1$:

$$q(F) = \|\eta_F(x, y, u, v)\|_1 = \sum_{(x,y,u,v)\in\Omega\times\Omega} \eta_F(x, y, u, v). \quad (1)$$

При этом определены операции объединения ($\cup$) и пересечения ($\cap$) форм. Однако решетка $(\Theta, \cup, \cap)$ с границами $(O, \varnothing)$ не совпадает с решеткой пытьевских форм. Во-первых, $\cap = \wedge$, но $\cup \neq \vee$. Во-вторых, $\varnothing$ не является разбиением (есть несуществующие пикселы ($\eta_F(x, y, x, y) = 0$)). Значит, решетка $(\Theta, \cup, \cap)$ включает локализованные формы [3]. В-третьих, операция $\cup$ незамкнута для мозаичных форм, поскольку $\eta_{F\cup G}$ не имеет «блочно-диагонального» вида. Таким образом, $(\cup, \cap)$-замыкание не совпадает с известными классами форм. Введем явно этот новый класс мозаичных форм.

Мозаичным покрытием $\mathbf{\Phi}$ кадра Ω площади S назовем пару разбиение-покрытие, или (эквивалентно) множество пар областей кадра вида:

$$\mathbf{\Phi} = \langle F, \Phi\rangle = \{\langle F_1, \Phi_1\rangle, \ldots, \langle F_n, \Phi_n\rangle\}, \quad (2)$$

таких, что: $F_1 \cup \ldots \cup F_n = \Omega; F_i \cap F_k = \varnothing; (\Phi_i = \varnothing)$ или $(\Phi_i \supseteq F_i); (F_i \nsubseteq \Phi_k) \Rightarrow (F_i \cap \Phi_k = \varnothing); (F_i \subseteq \Phi_k) \Leftrightarrow (F_k \subseteq \Phi_i), i = 1, \ldots, n; k = 1, \ldots, n$; где n – число пар $\langle$носитель, покрытие$\rangle$; $F = \{F_1, \ldots, F_n\}$ – набор носителей, составляющих базовое разбиение $\Omega; \Phi = \{\Phi_1, \ldots, \Phi_n\}$ – набор покрытий, составляющих покрытие Ω. В частности, при $F = \Phi$, $\mathbf{\Phi}$ – полное мозаичное разбиение кадра.

Для $\mathbf{\Phi} = \langle F, \Phi\rangle$ и $\mathbf{\Gamma} = \langle G, \Gamma\rangle = \{\langle G_1, \Gamma_1\rangle, \ldots, \langle G_l, \Gamma_l\rangle\}$ обозначим: $p_i, z_i, p_j, z_j, p_{ij}, z_{ij}$ – нормированые площади областей $F_i, \Phi_i, G_j, \Gamma_j, F_i \cap G_j, \Phi_i \cap \Gamma_j$ соответственно.

Реляционная форма $\mathbf{\Phi}$ описывается отношением покрытия:

$$\eta_\Phi(x, y, u, v) = \begin{cases} 1, \text{если } \exists i : ((x, y) \in F_i;) \text{ и } ((u, v) \in \Phi_i); \\ 0, \text{в противном случае} \end{cases} \quad (3)$$

Покрытия с одинаковыми реляционными формами считаются эквивалентными. Пусть $(\mathbf{\Phi})^*$ - операция приведения к каноническому виду, где нет носителей пустых ($\varnothing$) и с совпадающими покрытиями. Тогда:

$$\mathbf{\Phi} \cap \mathbf{\Gamma} = \mathbf{Y} = \langle W, Y\rangle = (\{\langle W_{ij} = F_i \cap G_j, Y_{ij} = \Phi_i \cap \Gamma_j\rangle\}_{j=1,\ldots,l;i=1,\ldots,n})^*,$$

$$\mathbf{\Phi} \cup \mathbf{\Gamma} = \mathbf{B} = \langle V, B\rangle = (\{\langle V_{ij} = F_i \cap G_j, B_{ij} = \Phi_i \cup \Gamma_j\rangle\}_{j=1,\ldots,l;i=1,\ldots,n})^*.$$

Можно доказать, что класс мозаичных покрытий является $(\cup, \cap)$ - замыканием множества локализованных мозаичных разбиений. Это позволяет ввести для покрытий все морфологические инструменты.

Мера простоты:

$$q(\Phi) = \|\eta_\Phi\|_{L1} = \sum_{i=1,\ldots,n} q(\Phi_i) = \sum_{i=1,\ldots,n} p_i z_i \tag{4}$$

Метрическое пространство форм ($\{0,1\}$-пространство Хэмминга в Θ):

$$\begin{aligned} d_{H\Theta}(\Phi,\Gamma) &= \|\eta_\Phi(x,y,u,v) - \eta_\Gamma(x,y,u,v)\|_{L1} = \\ &= \sum_{j=1,\ldots,l} \sum_{i=1,\ldots,n} p_{ij}(z_i + z_j - 2z_{ij}) = \\ &= q(\Phi) + q(\Gamma) - 2q(\Phi \cap \Gamma) = q(\Phi \cup \Gamma) - q(\Phi \cap \Gamma). \end{aligned} \tag{5}$$

Морфологический коэффициент корреляции форм (МККФ) для покрытий:

$$K_M^2(\Gamma,\Phi) = \sum_{i=1,\ldots,n} p_i \sum_{j=1,\ldots,m} q(\Gamma_j|\Phi_i) = \sum_{i=1,\ldots,n} \sum_{j=1,\ldots,m} p_{ij} z_{ij}/z_i, \tag{6}$$

где $q(\Gamma_j|\Phi_i) = q(\Gamma_j \cap \Phi_i)/q(\Phi_i)$ – условная простота Γ_j относительно Φ_i.

Обобщенное тождество простоты для покрытий (аналог [4]):

$$\begin{aligned} q(\Phi) = \|\eta_\Phi(x,y,u,v)\|_{L1} &= \sum_{i=1,\ldots,n} p_i z_i = d_H(\Phi,\varnothing) = \\ &= 1 - d_H(\Phi,O) = K_M^2(\Phi,O) = q(\Phi|O). \end{aligned} \tag{7}$$

Связь $\cup$ с $\vee$.Пусть операция $[\Phi]$ замыкания по связности последовательно находит в $\mathbf{\Phi}$ все области с пересекающимися покрытиями и заменяет их покрытия на объединение покрытий. Легко убедиться, что

$$F \vee G = ([F \cup G])^* \tag{8}$$

Таким образом, в данной работе предложена морфология мозаичных покрытий, согласованная с простотой как вероятностной мерой и включающая в качестве подкласса пытьевскую морфологию мозаичных разбиений.

[1] *Пытьев Ю. П., Чуличков А. И.* Методы морфологического анализа изображений // М: Физматлит, 2010. 336 с.

[2] *Визильтер Ю. В.* Морфологическая теория простоты // Математические методы распознавания образов: Тезисы докладов 20-й Всероссийской конференции, 2021.

[3] *Визильтер Ю. В., Желтов С. Ю., Бусурин В. И.* Современный морфологический анализ и его применение в авиационных системах технического зрения // М: Издво МАИ, 2020. 176 с.

[4] *Визильтер Ю. В., Выголов О. В., Желтов С. Ю.* «Формула Эйлера» для морфологического анализа мозаичных изображений // Интеллектуализация обработки информации: Тезисы докладов 13-й Международной конференции, 2020.

The Simplicity Theory and Mosaic Coverings

Vizilter Yuri[1]⋆ viz@gosniias.ru

[1]Moscow, GosNIIAS

In the Pyt'ev morphology [1] for any two shapes (tessellations of the frame Ω) F and G we can find the more complex shape $F \wedge G$ and less complex shape $F \vee G$. In paper [2] was proposed a "theory of simplicity" of forms based on the probability measure $q(F)$, defined as the $L1$-norm of the relational model $\eta_F(x, y, u, v) \in 0, 1$:

$$q(F) = \|\eta_F(x, y, u, v)\|_1 = \sum_{(x,y,u,v)\in\Omega\times\Omega} \eta_F(x, y, u, v). \quad (1)$$

We define the union ($\cup$) and intersection ($\cap$) of the shapes. However, the lattice $(\Theta, \cup, \cap)$ with boundaries $(O, \varnothing)$ does not match the lattice of Pyt'ev shapes. Firstly $\cap = \wedge$, but $\cup \neq \vee$. Secondly, $\varnothing$ is not a tessellation(has some non-existent pixels ($\eta_F(x, y, x, y) = 0$)). So, the lattice $(\Theta, \cup, \cap)$ included a set of localized shapes [3]. Thirdly, the operation $\cup$ is not closed for a set of mosaic shapes, because $\eta_{F\cup G}$ does not have "block-diagonal" view. Thus, $(\cup, \cap)$ operations does not coincide with any known class of shapes. So, let us introduce this new class of mosaic shapes in the evident form.

The mosaic cover $\mathbf{\Phi}$ of the frame Ω with area S is called a "tessellation-covering" pair, or (equivalently) a set of region pairs of the form:

$$\mathbf{\Phi} = \langle F, \Phi\rangle = \{\langle F_1, \Phi_1\rangle, \ldots, \langle F_n, \Phi_n\rangle\}, \quad (2)$$

such that: $F_1 \cup \ldots \cup F_n = \Omega; F_i \cap F_k = \varnothing; (\Phi_i = \varnothing)$ or $(\Phi_i \supseteq F_i); (F_i \nsubseteq \Phi_k) \Rightarrow (F_i \cap \Phi_k = \varnothing); (F_i \subseteq \Phi_k) \Leftrightarrow (F_k \subseteq \Phi_i), i = 1, \ldots, n; k = 1, \ldots, n$; where n – the number of pais ⟨support, cover⟩; $F = \{F_1, \ldots, F_n\}$ – set of supports, which form the basic partition of $\Omega; \Phi = \{\Phi_1, \ldots, \Phi_n\}$ – set of covers, which form the overall covering of Ω. In the particular case of $F = \Phi, \mathbf{\Phi}$ – is a complete tessellation of the frame.

For $\mathbf{\Phi} = \langle F, \Phi\rangle$ and $\mathbf{\Gamma} = \langle G, \Gamma\rangle = \{\langle G_1, \Gamma_1\rangle, \ldots, \langle G_l, \Gamma_l\rangle\}$ denote: $p_i, z_i, p_j, z_j, p_{ij}, z_{ij}$ – normalized areas of regions $F_i, \Phi_i, G_j, \Gamma_j, F_i \cap G_j, \Phi_i \cap \Gamma_j$ respectively.

The relational shape $\mathbf{\Phi}$ is described by the "covering relation":

$$\eta_\Phi(x, y, u, v) = \begin{cases} 1, \text{if } \exists i : ((x, y) \in F_i;) \text{ and } ((u, v) \in \Phi_i); \\ 0, \text{otherwise} \end{cases} \quad (3)$$

Coverings with the same relational models are considered as equivalent. Denote $(\mathbf{\Phi})^*$ the operation of bringing to a "canonical form" in which there are no supports with empty ($\varnothing$) or the same covers:

$$\mathbf{\Phi} \cap \mathbf{\Gamma} = \mathbf{Y} = \langle W, Y \rangle = (\{\langle W_{ij} = F_i \cap G_j, Y_{ij} = \Phi_i \cap \Gamma_j \rangle\}_{j=1,\ldots,l;i=1,\ldots,n})^*,$$

$$\mathbf{\Phi} \cup \mathbf{\Gamma} = \mathbf{B} = \langle V, B \rangle = (\{\langle V_{ij} = F_i \cap G_j, B_{ij} = \Phi_i \cup \Gamma_j \rangle\}_{j=1,\ldots,l;i=1,\ldots,n})^*.$$

One can prove that the set (class) of mosaic coverings is the $(\cup, \cap)$-closure of the set of localized mosaic tesselations. This we can introduce all morphological tools for coverings.

The complete order by simplicity and the simplicity measure:

$$q(\Phi) = \|\eta_\Phi\|_{L1} = \sum_{i=1,\ldots,n} q(\Phi_i) = \sum_{i=1,\ldots,n} p_i z_i \tag{4}$$

Metric space of shapes ($\{0,1\}$-Hamming space in Θ):

$$\begin{aligned} d_{H\Theta}(\Phi, \Gamma) &= \|\eta_\Phi(x,y,u,v) - \eta_\Gamma(x,y,u,v)\|_{L1} = \\ &= \sum_{j=1,\ldots,l} \sum_{i=1,\ldots,n} p_{ij}(z_i + z_j - 2z_{ij}) = \\ &= q(\Phi) + q(\Gamma) - 2q(\Phi \cap \Gamma) = q(\Phi \cup \Gamma) - q(\Phi \cap \Gamma). \end{aligned} \tag{5}$$

Morphological shape correlation coefficient (MSCC) for coverings:

$$K_M^2(\Gamma, \Phi) = \sum_{i=1,\ldots,n} p_i \sum_{j=1,\ldots,m} q(\Gamma_j|\Phi_i) = \sum_{i=1,\ldots,n} \sum_{j=1,\ldots,m} p_{ij} z_{ij}/z_i, \tag{6}$$

where $q(\Gamma_j|\Phi_i) = q(\Gamma_j \cap \Phi_i)/q(\Phi_i)$ – conditional simplicity of Γ_j with respect to Φ_i.

The "generalized simplicity identity" expression (analogue of [4]):

$$\begin{aligned} q(\Phi) = \|\eta_\Phi(x,y,u,v)\|_{L1} = \sum_{i=1,\ldots,n} p_i z_i = d_H(\Phi, \varnothing) = \\ 1 - d_H(\Phi, O) = K_M^2(\Phi, O) = q(\Phi|O). \end{aligned} \tag{7}$$

Finally, we state the connection between $\cup$ and $\vee$. Denote $[\Phi]$ the "connectivity closure" operation that finds the pairs of regions in $\mathbf{\Phi}$ with intersecting covers and replaces them with their union until there will be no pair with intersecting covers in Φ. It's easy to see that

$$F \vee G = ([F \cup G])^* \tag{8}$$

Thus, this work proposes the morphology of mosaic coverings, consistent with "simplicity" as a probabilistic measure and including the Pyt'ev morphology of mosaic tessellations as a subclass.

[1] *Pyt'ev Yu., Chulichkov A.* Morphological methods for image analysis // Moscow: Fizmatlit Publisher, 2010. 336 p.

[2] *Vizilter Yu.* The Morphological Simplicity Theory // Mathematical methods of pattern recognition: Theory and Applications: Book of abstract of the 20th Russian National, 2021.

[3] *Vizilter Yu., Zheltov S., Busurin V.* Modern morphological analysis and its application in aviation technical vision systems // Moscow: MAI, 2020. 176 p.

[4] *Vizilter Yu., Vygolov O., Zheltov S.* "Euler Identity" for Morphological Image Analysis // Intelligent Data Processing: Theory and Applications: Book of abstract of the 13th International Conference, Moscow, 2020.

Моделирование циклических процессов решениями кусочно-линейных разностных уравнений с постоянными коэффициентами по экспериментальным данным в виде временных рядов

Смирнов Владимир Юрьевич[1]⋆ smirnovvlyu@yandex.ru
Кузнецова Анна Викторовна[2] azfor@narod.ru
[1]Москва, ООО "Азфорус"
[2]Москва, ФГБУН ИБХФ им. Н.М. Эмануэля РАН

В работе предложен метод моделирования динамических циклических процессов, представленных временными рядами. Модель состоит из решений двух (или более) линейных разностных уравнений с постоянными коэффициентами. Предлагаемые модели относятся только к системам со свободными колебаниями (как ответ на задание начальных условий). Для систем с управляющими воздействиями при предлагаемом типе моделирования потребуется пересчитывать коэффициенты модели для каждого момента включения управляющего воздействия, которое в этом случае следует рассматривать как задание новых начальных условий. Интегральные кривые линейных разностных уравнений, входящих в модель, состыковываются по непрерывности. Переключение с одного уравнения на другое происходит, когда интегральная кривая достигает порога – границы в фазовом пространстве. Модель аппроксимирует по наименьшим квадратам экспериментальные данные (временной ряд) на которых она строится функциями вида $A \times e^{\alpha \times j \times \Delta t} \times \cos(\omega \times j \times \Delta t + \varphi) + B$. Такой выбор аппроксимирующих функций представляется наиболее соответствующим реальным динамическим процессам макромира – росту популяций (при отсутствии врагов и достаточной кормовой базе), выведению лекарств из организма, а также всем физическим процессам накопления / потребления взаимосвязанных ресурсов, когда скорость накопления / потребления пропорциональна имеющейся массе ресурсов. Использование модели с более чем одной системой линейных уравнений соответствует принципу управления по "узкому месту" – ситуации, когда некоторый ресурс находится в избытке и не влияет на динамику всего процесса в целом. В этом случае весь процесс распадается на отдельные звенья, которые могут быть описаны уравнениями 1-го – 2-го порядков.

Условие существования циклов

Алгоритм вычисления коэффициентов и начальных условий для аппроксимируюших моделей (авторегрессий) был описан нами в [1]. Необходимо отметить, что этот алгоритм использует все имеющиеся точки экспериментального временного ряда как для получения коэффициентов аппроксимирующей авторегресии, так и для ее начальных условий. Кроме того, алгоритм дает единственное решение и не использует какие-либо эвристики. В настоящей работе мы представляем главный результат – возможность конструирования замкнутого цикла в фазовом пространстве любой размерности с помощью авторегрессий

2-го (или больших) порядков. В [1] было показано, что авторегрессия, например 2-го порядка, может быть записана как:

$$Y(j) = A(j) \times Y_1 + B(j) \times Y_0 + C(j) \tag{1}$$

Тут коэффициенты $A(j)$, $B(j)$ и $C(j)$ являются переменными величинами, а Y_1 и Y_0 - постоянны и являются искомыми начальными условиями авторегрессии. Было показано, что для получения последовательности величин $A(j)$, $B(j)$ и $C(j)$ сначала необходимо вычислить коэффициенты обычной формы представления авторегрессии – из системы уравнений, формально соответствующей системе уравнений Юла-Уокера. Затем вычисляются все значения величин $A(j)$, $B(j)$ и $C(j)$, а по ним – значения Y_1 и Y_0 - из системы уравнений, минимизирующих среднеквадратичное отклонение (и по коэффициентам, и по этим начальным условиям).

Тот факт, что авторегрессия может быть представлена в форме (1), позволяет послать кривую - решение авторегрессии из любого начального положения (из точки Y_0 фазового пространства) в любую заданную точку $Y(n)$ за заданное число шагов (n). Необходимо лишь решить уравнение: $Y(n) = A(n) \times Y_1 + B(n) \times \times Y_0 + C(n)$ относительно Y1:

$$Y_1 = \frac{(Y(n) - B(n)) \times (Y_0 - C(n))}{A(n)} \tag{2}$$

Тогда авторегрессия в форме (1) и с начальными условиями Y_0 и Y_1 примет на (n)-м шаге в точности значение $Y(n)$.

Как результат – есть возможность получить замкнутый цикл в фазовом пространстве любой размерности – надо лишь послать замыкающую ветвь второго уравнения из некоторой точки ветви 1-го процесса в его начальную точку.

Кроме того, такой тип моделирования позволяет перевести моделируемую систему в заданное состояние - точку-цель из любого текущего положения даже в случае, когда целевое состояние меняется с течением времени.

На рисунке 1 представлен результат компьютерного моделирования циклического процесса на фазовой плоскости X-Y (Y – вертикальная координата). В работе также проведен анализ скорости сходимости интегральных кривых к устойчивому циклу, анализ устойчивости аппроксимации экспериментальных циклических данных, и найдена зависимость между коэффициентами авторегрессии и коэффициентами представления ее решения в виде кривой, зависящей от времени.

[1] *Загоруйко Н. Г.* Гипотезы компактности и λ-компактности в методах анализа данных // Сиб. журн. индустр. матем., 1998. Т. 1(1). С. 114–126.

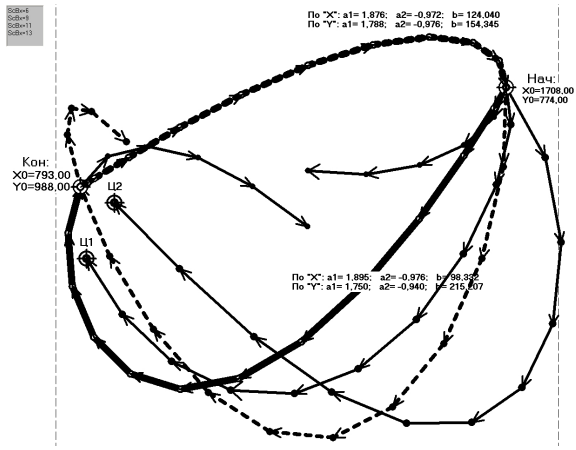


Рис. 1. The result of computer simulation of a cyclic process on the phase plane X-Y

Modeling of Cyclic Processes (of time series type experimental date) by Solving of Piecewise Linear Differential Equations with Constant Coefficients

Smirnov Vladimir[1]⋆ smirnovvlyu@yandex.ru
Kuznetsova Anna[2] azfor@narod.ru
[1]Moscow, OOO "Azforus"
[2]Moscow, Federal State Budgetary Institution of Science Institute of Biochemical Physics named after N.M. Emanuel of the Russian Academy of Sciences

This paper presents a method for dynamic cyclic processes modeling, represented by time series. The model is obtained in form of the solutions of two (or more) linear differential equations with constant coefficients.

Proposed models can be applied only to systems with free reactions (as a response for setting initial conditions). For systems with forcing impact the proclaimed type of modeling will cause the recalculation of model's coefficients for each new moment of forcing impact, considered to be new initial conditions. Integral curves of linear differential equations systems are connected by continuity. Switching from one system of equations to another is obtained then integral curves reaches some threshold - frontier in phase sheet (space).

The solution of linear differential equations approximates experimental data (in the form of time series) by functions of $A \times e^{\alpha \times j \times \Delta t} \times \cos(\omega \times j \times \Delta t + \varphi) + B$ shape (or sign-alternating functions) on condition of minimizing the RMS deviation.

Such choice of approximating functions is considered to be the most related to a physical sense of many real dynamical processes of macroworld – the growth of population of some species (under condition of enemies absence and a lot of food), elimination of medicines from organism and to all physical processes of accumulation or consumption some interconnected resources, when the speed of accumulation / consumption is proportional the amount mass of resources.

Using of model with more than one system of linear differential equations connected with principle of control by "narrow place" – the situation, when some resource presented in more then enough amount and not influence for a dynamic of whole process. In such case whole process may be considered as a number of chains which can be described by equations of 1-st or 2-nd order.

Cycle existing conditions

Algorithm calculation of coefficients and initial conditions for such models (autoregressions) was described in [1]. Necessary to notice that this algorithm uses all points of experimental time series data for calculation of coefficients and initial conditions for equation in autoregression form. In addition to it, algorithm gives the single solution and not needs in any heuristic methods.

In this paper we would like to present the main result – ability to construct the close cycle in phase space of any dimension with help of autoregressions of the 2-nd (or more) order.

It was shown in [1] that autoregression of 2-nd order, for example can be written as:

$$Y(j) = A(j) \times Y_1 + B(j) \times Y_0 + C(j) \tag{1}$$

The coefficients $A(j)$, $B(j)$ and $C(j)$ have variable meanings, but Y_1 and Y_0 are the constant initial conditions to be found.

Sequence of $A(j)$, $B(j)$ and $C(j)$ can be calculated after calculation of usual form autoregression coefficients, and these coefficients of usual form must be obtained as a solution of the equation Yule-Walker system type. After it the initial conditions Y_0 and Y_1 must be calculated as a solution of equations system, minimizing the RMS approximation deviation (both for coefficients and initial conditions).

Because autoregression can be represented in (1) form, it makes possible to send a modeling object from any initial condition (point Y_0 in phase space) to a prescribed final point $Y(n)$ in phase space for a prescribed amount of steps (n). It is necessary to solve the equation: $Y(n) = A(n) \times Y_1 + B(n) \times Y_0 + C(n)$ and found here the initial condition Y_1:

$$Y_1 = \frac{(Y(n) - B(n)) \times (Y_0 - C(n))}{A(n)} \tag{2}$$

Then autoregression in form (1) with Y_0 and Y_1 initial meanings, calculated in such way takes the exact meaning $Y(n)$ at step n.

As a result were found conditions for a constructing closed cycle, consisting of two (or more) branches in phase space of any dimension – it is necessary to send the brunch of 2-nd process from some point of it to the start point of 1-st process to archive the closed cycle.

The result of computer modeling of cyclic process on a phase plate X-Y is shown in figure 1. The analysis of stability and convergence rate of the solution for such closed cycles is made. The essential influence of equation's initial conditions to a shape of integral curves is underlined in paper. The relations between the coefficients of autoregression and coefficients of formula in continuous form for integral curve is shown in paper too.

In additional, such modeling let us to archive the aim point from any start point also let us to get aim in a case of time changing aim position.

[1] *Smirnov V. Yu., Kuznetsova A. V.* Approximation of Experimental Data by Solving Linear Differential Equations with Constant Coefficients (in Particular, by Exponentials and Exponential Cosines) // Pattern Recognition and Image Analysis, 2017. Vol. 27(2). Pp. 175–183.

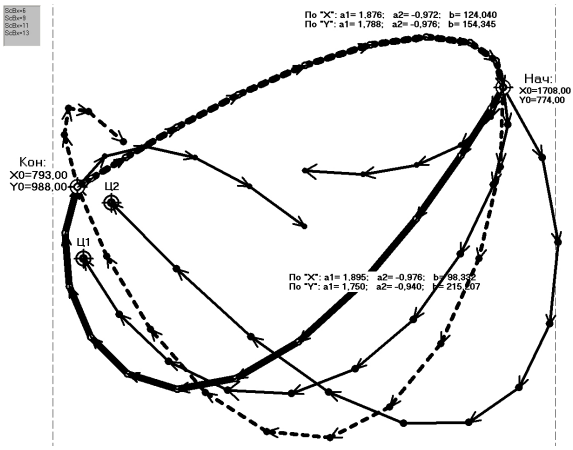


Fig. 1. The result of computer simulation of a cyclic process on the phase plane X-Y

Нейросетевой метод решения задачи обобщённого неметрического многомерного шкалирования

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

Задачи многомерного шкалирования возникают, когда необходимо понизить размерность векторных представлений объектов или реализовать матрицу близостей набора объектов в векторном пространстве. С практической точки зрения это может быть нужно для визуализации и дальнейшего анализа, более компактного представления и уменьшения места для хранения либо более быстрого вычисления, например, скалярного произведения. В задачах перевода естественных языков, рекомендательных системах и системах информационного поиска актуально погружение набора объектов в векторное пространство, поскольку в этом случае становятся доступны структуры данных для быстрого приближённого поиска ближайших соседей [1, 2].

Постановки задач многомерного шкалирования условно разделяются на два класса, в завимости от того, требуется ли сохранить значения близостей набора объектов или порядки близостей. Когда необходимо сохранить порядки близостей, могут быть заданы как сами близости, так и только их порядки, например, в виде матрицы, где в строках и столбцах находятся пары объектов из набора, а в ячейках знаки между различиями в одной и другой парах объектов (см. Рис. 1).

	(i, j)	.	.	.	(k, l)
(i, j)	#	<	>	<	>
.	#	#	>	>	<
.	#	#	#	<	>
.	#	#	#	#	<
(k, l)	#	#	#	#	#

Рис. 1. Матрица порядков близостей

Когда заданы только порядки близостей и необходимо найти такие векторные представления объектов, чтобы для некоторой функции расстояния новые порядки близостей не изменились (1), то в случае когда показателем качества выступает доля сохранённых порядков близостей, такая постановка называется задачей обобщённого неметрического многомерного шкалирования, Generalized

Non-metric Multidimensional Scaling, GNMDS [3, 4]

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

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

Задача GNMDS может быть эффективно решена с помощью нейронной сети, состоящей из одного полносвязного слоя (см. Рис. 2). Изначально веса слоя и векторные представления объектов задаются случайным образом. В ходе обучения, по матрице порядков близостей для каждой четвёрки объектов, состоящей из двух пар объектов, определяется, верно ли сохранён порядок близостей у векторных представлений объектов, полученных в результате применения полносвязного слоя к изначальным векторным представлениям. Количество ошибок и является функцией потерь, которая может быть приближена (2)

$$max(0,\ e(g(i),\ g(j)) - e(g(k),\ g(l)) + margin) \tag{2}$$

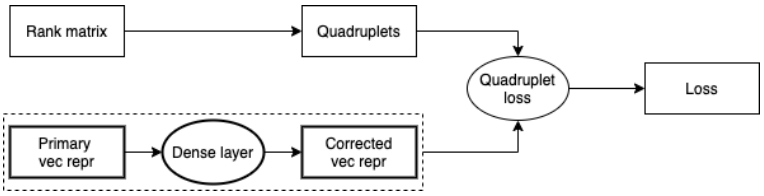


Рис. 2. Общая схема решения задачи GNMDS

В результате решения описанной задачи оптимизации может быть получен такой набор векторных представлений объектов, что для него приведённая функция потерь равна нулю — это означает, что сохранены все порядки близостей, а значит получено точное решение задачи GNMDS. В случае, когда не удаётся найти нуль функции потерь, полученное решение является приближённым — мерой точности решения является доля верно сохранённых порядков близостей. Условия, при которых получается точное решение, являются вопросом для дальнейшего исследования.

Описанный метод реализован в виде библиотеки на языке python [5].

Работа выполнена при поддержке некоммерческого Фонда развития науки и образования «Интеллект».

[1] *Johnson J., Douze M., Jégou H.* Billion-scale similarity search with GPUs // arXiv preprint arXiv:1702.08734, 2017.

[2] *Bernhardsson E.* Annoy // github.com/spotify/annoy

[3] *Agarwal S. et al* Generalized Non-metric Multidimensional Scaling // Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, 2007. Pp. 11–18.

[4] *Bronstein A. et al* Generalized Multidimensional Scaling: A Framework for Isometry-Invariant Partial Surface Matching // Proceedings of the National Academy of Sciences, 2006. Pp. 1168–72.

[5] *Kolosov A.* obj2vec // github.com/obj2vec/gnmds

Neural network method for solving the problem of generalized non-metric multidimensional scaling

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

Multidimensional scaling problems arise when it is necessary to reduce the dimension of vector representations of objects or to embed a set of objects similarity matrix in vector space. From a practical point of view, this may be necessary for visualization and further analysis, for a more compact representation and less storage space or for a faster calculation, for example, a dot product. In the tasks of translating natural languages, recommender systems and information retrieval systems, it is important to embed a set of objects in a vector space, since in this case data structures become available for a fast approximate search for nearest neighbors [1, 2].

Problem statements for multidimensional scaling are conventionally divided into two classes, depending on whether it is required to preserve the values of the similarities of a set of objects or orders of similarities. When it is necessary to preserve the orders of similarities, both the similarities themselves and only their orders can be specified, for example, in the form of a matrix, where in rows and columns there are pairs of objects from a set, and in cells there are signs between differences in one and another pair of objects (see Fig. 1).

	(i, j)	.	.	.	(k, l)
(i, j)	#	<	>	<	>
.	#	#	>	>	<
.	#	#	#	<	>
.	#	#	#	#	<
(k, l)	#	#	#	#	#

Fig. 1. Rank similarity matrix

When only the orders of similarities are given and it is necessary to find such vector representations of objects so that for a certain distance function the new orders of similarities do not change (1), then in the case when the quality metric is the fraction of preserved orders of similarities, such a formulation is called the problem of Generalized Non-metric Multidimensional Scaling, GNMDS [3, 4]

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

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

The GNMDS problem can be effectively solved using a neural network consisting of one fully connected layer (see Fig. 2). Initially, layer weights and vector representations of objects are set randomly. When training, using the rank similarity matrix for each quadruplet of objects, consisting of two pairs of objects, it is determined whether the order of similarity in vector representations of objects obtained as a result of applying a fully connected layer to the original vector representations is correctly preserved. The number of errors is a loss function that can be approximated (2)

$$max(0,\ e(g(i),\ g(j)) - e(g(k),\ g(l)) + margin) \tag{2}$$

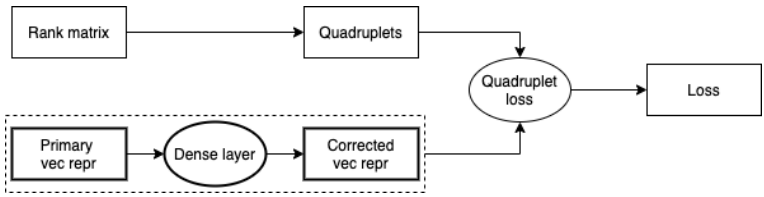


Fig. 2. General scheme for solving the GNMDS problem

As a result of solving the described optimization problem, a set of vector representations of objects can be obtained such that the given loss function is equal to zero, which means that all orders of similarities are preserved, which means that an exact solution of the GNMDS problem is obtained. In the case when it is not possible to find the zero of the loss function, the obtained solution is approximate - the measure of the accuracy of the solution is the fraction of correctly preserved orders of similarities. The conditions under which the exact solution is obtained are a question for further research.

The described method is implemented as a python library [5].

The research is carried out with the financial support of the Intellect Nonprofit Foundation for the Development of Science and Education.

[1] *Johnson J., Douze M., Jégou H.* Billion-scale similarity search with GPUs // arXiv preprint arXiv:1702.08734, 2017.

[2] *Bernhardsson E.* Annoy // github.com/spotify/annoy

[3] *Agarwal S. et al* Generalized Non-metric Multidimensional Scaling // Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, 2007. Pp. 11–18.

[4] *Bronstein A. et al* Generalized Multidimensional Scaling: A Framework for Isometry-Invariant Partial Surface Matching // Proceedings of the National Academy of Sciences, 2006. Pp. 1168–72.

[5] *Kolosov A.* obj2vec // github.com/obj2vec/gnmds

Дифференцируемый алгоритм поиска архитектуры с контролем сложности

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

В работе рассматривается задача оптимизации модели глубокого обучения с контролем сложности архитектуры. Предлагаемый метод основан на дифференцируемом алгоритме поиска архитектуры (DARTS) [3]. Структурные параметры задаются гипесетью [5], зависящей от коэффициента, задающего сложность архитектуры. Под гиперсетью понимается модель, порождающая параметры оптимизируемой модели. Структура модели представлена в виде ориентированного ациклического графа, где ребра соответствуют нелинейным отображениям на выборке, а вершины — промежуточными представлениями выборки под действием этих отображений. Структурные параметры, соответствующие определенному ребру подчинены распределению Gumbel-Softmax [4].

Вычислительный эксперимент проводится на выборке Fashion-MNIST. Сравниваются модели, полученные с помощью предложенного метода, модели, полученные с помощью DARTS, а также модели со случайными архитектурами. Результаты показали, что предложенный метод сопоставим по точности с DARTS, однако позволяет контролировать сложность архитектуры модели, изменяя коэффициент, задающий сложность. Чем выше значение коэффициента, тем проще получаемая архитектура.

[1] *Bakhteev Yu., Strijov V.* Comprehensive analysis of gradient-based hyperparameter optimization algorithms // Annals of Operations Research, 2020. Vol. 289(1). Pp. 51–65.

[2] *Ha D., Dai A., Le Q.* Hypernetworks // arXiv preprint arXiv:1609.09106, 2016.

[3] *Liu P., Simonyan K., Yang H.* Darts: Differentiable architecture search // arXiv preprint arXiv:1806.09055, 2018.

[4] *Maddison C., Mnih A., Teh Y.* The concrete distribution: A continuous relaxation of discrete random variables // arXiv preprint arXiv:1611.00712, 2016.

[5] *Гребенькова О. С., Бахтеев О. Ю, Стрижов В. В.* Вариационная оптимизация модели глубокого обучения с контролем сложности // Информатика и её применения, 2021. Т. 15(1). С. 42–49.

Differentiable architecture search with model complexity control

Yakovlev Konstantin[1]⋆ iakovlev.kd@phystech.edu
Grebenkova Olga[1] grebenkova.os@phystech.edu
Bakhteev Oleg[1,2] bakhteev@phystech.edu
Strijov Vadim[1,2] strijov@phystech.edu
[1]Moscow, Moscow Institute of Physics and Technology
[2]Moscow, FRCCSC of the Russian Academy of Sciences

The paper investigates the problem of optimizing a deep learning model with the control of complexity of the architecture. The proposed method is based on a differentiable architecture search algorithm (DARTS) [3]. The structural parameters are generated by a hypernetwork [5], depending on a regularization parameter. The hypernetwork is a model that generates parameters of the optimized model. The structure of the model is presented in the form of a directed acyclic graph, where the edges correspond to nonlinear maps on the objects, and the vertices are intermediate representations of the objects under the action of these maps. The structural parameters corresponding to a certain edge are taken from Gumbel-Softmax [4] distribution.

The computational experiment is performed on a dataset Fashion-MNIST. Models obtained using the proposed method, models obtained using DARTS, as well as models with random architectures are compared. The results showed that the proposed method is comparable to DARTS, but allows us to control the complexity of the model architecture by changing the regularization parameter. The higher the parameter is, the simpler the resulting architecture.

[1] *Bakhteev Yu., Strijov V.* Comprehensive analysis of gradient-based hyperparameter optimization algorithms // Annals of Operations Research, 2020. Vol. 289(1). Pp. 51–65.

[2] *Ha D., Dai A., Le Q.* Hypernetworks // arXiv preprint arXiv:1609.09106, 2016.

[3] *Liu H., Simonyan K., Yang H.* Darts: Differentiable architecture search // arXiv preprint arXiv:1806.09055, 2018.

[4] *Maddison C., Mnih A., Teh Y.* The concrete distribution: A continuous relaxation of discrete random variables // arXiv preprint arXiv:1611.00712, 2016.

[5] *Grebenkova O., Bakhteev O., Strijov V.* Variational Optimization of a Deep Learning Model with Complexity Control // Informatics and Applications, 2021. Vol. 15(1). Pp. 42–49.

Градиентные методы оптимизации метапараметров в задаче дистилляции знаний

Горпинич Мария[1]⋆ gorpinich.m@phystech.edu
Бахтеев Олег Юрьевич[1,2] bakhteev@phystech.edu
Стрижов Вадим Викторович[1,2] strijov@ccas.ru

[1]Москва, Московский физико-технический институт (национальный исследовательский университет)
[2]Москва, Вычислительный центр имени А.А. Дородницына Федерального исследовательского центра «Информатика и управление» Российской академии наук

В работе рассматривается задача дистилляции модели глубокого обучения. Дистилляция знаний — это задача оптимизации, в которой учитывается как информация, содержащаяся в исходной выборке, так и информация, содержащаяся в модели более сложной структуры, называемой моделью-учителем. Функция потерь состоит из двух слагаемых: правдоподобия исходной выборки и правдоподобия выборки дистилляции. Под метапараметрами модели понимаются параметры оптимизационной задачи дистилляции, а именно, коэффициент перед слагаемым дистилляции в функции потерь и параметр температуры. Параметры модели-ученика оптимизируются путем переноса знаний от модели-учителя. Задачу дистилляции с оптимизацией метапараметров модели-ученика можно представить в виде двухуровневой задачи. На первом уровне оптимизируются метапараметры, на втором — параметры модели. Двухуровневая задача решается градиентными методами. Для уменьшения вычислительной сложности задачи траектория оптимизации частично предсказывается линейной моделью.

Для анализа данного метода проводится вычислительный эксперимент на синтетической выборке, выборках изображений CIFAR-10 и Fashion-MNIST. По результатам эксперимента можно сделать вывод об эффективности использования градиентных методов в задаче назначения метапараметров для функции потерь в задаче дистилляции. Также был проанализирован предложенный метод аппроксимации траектории оптимизации метапараметров с помощью линейных моделей.

[1] *Hinton G., Vinyals O., Dean J.* Distilling the Knowledge in a Neural Network // CoRR, 2015.

[2] *Luketina J., Berglund M., Greff K., Raiko T.* Scalable Gradient-Based Tuning of Continuous Regularization Hyperparameters // CoRR, 2015.

[3] *Bakhteev O., Strijov V.* Comprehensive analysis of gradient-based hyperparameter optimization algorithms // Ann. Oper. Res, 2020. Pp. 51–65.

Gradient-based metaparameter optimization in knowledge distillation task

Gorpinich Mariya[1]⋆ gorpinich.m@phystech.edu
Bakhteev Oleg[1,2] bakhteev@phystech.edu
Strijov Vadim[1,2] strijov@ccas.ru
[1]Moscow, Moscow Institute of Physics and Technology
[2]Moscow, Dorodnicyn Computing Center RAS

The paper investigates the knowledge distillation problem for deep learning models. Knowledge distillation is a model parameter optimization problem that takes into account not only information contained in the initial dataset but also the information contained in the model with a more complex structure that is called the teacher model. The loss function has the initial dataset likelihood term and distillation dataset likelihood term. Metaparameters of this problem are the parameters of the optimization problem, namely, the coefficient of the distillation term in loss function and a temperature factor. The student model parameters are optimized by transferring the knowledge of a teacher model. The knowledge distillation problem with metaparameter optimization is a bi-level problem. Metaparameters are optimized on the first level and model parameters are optimized on the second. Gradient methods are used to solve the bi-level problem. To make the solution less computationally expensive we alternate gradient optimization steps and optimization trajectory prediction with linear models.

The method is evaluated using the synthetic dataset, Fashion-MNIST dataset, and CIFAR-10 dataset. The computational experiment showed the effectiveness of gradient-based optimization for selecting metaparameters of the distillation loss function. The possibility of optimization path approximation using linear models was analyzed.

[1] *Hinton G., Vinyals O., Dean J.* Distilling the Knowledge in a Neural Network // CoRR, 2015.

[2] *Luketina J., Berglund M., Greff K., Raiko T.* Scalable Gradient-Based Tuning of Continuous Regularization Hyperparameters // CoRR, 2015.

[3] *Bakhteev O., Strijov V.* Comprehensive analysis of gradient-based hyperparameter optimization algorithms // Ann. Oper. Res, 2020. Pp. 51–65.

Регуляризация параметров нейронной сети на основе неравенства Рисса

Григорьев Алексей Дмитриевич[1]⋆ grigorev.ad@phystech.edu
Гнеушев Александр Николаевич[1,2] gneushev@ccas.ru
[1]Москва, Московский физико-технический институт (НИУ)
[2]Москва, Федеральный исследовательский центр "Информатика и управление" РАН

Нейронные сети прямой связи с большим числом параметров позволяют описывать сложные нелинейные зависимости с высокой точностью путем оптимизации весов модели методом обратного распространения ошибки. Высокая размерность пространства параметров зачастую приводит к коррелированности нейронов сети, что уменьшает обобщающую способность модели и ведет к неэффективному использованию вычислительных ресурсов. В работе ставится задача регуляризации параметров нейросетевой модели для уменьшения ее избыточности и повышения устойчивости. Предлагается рассматривать веса слоя нейросети как реализации семейства базисных функций гильбертова пространства, таких что проекция входной функции на этот базис устойчива и полна. В отличии от известных методов, основанных на ортогонализации параметров, предлагается более общий подход, при котором после обучения веса каждого слоя образовывают функциональный фрейм, соответствующий базису Рисса. Избыточность базиса Рисса свойственна нейронной сети и позволяет точнее описывать ее веса. Полнота и устойчивость фрейма аналогичны свойствам ортонормированного базиса. Таким образом, в работе формулируется новая функция потерь Рисса, реализующая слабые ограничения на параметры модели, но придающая базисные свойства нейросетевому слою. Вычислительные эксперименты, проведенные на выборках CIFAR10, CIFAR100 и ImageNet, показали высокую эффективность предложенного метода обучения по сравнению с альтернативными подходами регуляризации параметров. Исследована зависимость качества модели от значений границ фрейма. Задача ортогонализации параметров модели обобщена предложенным методом регуляризации Рисса с соответствующими границами фрейма.

Neural network parameters regularization based on Riesz inequality

Grigorev Alexey[1]⋆ grigorev.ad@phystech.edu
Gneushev Alexander[1,2] gneushev@ccas.ru
[1]Moscow, Moscow Institute of Physics and Technology
[2]Moscow, Federal Research Center "Computer Science and Control" of RAS

Feedforward neural networks with a large number of parameters are capable of describing complex nonlinear dependencies with high accuracy by optimizing the model parameters using the backpropagation algorithm. The high dimensionality of the parameter space often leads to the network neurons correlation which reduces the generalization ability of the model and leads to inefficient use of computational resources. The paper considers the problem of neural network parameters regularization to reduce redundancy and increase the stability of the model. The weights of a neural network layer are proposed to be considered as an implementation of a family of basis functions in the Hilbert space such that the projection of the input function onto this basis is stable and complete. In contrast to the well-known methods based on parameters orthogonalization, a more general approach is proposed; the method implies that after training the parameters of each layer form a functional frame corresponding to the Riesz basis. Redundancy of the Riesz basis is inherent in a neural network and allows a more accurate description of its parameters. The completeness and stability of the frame are similar to those of the orthonormal basis. Thus, the paper formulates a new Riesz loss function that implements weak constraints on the model parameters but provides basis properties to the neural network layer. Computational experiments that were carried out on the CIFAR10, CIFAR100, and ImageNet datasets showed the high efficiency of the proposed method in comparison with alternative approaches to parameters regularization. The dependence of model quality on the frame bounds is investigated. The problem of model parameters orthogonalization is generalized by the proposed Riesz regularization method with the corresponding frame bounds.

Порождение моделей заданной сложности с использованием байесовских гиперсетей

Гребенькова Ольга Сергеевна[1]⋆ grebenkova.os@phystech.edu
Бахтеев Олег Юрьевич[1,2] bakhteev@phystech.edu
Стрижов Вадим Викторович[1,2] strijov@phystech.edu
[1]Москва, Московский физико-технический институт
[2]Москва, Вычислительный центр им. А.А. Дородницына ФИЦ ИУ РАН

В работе рассматривается задача оптимизации модели глубокого обучения. Построение модели заданной сложности — одна из фундаментальных проблем глубокого обучения, так как по построению данное семейство моделей имеет избыточное число параметров [1]. В работе предполагается, что сложность модели задана заранее. Сложность модели интерпретируется как коэффицент, настраиваемый при использовании модели в зависимости от эксплутационных требований.

Выбор модели глубокого обучения вычислительно сложный процесс. Мы предлагаем вместо оптимизации модели с заранее выбранным гиперпараметром, контролирующим сложность, оптимизировать сразу семейство моделей. Предлагаемый метод заключается в представлении параметров модели глубокого обучения в виде гиперсети [2]. Гиперсеть — функция, которая порождает параметры желаемой модели. Другими словами, гиперсеть — это отображение из множества переменных, отвечающих за сложность модели, во множество параметров модели.

В данной работе используется байесовский подход. Вводятся вероятностные предположения о параметрах моделей глубокого обучения. Для проверки возможности гиперсети порождать модели разной сложности были реализованы разные методы прореживания моделей [1, 5]. В данной статье исследуется упрощенный случай, когда предполагается, что параметры модели распределены нормально [1]. Были исследованы две функции для оптимизации модели, основанные на вариационном байесовском подходе [1, 6]. Также было проведено сравнение этих функций с детерминированной оптимизацией модели, где использовалась l_2 регуляризация. Проведен анализ их свойств для нашего метода.

Вычислительный эксперимент проводился на выборке рукописных цифр MNIST [3] и на выборке CIFAR-10 [4]. Полученные в результате эксперимента гиперсети порждают как простые, так и сложные модели в зависимости от требуемой сложности. Эти модели имеют те же свойства, что и модели, обученные напрямую, но для их получения требуется меньшее количество вычислительных ресурсов. Кроме того, они более стабильны с точки зрения удаления параметров.

Работа выполнена при поддержке Российского фонда фундаментальных исследований, проект No 19-07-00875.

[1] *Graves, A.* Practical Variational Inference for Neural Networks // Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems, 2011. Pp. 2348–2356.

[2] *Ha A., Dai M., Le Q.* HyperNetworks // The International Conference on Learning Representations (ICLR), 2017.

[3] *LeCun Y., Cortes C., Burges C.* The MNIST dataset of handwritten digits // Available at: `http://yann.lecun.com/exdb/mnist/index.html`, 1998.

[4] *Krizhevsky A., Vinod N., Hinton G.* The CIFAR-10 (Canadian Institute for Advanced Research) dataset // Available at `http://www.cs.toronto.edu/~kriz/cifar.html`

[5] *Song H., Pool J., Tran J., Dally W.* Learning both Weights and Connections for Efficient Neural Network // Advances in Neural Information Processing Systems, 2015. Vol. 28.

[6] *Sambasivan R., Das S., Sahu S., Dally W.* A Bayesian perspective of statistical machine learning for big data // Computational Statistics, 2020. Vol. 35(3). Pp. 893–930.

Model selection using Bayesian hypernetworks

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

The paper considers the problem of optimizing a deep learning model. Optimal model selection is the fundamental problem of deep learning, since this family of models has an excessive number of parameters implicitly [1]. The paper assumes that the complexity of the model is given. We consider the model complexity as a coefficient assigned during model final training depending on the operational requirements.

The deep model selection procedure is expensive in calculations. Instead of optimizing a model with some predefined hyperparameter value that controls the model complexity, we propose to optimize a family of models. We propose to represent the parameters of the model in the form of a hypernetwork. A hypernetwork is a function, which generates the parameters of the desired model [2]. In other words, a hypernetwork is a mapping from a set of variables responsible for the complexity of a desired model to a set of its parameters.

This paper uses a Bayesian approach. We introduce a probabilistic assumptions about the distribution of parameters of the deep learning model. To demonstrate that we obtain models of different complexity by optimized hypernetworks, we implement the model pruning methods [1, 5]. This paper investigates a simple case when the model parameters are assumed to be distributed with a Gaussian distribution [1]. We investigate two forms of model optimization functions that are based on variational Bayesian approach [1, 6]. We compare them with a simple deterministic model optimization with l_2 regularization and analyze their properties for our optimization method.

The computational experiment is carried out on a sample of handwritten digits MNIST [3] and CIFAR-10 [4]. The resulting hypernetworks generate both simple and complex models depending on the required model properties. These models have the same properties as models trained directly but use less computational resources. Furthermore, they are more stable in terms of deleting parameters.

This work was supported by the Russian Foundation for Basic Research, project No. 19-07-00875.

[1] *Graves, A.* Practical Variational Inference for Neural Networks // Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems, 2011. Pp. 2348–2356.

[2] *Ha A., Dai M., Le Q.* HyperNetworks // The International Conference on Learning Representations (ICLR), 2017.

[3] *LeCun Y., Cortes C., Burges C.* The MNIST dataset of handwritten digits // Available at: `http://yann.lecun.com/exdb/mnist/index.html`, 1998.

[4] *Krizhevsky A., Vinod N., Hinton G.* The CIFAR-10 (Canadian Institute for Advanced Research) dataset // Available at `http://www.cs.toronto.edu/~kriz/cifar.html`

[5] *Song H., Pool J., Tran J., Dally W.* Learning both Weights and Connections for Efficient Neural Network // Advances in Neural Information Processing Systems, 2015. Vol. 28.

[6] *Sambasivan R., Das S., Sahu S., Dally W.* A Bayesian perspective of statistical machine learning for big data // Computational Statistics, 2020. Vol. 35(3). Pp. 893–930.

Вычислительные технологии поиска низкопотенциальных состояний кластеров Морса размерностей от 460 до 690 атомов

Сороковиков Павел Сергеевич[1]⋆ sorokovikov.p.s@gmail.com
Горнов Александр Юрьевич[1] gornov@icc.ru
[1]Иркутск, Институт динамики систем и теории управления имени В.М. Матросова СО РАН

В настоящее время продолжает возрастать интерес к проблеме получения сверхмелкодисперсных структур атомных кластеров. Цель подобных исследований заключается в нахождении кластерной структуры с минимальной потенциальной энергией. Задачи поиска низкопотенциальных состояний атомных кластеров заключаются в минимизации невыпуклых функций [1], характеризуются чрезвычайно быстрым увеличением количества локальных оптимумов с ростом числа переменных. Поэтому для их решения требуется использование специальных подходов, учитывающих специфику исследуемых структур.

Для решения задач оптимизации потенциалов атомных кластеров реализованы вычислительные технологии, включающие специализированные алгоритмы глобального и локального поиска. Библиотеку глобализованных оптимизационных алгоритмов составляют модификации алгоритмов MSBH («Monotonic Sequence Basin-Hopping»), «парабол», Пауэлла, туннельного поиска, Розенброка, Лууса–Яаколы, Растригина, «стохастических покрытий», табу-поиска, дифференциальной эволюции, генетического поиска, биогеографии, поиска гармоний, опыления цветков, поиска по принципу «учитель-ученик», роя светлячков и других. Коллекция алгоритмов локальной оптимизации включает в себя квазиньютоновский алгоритм L-BFGS, методы сопряжённых градиентов, декомпозиционный «рейдер-метод» и прочие.

С применением разработанных вычислительных технологий проведено численное исследование кластеров Морса со сверхбольшим числом частиц (атомов). Минимизируемый функционал имеет следующую структуру: $f(x) = \sum\limits_{i=1}^{N} \sum\limits_{j=i+1}^{N} e^{\rho(1-r_{ij})} \left(e^{\rho(1-r_{ij})} - 2\right)$, где N – количество частиц, r_{ij} – расстояние между атомами i и j. К настоящему времени опубликованы результаты расчетов для кластеров Морса из 5–147 (в базе данных [1]), 148–240 (в работах групп из Китая [2] и Португалии [3]), 241–450 атомов (в ранних публикациях авторов и А.С. Аникина). В данной работе проведены системные численные расчеты нахождения низкоэнергетических состояний кластеров из 460–690 частиц с шагом 10 при $\rho = 3$. В Таблице 1 приведены наилучшие найденные значения целевой функции при указанных размерностях. В результате исследования полученных результатов расчетов была выявлена достаточно регулярная закономерность изменения наилучшего из известных («best of known») значения в зависимости от количества частиц в кластере. Авторам неизвестно о других попытках числен-

ного решения задач оптимизации конформации атомных кластеров Морса для рассматриваемых размерностей.

Таблица 1. Полученные значения потенциальной функции

N	Значение	N	Значение	N	Значение
460	-6243.648248	540	-7548.160216	620	-8876.820224
470	-6403.362570	550	-7700.516816	630	-9046.997894
480	-6559.388537	560	-7881.460360	640	-9208.003264
490	-6718.903821	570	-8034.433295	650	-9367.283286
500	-6893.650168	580	-8210.328570	660	-9549.828351
510	-7054.414052	590	-8374.874243	670	-9708.553979
520	-7216.954715	600	-8544.026103	680	-9882.546522
530	-7378.800085	610	-8704.551373	690	-10061.910151

[1] *Wales D., Doye J.* The Cambridge Energy Landscape Database // URL http://www-wales.ch.cam.ac.uk/CCD.html.

[2] *Feng Y., Cheng L., Liu H.* Putative global minimum structures of Morse clusters as a function of the range of the potential: 161⩽n⩽240 // The Journal of Physical Chemistry A, 2009. Vol. 113(49). Pp. 13651–13655.

[3] *Marques J., Pais A., Abreu P.* On the use of big-bang method to generate low-energy structures of atomic clusters modeled with pair potentials of different ranges // Journal of Computational Chemistry, 2012. Vol. 33(4). Pp. 442–452.

Computational technologies for the search for low-potential states of Morse clusters with dimensions from 460 to 690 atoms

Sorokovikov Pavel[1]⋆ sorokovikov.p.s@gmail.com
Gornov Alexander[1] gornov@icc.ru
[1]Irkutsk, Matrosov Institute for System Dynamics and Control Theory of SB RAS

At present, interest continues to grow in the problem of obtaining ultrafine structures of atomic clusters. The purpose of such studies is to find the cluster structure with the minimum potential energy. The problems of searching for low-potential states of atomic clusters consist in minimizing non-convex functions [1] and are characterized by an extremely rapid increase in the number of local extrema with a rise in the number of variables. Therefore, their investigation requires specific approaches that take into account the specifics of the structures under study.

Computational technologies have been implemented, including specialized algorithms for global and local search, to solve the problems of optimizing the potentials of atomic clusters. The library of globalized optimization algorithms consists of modifications of the MSBH ("Monotonic Sequence Basin-Hopping"), "parabolas", Powell, tunnel search, Rosenbrock, Luus-Jaakola, Rastrigin, "stochastic coverings", tabu search, differential evolution, genetic search, biogeography, harmony search, flower pollination, teaching–learning-based search, firefly swarm algorithms, and others. The collection of local optimization algorithms includes the L-BFGS quasi-Newtonian algorithm, conjugate gradient methods, decomposition "raider method", and others.

A numerical study of Morse clusters with an extremely large number of particles (atoms) has been carried out using the developed computational technologies. The functional to be minimized has the following structure: $f(x) = \sum_{i=1}^{N} \sum_{j=i+1}^{N} e^{\rho(1-r_{ij})} \left(e^{\rho(1-r_{ij})} - 2\right)$, where N is the number of particles, r_{ij} is the distance between atoms i and j. To date, the results of calculations have been published for Morse clusters with 5–147 (in the database [1]), 148–240 (in the works of groups from China [2] and Portugal [3]), 241–450 atoms (in the early publications of the authors and A.S. Anikin). In this work, we performed systemic numerical calculations of finding the low-energy states of clusters of 460–690 particles with a step of 10 at $\rho = 3$. Table 1 shows the best-found values of the objective function for the specified dimensions. As a result of the study of the obtained calculation results, a fairly regular pattern of changes in the "best of known" value was revealed depending on the number of particles in the cluster. The authors aren't aware of other efforts to numerically solve the problems of optimizing the conformation of Morse atomic clusters for the dimensions under consideration.

[1] *Wales D., Doye J.* The Cambridge Energy Landscape Database // URL http://www-wales.ch.cam.ac.uk/CCD.html.

Table 1. The obtained values of the potential function

N	Value	N	Value	N	Value
460	-6243.648248	540	-7548.160216	620	-8876.820224
470	-6403.362570	550	-7700.516816	630	-9046.997894
480	-6559.388537	560	-7881.460360	640	-9208.003264
490	-6718.903821	570	-8034.433295	650	-9367.283286
500	-6893.650168	580	-8210.328570	660	-9549.828351
510	-7054.414052	590	-8374.874243	670	-9708.553979
520	-7216.954715	600	-8544.026103	680	-9882.546522
530	-7378.800085	610	-8704.551373	690	-10061.910151

[2] *Feng Y., Cheng L., Liu H.* Putative global minimum structures of Morse clusters as a function of the range of the potential: 161⩽n⩽240 // The Journal of Physical Chemistry A, 2009. Vol. 113(49). Pp. 13651–13655.

[3] *Marques J., Pais A., Abreu P.* On the use of big-bang method to generate low-energy structures of atomic clusters modeled with pair potentials of different ranges // Journal of Computational Chemistry, 2012. Vol. 33(4). Pp. 442–452.

Генеративная модель автокодировщиков, обучающихся на изображениях представленых выборкмаи отсчетов

Анциперов Вячеслав Евгеньевич antciperov@cplire.ru
Москва, Инст. радиотехники и электроники им.В.А.Котельникова РАН

В данной работе мы предлагаем новый подход к обучению автокодировщиков по изображениям, заданных специальным образом - с помощью представлений в виде выборки осчетов (выборочных представлений). Эти представления тесно связаны с концепцией идеального изображения. В первой половине работы мы обсуждаем как саму концепцию, так и ее редукцию к используемым выборочным представлениям. Вторая часть работы посвящена генеративной модели обучающихся автоэнкодеров в условиях, когда входные данные имеют вид выборочного представления. Для этого случая мы приводим результаты формализации генеративной модели обучения, оптимизации процедур кодирования / декодирования и их связь с известным в статистической теории методом максимального правдоподобия Фишера.

В нескольких предыдущих статьях [1, 2] мы предложили представление изображений выборками случайных отсчетов (т.н. статистикой отсчетов [3]). Такое представление мотивировано, с одной стороны, развитием SPAD-матриц на основе однофотонных лавинных диодов, которые регистрируют излучение в виде дискретного набора фотоотсчетов, а с другой стороны, постоянно растущей тенденцией адаптация зрительного восприятия человека к цифровой обработке изображений. Обе тенденции предполагают аналогичную структуру формирования изображения - чувствительную поверхность, состоящую из очень большого количества детекторов/пикселей. При этом предполагается, что эти детекторы настолько малы, что каждый из них может детектировать одиночный фотон падающего излучения. Перечисленные общие черты могут быть положены в основу концепции идеального устройства визуализации, обобщающего, помимо упомянутых SPAD-матриц, также сетчатку глаза с фоторецепторами, фотопластинки с желатино-серебряной эмульсией и пр.

Формально под идеальным устройством формирования изображения мы подразумеваем $2D$-поверхность Ω с координатами $\boldsymbol{x} = (x_1, x_2)$, на которой идентичные точечные детекторы, или "джоты" в терминах [4] расположены вплотную друг к другу. Точечные детекторы по определению имеют исчезающе маленькую площадь da светочувствительных поверхностей. Соответственно, общее количество детекторов составляет $N = A/da$, где A - общая площадь поверхности Ω. В предположении, что A фиксировано и $da \to 0$, число N полагается очень большим: $N \to \infty$. Таким образом, идеальное устройство формирования изображения - это почти непрерывная чувствительная поверхность Ω с координатами $\boldsymbol{x}$, определяющими положение идеальных точечных детекторов.

Основываясь на концепции идеального устройства формирования изображения, можно сформулировать модель собственно идеального изображения как

результирующий набор (фото) отсчетов процесса регистрации излучения. А именно, под идеальным изображением мы подразумеваем (упорядоченное) множество $X = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_n}), \boldsymbol{x_i} \in \Omega$ из n случайных отсчетов, зарегистрированных идеальным устройством. Мы используем название "идеальное изображение"для предлагаемой конструкции вслед за авторами [5]. Рассматривая X как генеральную совокупность отсчетов, мы предлагаем использовать только ее небольшую случайную выборку X_k заданного размера $k \ll n$ для представления изображения. Мы называем это X_k представлением выборкой случайных отсчетов, или просто выборочным представлением.

Полное статистическое описание выборочного представления $X_k = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_k})$ может быть получено из модели идеального изображения [**?**, **?**] и с высокой точностью плотность распределения вероятностей X_k имеет вид:

$$\rho(X_k|I(\boldsymbol{x})) = \prod_{j=1}^{k} \rho(\boldsymbol{x_j}|I(\boldsymbol{x})), \tag{1}$$

$$\rho(\boldsymbol{x_j}|I(\boldsymbol{x})) = \frac{I(\boldsymbol{x_j})}{W}, \quad W = \iint\limits_{\Omega} I(\boldsymbol{x})da$$

где $I(\boldsymbol{x})$ обозначает регистрируемую интенсивность.

Осуществив необходимую формализацию генеративной модели в случае, когда автонкодировщики обучаются на изображениях, заданных в форме выборочного представления (1), мы обнаружили, что процесс кодирования-декодирования может быть естественным образом реализован в виде итерационной вычислительной процедуры. Эта процедура чем-то напоминает известный EM-алгоритм [6] и состоит в двух-шаговых вычислениях на каждой из повторяющихся итераций:

Кодирование $\mathcal{G} \to \mathcal{F} : \rho(\boldsymbol{x}|\boldsymbol{\theta}) \to \boldsymbol{f}$:

$$\boldsymbol{f}(X_k, \boldsymbol{\theta}) = \\ arg \max_{j \in \{1,\ldots,K\}} \left(\rho(\boldsymbol{x}_1, j|\boldsymbol{\theta}), ..., \rho(\boldsymbol{x}_k, j|\boldsymbol{\theta})\right), \tag{2}$$

Декодирование $\mathcal{F} \to \mathcal{G} : \boldsymbol{f} \to \rho(\boldsymbol{x}|\boldsymbol{\theta})$:

$$\boldsymbol{\theta}_{ML} = arg \max_{\boldsymbol{\theta} \in \Theta} \left(\rho(X_k, \boldsymbol{f}|\boldsymbol{\theta})\right) \\ \Rightarrow \boldsymbol{s_f}\left(\boldsymbol{\theta}_{ML}, X_k)\right) = \mathbf{0}. \tag{3}$$

где $\boldsymbol{s_f}$ - вклад в правдоподобие, $\boldsymbol{\theta}_{ML}$ - оценка максимального правдоподобия, последнее уравнение в (3) является достаточным условием Фишера максимального правдоподобия.

[1] *Antsiperov V.* Maximum Similarity Method for Image Mining // Pattern Recognition. ICPR 2021. Lecture Notes in Computer Science, 2021. Vol. 12665. Pp. 301–313.

[2] *Antsiperov V.* Representation of Images by the Optimal Lattice Partitions of Random Counts // Pattern Recognition and Image Analysis, 2021. Vol. 31(3). Pp. 381–393.

[3] *Fox M.* Quantum Optics: An Introduction. // Oxford, NY: Oxford University Press, 2006.

[4] *Fossum E.* The Invention of CMOS Image Sensors: A Camera in Every Pocket. // Pan Pacific Microelectronics Symposium, 2020. Pp. 1–6.

[5] *Pal N., Pal S.* Image model, Poisson distribution and object extraction. // J. Pattern Recognition and AI, 1991. Vol. 5(3). Pp. 459–483.

[6] *Gupta M.R.* Theory and Use of the EM Algorithm. // Foundations and Trends in Signal Processing, 2010. Vol. 1(3). Pp. 223–296.

Generative model of autoencoders learning images by count sample representations

Antsiperov Vyacheslav antciperov@cplire.ru

[1]Moscow, Kotelnikov Institute of Radioengineering and Electronics RAS

In this work, we propose a new approach to autoencoders learning by images that are given in a special way – by special count sample (sampling) representations. These representations are closely related to the concept of an ideal image. The first half of the work in detail discussed this relation. The second part is devoted to generative model of learning autoencoders providing that input is given as an image sampling representation. The questions of generative model formalization, optimization of encoding / decoding procedures and connection with the renowned Fisher's method of maximum likelihood in statistical theory are presented below for this case.

In several previous papers [1, 2] we proposed the representation of images by samples of random counts (i. e. by the counting statistics [3]). This representation is motivated, on the one hand, by progress in the SPAD (single photon avalanche diodes) video matrices, that register radiation in the form of a discrete set of photocounts and on the other hand, by the ever-increasing trend in the adaptation of human visual perception for digital image processing. Both trends incorporate almost the same image forming structure – sensitive surface consisting of a very large number of detectors/pixels. These detectors are assumed to be small enough, so that each of them can detect the single photon of incident radiation. So, the listed features can be taken as the basis of the concept of ideal imaging device, generalizing besides mentioned SPAD sensors also the retina with photoreceptors, photographic plates with gelatin-silver emulsion, etc.

Formally, by an ideal imaging device we assume a $2D$-surface Ω with coordinates $\boldsymbol{x} = (x_1, x_2)$, on which identical point detectors, or "jots" in terms of [4] are allocated close to each other. Point detectors have by a definition a vanishingly small area da of light-sensitive surfaces. Accordingly, the total number of detectors is $N = A/da$, where A is the total area of surface Ω. Under the assumption that A is fixed and $da \to 0$, the number N is assumed to be arbitrarily large: $N \to \infty$. Thus, the ideal imaging device is an almost continuous sensitive surface Ω, with the coordinates $\boldsymbol{x}$, specifying positions of its ideal point detectors.

Based on the ideal imaging device concepts, it is possible to formulate the model of an ideal image as a resultant set of (photo) counts of a radiation registration process. Namely, under the ideal image we mean the (ordered) set $X = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_n}), \boldsymbol{x_i} \in \Omega$ of n random counts registered by the ideal imaging device. We use the name "ideal image" for the proposed construction, following the authors of [5], which introduced this term in the early 90-s. Considering X as a certain general population of counts, we propose to use only its small random sample X_k of fixed size $k \ll n$ to represent the image. Obviously, in full agreement with the

classical statistical paradigm, such a "sample" representation will still represent the ideal image X. We call this X_k representation by a sample of random counts, or, in short, the sampling representation.

The complete statistical description of sampling representation $X_k = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_k})$ can be deduced from the ideal image model [1, 2] and with high accuracy the probability distribution density of X_k has the form:

$$\rho(X_k|I(\boldsymbol{x})) = \prod_{j=1}^{k} \rho(\boldsymbol{x_j}|I(\boldsymbol{x})), \tag{1}$$

$$\rho(\boldsymbol{x_j}|I(\boldsymbol{x})) = \frac{I(\boldsymbol{x_j})}{W}, \quad W = \iint\limits_{\Omega} I(\boldsymbol{x})da.$$

where $I(\boldsymbol{x})$ denotes the intensity registered.

Basing on the generative model formalization in the case of autoencoders learning the images given in the form of sampling representation (1), we found that the encoding–decoding process can be naturally implemented in the form of a recurrent computational procedure. This procedure is somewhat reminiscent of the well-known EM-algorithm [6] and consists in the the following two-step recurrent iterations:

Encoding $\mathcal{G} \to \mathcal{F} : \rho(\boldsymbol{x}|\boldsymbol{\theta}) \to \boldsymbol{f}$:

$$\boldsymbol{f}(X_k, \boldsymbol{\theta}) = \tag{2}$$
$$arg \max_{j \in \{1,\ldots,K\}} \left(\rho(\boldsymbol{x}_1, j|\boldsymbol{\theta}), ..., \rho(\boldsymbol{x}_k, j|\boldsymbol{\theta})\right),$$

Decoding $\mathcal{F} \to \mathcal{G} : \boldsymbol{f} \to \rho(\boldsymbol{x}|\boldsymbol{\theta})$:

$$\boldsymbol{\theta}_{ML} = arg \max_{\boldsymbol{\theta} \in \Theta} \left(\rho(X_k, \boldsymbol{f}|\boldsymbol{\theta})\right) \tag{3}$$
$$\Rightarrow \boldsymbol{s_f}\left(\boldsymbol{\theta}_{ML}, X_k\right)) = \boldsymbol{0}$$

where $\boldsymbol{s_f}$ is the score, $\boldsymbol{\theta}_{ML}$ – maximum likelihood estimation, so the last equation in (3) is the Fischer's sufficient condition for maximal likelihood.

[1] *Antsiperov V.* Maximum Similarity Method for Image Mining // Pattern Recognition. ICPR 2021. Lecture Notes in Computer Science, 2021. Vol. 12665. Pp. 301–313.

[2] *Antsiperov V.* Representation of Images by the Optimal Lattice Partitions of Random Counts // Pattern Recognition and Image Analysis, 2021. Vol. 31(3). Pp. 381–393.

[3] *Fox M.* Quantum Optics: An Introduction. // Oxford, NY: Oxford University Press, 2006.

[4] *Fossum E.* The Invention of CMOS Image Sensors: A Camera in Every Pocket. // Pan Pacific Microelectronics Symposium, 2020. Pp. 1–6.

[5] *Pal N., Pal S.* Image model, Poisson distribution and object extraction. // J. Pattern Recognition and AI, 1991. Vol. 5(3). Pp. 459–483.

[6] *Gupta M.R.* Theory and Use of the EM Algorithm. // Foundations and Trends in Signal Processing, 2010. Vol. 1(3). Pp. 223–296.

Сверточный иерархический нейросетевой классификатор

Гаджиев Исмаил Маратович[1]⋆ ismailgadzhievff@gmail.com
Доленко Сергей Анатольевич[2] 0000-0001-6214-3195
[1]Москва, НИИ ядерной физики имени Д.В.Скобельцына
[2]Москва, МГУ имени М.В.Ломоносова

Доклад посвящен разработке и исследованию сверточной модификации алгоритма иерархического нейросетевого классификатора (ИНК) [1]. Алгоритм выполняет рекурсивную классификацию путем построения дерева классов, в каждом узле которого отдельной нейронной сетью распознается подмножество всех классов (каждый дочерний узел содержит подмножество классов узла-родителя). Построение дерева классов происходит в процессе обучения за счет слияния классов, которые классификатор в узле путает чаще всего. Алгоритм призван объединить сильные стороны деревьев решений и нейронных сетей, в частности, уменьшить количество параметров при сохранении достаточной сложности модели.

Проблема уменьшения количества параметров особенно актуальна для задач компьютерного зрения – модели, выдающие наиболее качественные результаты для «тяжелых» задач, требуют оптимизации большого количества параметров. Например, лучшие результаты на задаче ImageNet показаны в статье [2]. Модель использовала 2440 млн параметров. Однако, в работе [3] сформулирована гипотеза о том, что до 90% весов нейронной сети являются неинформативными – авторы выбирают случайную подсеть исходной сети и получают схожее качество работы.

Уменьшение количества параметров может позволить эффективное использование модели на «слабых» вычислителях с ограниченной памятью и ограниченными вычислительными возможностями, например, на мобильных телефонах, что является весьма актуальным.

Модификация ИНК, представленная в докладе, - это попытка применения ИНК к задаче распознавания изображений путем замены классификаторов в узлах дерева на сверточные нейронные сети. Желаемым эффектом использования иерархии классов является уменьшение числа параметров модели. В отличие от оригинального ИНК обучение происходит в две стадии. На первой стадии строится дерево классов за счет «слабых» сверточных сетей с малым количеством параметров, на второй происходит замена сверточных сетей в узлах на более сильные с их обучением на подмножествах исходного набора.

Опишем алгоритм построения дерева классов для сверточного ИНК (СИНК) (первая стадия). В каждом узле дерева находится нейронная сеть, которая обучается распознавать примеры из некоторого подмножества всех классов. В процессе обучения сети классы, которые сеть чаще всего путает, объединяются.

Для определения того, какие классы сеть путает между собой, используется процедура «голосования» примеров каждого класса; объединяются те клас-

сы, большинство примеров в которых «проголосовали» в один и тот же класс. После голосования желаемые ответы для объединяемых классов делаются одинаковыми, а обучение сети продолжается с переразмеченными таким образом примерами. В случаях, если дальнейшее слияние классов невозможно, или если прошло установленное максимальное количество эпох без слияний, процедура обучения останавливается.

В результате классы оказываются объединенными в группы в соответствии с тем, как сеть в узле их путает, а эффективность и качество обучения после объединения классов существенно возрастают. Затем на каждой из полученных групп обучается новая нейронная сеть – таким образом рекурсивно выстраивается дерево классов.

Как отмечалось ранее, после этого производится вторая стадия обучения – модели в узлах заменяются на более сложные, и производится повторное обучение.

Алгоритм реализован с помощью библиотеки Tensorflow [4].

Для тестирования алгоритма использована эталонная задача CIFAR-10 [5].

В докладе обсуждаются методы, позволяющие улучшить качество работы СИНК. Одним из гиперпараметров СИНК является порог по активациям нейронов при голосовании (голосуют только те примеры, максимальная активация выходных нейронов для которых превышает порог). Этот гиперпаметр наследуется сверточной модификацией от оригинала. Обсуждается целесообразность использования порога по активациям в рамках сверточной модификации.

Важно подчеркнуть, что дерево классов, полученное на первой стадии обучения СИНК, может быть переиспользовано для других задач. Например, в модели YOLO9000 [6], предназначенной для сегментирования объектов на изображениях, используется иерархическая относительно классов функция потерь, где ошибки считаются на всех уровнях дерева классов. Иерархия классов выстраивается на основе семантики слов, обозначающих этих классы, полученной из словарей. Однако семантическое родство меток классов не означает визуальное родство примеров. Поэтому целесообразной является возможность использования иерархии классов, полученной адаптивно из данных.

Обсуждаются также дальнейшие направления исследования и совершенствования алгоритма, т.к. полученные результаты по качеству распознавания пока далеки от идеальных, сильно уступая результатам, получаемым с помощью глубоких сверточных нейронных сетей.

[1] *Svetlov V., Persiantsev I., Shugay J., Dolenko S.* A new implementation of the algorithm of adaptive construction of hierarchical neural network classifier // Optical Memory and Neural Networks (Information Optics), 2015. Vol. 24. Pp. 288-294.

[2] *Dai Z., Liu H., Le Q., Tan M.* CoAtNet: Marrying Convolution and Attention for All Data Sizes // arXiv:2106.04803, 2021.

[3] *Frankle F., Carbin M.* CoAtNet: Marrying Convolution and Attention for All Data Sizes // arXiv:1803.03635, 2019.

[4] Tensorflow library // https://www.tensorflow.org/.

[5] CIFAR-10 dataset // https://www.cs.toronto.edu/~kriz/cifar.html.

[6] *Redmon J., Farhadi A.* YOLO9000: Better, Faster, Stronger // arXiv:1612.08242, 2016.

Convolutional hierarchical neural network classifier

Gadzhiev Ismail[1]⋆ ismailgadzhievff@gmail.com
Dolenko Sergey[2] 0000-0001-6214-3195
[1]Moscow, D.V.Skobeltsyn Institute of Nuclear Physics
[2]Moscow, M.V. Lomonosov Moscow State University

The report is devoted to the development and research of the convolutional modification of the hierarchical neural network classifier (HNNC) algorithm [1]. The algorithm performs recursive classification by building a class tree, in each node of which a subset of all classes is classified by a separate neural network (each child node contains a subset of the classes of the parent node). The construction of a class tree occurs in the training process by merging classes that the classifier in the node confuses most often. The algorithm is designed to combine the benefits of decision trees and neural networks, in particular, to reduce the number of parameters while maintaining sufficient complexity of the model.

The problem of reducing the number of parameters is especially relevant for computer vision tasks - models that produce the highest quality results for "difficult" tasks require optimization of a large number of parameters. For example, the best results on the ImageNet problem are shown in article [2]. The model used 2440 million parameters. However, in [3], a hypothesis is formulated that up to 90% of the neural network weights are uninformative. The authors choose a random subnet of the original network and get a similar quality of results.

Reducing the number of parameters can allow the effective use of the model on "weak" computers with limited memory and limited computing capabilities, for example, on mobile phones, which is very important.

We present the HNNC modification, that is an attempt to apply the HNNC to the image recognition problem by replacing the classifiers in the tree nodes with convolutional neural networks. The desired effect of using a class hierarchy is to reduce the number of model parameters. Unlike the original HNNC, training consists of two stages. At the first stage, a class tree is built with the use of "weak" convolutional networks with a small number of parameters; at the second stage, the convolutional networks at the nodes are replaced with stronger ones, with their training on subsets of the original set.

Let us describe the algorithm for constructing a class tree for convolutional HNNC (CHNNC) (first stage). At each node of the tree, there is a neural network that learns to classify patterns from a subset of all classes. In the process of training, the classes that the network confuses most often are combined.

To determine which classes the network confuses with each other, the procedure of "voting" of the patterns of each class is used; those classes are combined, most of the patterns in which "voted" in the same class. After voting, the desired answers for the merged classes are made the same, and the training of the network continues with the patterns re-labeled in this way. In cases where further merging of classes

is impossible, or if the specified maximum number of epochs has passed without merging, the training procedure stops.

As a result, the classes are combined into groups in accordance with how the network at the node confuses them, and the efficiency and quality of learning after the classes are combined significantly increase. Then, on each of the obtained groups, a new neural network is trained - in this way a class tree is recursively built.

As noted before, after this, the second stage of training is performed - the models in the nodes are replaced with more complex ones, and the training is repeated.

The algorithm is implemented using the Tensorflow library [4].

We use CIFAR-10 image classification task [5] to test the algorithm.

We also discuss methods to improve the quality of the CHNNC. One of the CHNNC hyperparameters is the threshold for neuron activation during voting (only those patterns vote for which the maximum activation of the output neurons exceeds the threshold). This hyperparameter is inherited from the original. The expediency of using the threshold for activations in the framework of convolutional modification is discussed.

It is important to emphasize that the class tree obtained at the first stage of CHNNC training can be reused for other tasks. For example, in the YOLO9000 model [6], designed for segmenting objects in images, a hierarchical loss function relative to classes is used, where errors are calculated at all levels of the class tree. The class hierarchy is built on the basis of the semantics of words denoting these classes, obtained from dictionaries. However, the semantic affinity of class labels does not imply visual affinity of the patterns. Therefore, it is expedient to be able to use the class hierarchy obtained adaptively from the data.

Further directions of research and improvement of the algorithm are also discussed, since the obtained results in terms of recognition quality are still far from ideal, much inferior to the results obtained using deep convolutional neural networks.

[1] *Svetlov V., Persiantsev I., Shugay J., Dolenko S.* A new implementation of the algorithm of adaptive construction of hierarchical neural network classifier // Optical Memory and Neural Networks (Information Optics), 2015. Vol. 24. Pp. 288-294.

[2] *Dai Z., Liu H., Le Q., Tan M.* CoAtNet: Marrying Convolution and Attention for All Data Sizes // arXiv:2106.04803, 2021.

[3] *Frankle F., Carbin M.* CoAtNet: Marrying Convolution and Attention for All Data Sizes // arXiv:1803.03635, 2019.

[4] Tensorflow library // `https://www.tensorflow.org/`.

[5] CIFAR-10 dataset // `https://www.cs.toronto.edu/~kriz/cifar.html`.

[6] *Redmon J., Farhadi A.* YOLO9000: Better, Faster, Stronger // arXiv:1612.08242, 2016.

Априорное распределение параметров в задачах выбора моделей глубокого обучения

Грабовой Андрей Валериевич[1]⋆ grabovoy.av@phyesstech.edu
Стрижов Вадим Викторович[1,2] strijov@phystech.edu
[1]Москва, Московский физико-технический институт (национальный исследовательский университет)
[2]Москва, Федеральный исследовательский центр «Информатика и управление» Российской академии наук (ФИЦ ИУ РАН)

В работе исследуется проблема снижения сложности аппроксимирующих моделей машинного обучения. Рассматриваются подходы, которые основанные на дистилляции моделей глубокого обучения. Предлагается вероятностное обоснование методов дистилляции и привилегированного обучения, а также обобщение классической дистилляции, используя связный байесовский вывод. Для снижения размерности пространства параметров при выборе модели используется информация об их априорном и апостериорном распределениях. Предложен метод назначения априорного распределения параметров ученика на основе апостериорного распределения параметров модели учителя. Пространства параметров учителя и ученика в общем случае не совпадают. Предлагается механизм приведения пространства параметров модели учителя к пространству параметров модели ученика путем выравнивания структуры модели учителя.

Предложен метод локального выравнивания структур в рамках одного семейства нейросетевых моделей: полносвязные сети, рекуррентные сети. Рассмотренные структуры семейств задаются числом слоев и размерностью промежуточных пространств после каждого слоя. Следовательно, каждая структура задается последовательностью натуральных чисел. Каждое число соответствует размерности промежуточного пространства, а длина последовательности соответствует числу слоев нейросети. Пространство структур является пространством последовательностей натуральных чисел.

В работе предложен метод локального выравнивания структуры с сохранением апостериорного распределения вектора параметров модели учителя. Вводиться множество структур, которое при помощи последовательности локальных преобразований получается из исходной структуры модели учителя.

Проводится теоретический анализ предложенного механизма сопоставления структур. Вычислительный эксперимент проводится на синтетических данных, а также на выборках FashionMNIST и Twitter Sentiment Analysis.

Настоящая работа содержит результаты проекта Математические методы интеллектуального анализа больших данных, выполняемого в рамках реализации Программы Центра компетенций Национальной технологической инициативы «Центр хранения и анализа больших данных», поддерживаемого Министерством науки и высшего образования Российской Федерации по Договору МГУ им. М.В. Ломоносова с Фондом поддержки проектов Национальной тех-

нологической инициативы от 11.12.2018 No13/1251/2018. Работа выполнена при поддержке Российского фонда фундаментальных исследований (проекты No19-07-01155, No 19-07-00875).

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

Prior distribution of parameters for the deep learning model selection problem

Grabovoy Andrey[1,2] grabovoy.av@phyesstech.edu
Strijov Vadim[2]⋆ strijov@phystech.edu
[1]Moscow, Moscow Institute of Physics and Technology (national research university)
[2]Moscow, Federal Research Center "Informatics and management" Russian Academy of Sciences

This research studies the approximating machine learning models' complexity reduction problem. It analyzes distillation methods for deep learning models. It proposes to generalize the original distillation with the probabilistic techniques. It generalizes this probabilistic technique for the case of Bayesian inference. The authors propose to reduce dimensionality of the parameter space using the prior and posterior distributions of the model parameters. They assign the prior distribution of the student parameters to the posterior distribution of the teacher parameters. The paper analyzes the case when the teacher and student parameter spaces do not match. To align the spaces the authors perform local transformations of the teacher's structure.

A method for the local alignment of structures was proposed. It includes the fully connected network and the recurrent networks. The model structure is specified by the number of layers and the number of dimensions of intermediate spaces after each layer. So any structure is specified by a sequence of natural numbers. Each number in the sequence corresponds to the dimension of intermediate space. But the length of the sequence corresponds to the number of layers in the neural network.

The paper proposes a method for local structure transformation. It keeps the posterior distribution of the teacher parameters. It introduces a set of structures, which starts with the initial structure of the teacher and follows with the sequence of local transformations.

The paper carries the theoretical analysis of the proposed structure alignment method. The computational experiment is carried out on synthetic data, as well as on FashionMNIST and Twitter Sentiment Analysis samples.

This article contains the results of the project Mathematical Methods for Mining Big Data, carried out within the framework of the Program of the Competence Center of the National Technological Initiative "Big Data Storage and Analysis Center", supported by the Ministry of Science and Higher Education of the Russian Federation under the Agreement of Moscow State University. M.V. Lomonosov with the Fund for Support of Projects of the National Technological Initiative of 11.12.2018 No.13 / 1251/2018. This work was supported by the Russian Foundation for Basic Research (projects No. 19-07-01155, No. 19-07-00875).

[1] *Grabovoy A., Strijov V.,* Bayesian Distillation For Neural Network // Automat. Remote Control, 2021.

Методика оценки степени несепарабельности функции

Горнов Александр Юрьевич[1]⋆ gornov@icc.ru
Зароднюк Татьяна Сергеевна[1] tz@icc.ru
[1]Иркутск, Институт динамики систем и теории управления им. В.М. Матросова СО РАН

Функцию принято называть сепарабельной, если ее можно представить как сумму функций одного переменного. На практике достаточно часто встречаются модели, описываемые близкими к сепарабельным функциями («квазисепарабельными»). Под степенью несепарабельности будем понимать среднее соотношение суммы модулей внедиагональных элементов к сумме модулей диагональных.

В докладе обсуждаются две методики оценки степени несепарабельности многомерной функции, опирающиеся на два недерминированных алгоритма. Первый алгоритм построен по «минимальному принципу»: с применением заданной вычислительной схемы стоится аппроксимация матрицы Гессе, прямой анализ которой, выполненный в наборе пробных точек, и дает усредненное значение искомого показателя. Второй алгоритм построен по «память-экономичной» схеме: в каждой пробной точке вычисляется градиент функции, производится смещение пробной точки на априори заданную величину вдоль направления градиента; в полученной точке снова вычисляется градиент; на разности градиентов и строится локальная оценка. Полученная по такой схеме оценка совпадает с оценкой, достигаемой первым алгоритмом, однако в данном случае нет необходимости хранить квадратичную матрицу.

Представляется, что предложенная методика может быть удобным инструментом для априорного изучения свойств моделей, применяемых в задачах анализа данных. Приводятся результаты вычислительных экспериментов.

Работа выполнена в рамках госзадания Минобрнауки России по проекту "Теория и методы исследования эволюционных уравнений и управляемых систем с их приложениями"(№ гос регистрации: 121041300060-4)

Technique for estimating the inseparability degree of a function

Gornov Alexander[1]⋆ gornov@icc.ru
Zarodnyuk Tatiana[1] tz@icc.ru
[1]Irkutsk, Matrosov Institute for System Dynamics and Control Theory of SB RAS

A function is usually called separable if it can be represented as a sum of functions of one variable. In practice, quite often there are models described by functions close to separable ("quasi-separable"). By the degree of inseparability we mean the average ratio of the sum of the modules of the off-diagonal elements to the sum of the modules of the diagonal ones.

The report discusses two methods for estimating the degree of the multidimensional function inseparability, based on two non-deterministic algorithms. The first algorithm is constructed according to the "minimal principle": using a given computational scheme, the approximation of the Hessian matrix is created, the direct analysis of which, performed in a set of sample points, gives the average value of the required indicator. The second algorithm is built according to the "memory-economical" scheme: at each test point, the gradient of the function is calculated, the test point is shifted by a predetermined value along the direction of the gradient; the gradient is calculated again at the obtained point; the difference between the gradients is used to construct the local estimate. The estimate obtained by this scheme coincides with the estimate achieved by the first algorithm, but in this case there is no need to store the quadratic matrix.

It seems that the proposed technique can be a convenient tool for a priori study of the models properties used in data analysis problems. The results of computational experiments are presented.

This work was carried out within the state assignment of the Ministry of Education and Science of Russia under the project "Theory and Methods of Research of Evolutionary Equations and Controlled Systems with their Applications" (State Registration No. 121041300060-4).

Детектирование периодических решений с помощью алгоритма BFOA на неполных картах Пуанкаре

Ручкин Константин construchk@gmail.com
Ручкин Александр alex3005r@gmail.com

В данной статье рассматривается задача нахождения периодических решений динамических систем [1]. Детектирование периодических решений проводится путем анализа плоских карт (сечений) Пуанкаре на наличие замкнутых траекторий специального типа - таких как круги, эллипсы и подобные фигуры. Специфика применения алгоритма построения сечений Пуанкаре такова, что данные траектории являются замкнутыми лишь на достаточно большом интервале времени работы алгоритма (а иногда и на бесконечном времени). При ограниченном времени эти траектории не являются замкнутыми, однако они должны быть идентифицированы как замкнутые траектории. Построению такого алгоритма посвящена данная работа.

Для оптимального поиска замкнутых траекторий мы используем эволюционный метод бактериальной поисковой оптимизации (BFOA). Построена целевая функция и разработана адаптивная версия алгоритма BFOA для поиска почти замкнутых траекторий по всему изображению карт Пуанкаре. Минимизация построенной целевой функции с помощью BFOA приводит к более быстрому и достаточно эффективному результату. Таким образом в работе показано, что распознавание периодических решений динамических систем возможно проводит численно-аналитическими методами за достаточно короткое время.

[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, 2014.

Detecting periodic solutions using the BFOA algorithm on the incomplete cards Poincare

Ruchkin Constantin construchk@gmail.com
Ruchkin Alexander alex3005r@gmail.com

This article discusses the problem of finding periodic solutions of dynamical systems. Detection of periodic solutions is carried out by analyzing Poincare plane maps (sections) for the including of closed trajectories of a special type, such as circles, ellipses, and similar figures. The specificity of the application of the algorithm for constructing Poincare sections is such that these trajectories are closed only over a sufficiently long time interval of the algorithm's operation (and sometimes over infinite time). For limited time, these paths are not closed, but they must be identified as closed paths. This work is devoted to the construction of such an algorithm.

For the optimal search for closed trajectories, we use the evolutionary bacterial search engine optimization (BFOA) method. An objective function has been constructed and an adaptive version of the BFOA algorithm has been developed to search for almost closed trajectories over the entire image of Poincare maps. The minimization of the constructed objective function using BFOA leads to a faster and more efficient result. Thus, it is shown in the work that the detection of periodic solutions of dynamical systems can be carried out by numerical-analytical methods for a shortly time.

[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, 2014.

Q-поиск: удачный метод для задачи безусловной минимизации

Горнов Александр Юрьевич[1,2] gornov@icc.ru

[1]Иркутск, Институт динамики систем и теории управления им. В.М. Матросова СО РАН

[2]Москва, Московский физико-технический институт (национальный исследовательский университет)

Практический успех метода оптимизации зависит от множества разнохарактерных факторов, многие из которых невозможно заранее предугадать и промерить. Приято считать «удачным» метод оптимизации, который показывает хорошие вычислительные результаты на многих разнообразных экстремальных задачах.

В докладе обсуждается развиваемый нами в последние годы эвристический алгоритм, названный – «Q-поиск». Основная идея подхода восходит к классическим работам Ю.Г. Евтушенко, в которых впервые применялась техника неравномерных покрытий, далее распространенная автором на многомерные невыпуклые задачи. В предложенном алгоритме нет накопления информационной базы проб и/или разбиений, что позволяет снять чрезмерно высокие требования к используемой памяти – конечно, за счет «потери гарантий». Другой «идеологической основой» можно считать поисковый метод Лууса–Яаколы или метод отсечения эллипсоидов, которые послужили моделью для механизма сжатия поисковых «брусов», позволяющего постепенно переходить от глобального сканирования к локальному уточнению. Сформированный алгоритм оснащен средствами «смягчения жесткости» отсечений, что позволяет понижать риски стартовой потери глобального решения, характерные для метода Лууса–Яаколы. В завершение конструкции организован двухуровневый вычислительный процесс, позволяющий использовать периодический перезапуск по формату «basin hopping», что также существенно повышает надежность достижения результата в невыпуклых задачах. Алгоритм требует всего одной пробы на итерации сжатия и, следовательно, максимально «информационно эффективен». Приводятся результаты вычислительных экспериментов.

Работа выполнена при поддержке Минобрнауки России (№ гос. регистрации: 121041300060-4).

Q-Search: a successful method for the unconstrained minimization problem

Gornov Alexander[1,2] gornov@icc.ru
[1]Irkutsk, Matrosov Institute for System Dynamics and Control Theory of SB RAS
[2]Moscow Institute of Physics and Technology (National Research University)

The practical success of the optimization method depends on many factors, many of which cannot be predicted and measured in advance. It is considered an optimization method "successful", which shows good computational results on many different extremal problems.

The report discusses a heuristic algorithm, which we have been developing in recent years, called "Q-search". The main idea of the approach goes back to the classical works of Yu.G. Evtushenko, in which the technique of nonuniform coverings was first applied, which was further extended by the author to multidimensional nonconvex problems. The proposed algorithm does not accumulate the information base of samples and/or partitions, which made it possible to remove excessively high requirements for the used memory (of course, due to the "loss of guarantees"). Another "ideological basis" can be considered the Luus–Jaakola search method or the ellipsoid cut-off method, which served as a model for the compression mechanism of the search "boxes". This technique allows a gradual transition from global scanning to local refinement. The constructed algorithm is equipped with a means of "softening the rigidity" of cutoffs, which makes it possible to reduce the risks of the initial loss of the global solution. Such risks are typical for the Luus–Yaakola method. At the end a two-level computational process is organized, which allows the use of a periodic restart in the "basin hopping" format. It also significantly increases the reliability of achieving results in non-convex problems. The algorithm requires only one sample per compression iteration and, therefore, is "informationally efficient" as much as possible. The results of computational experiments are presented.

This work was supported by the Ministry of Education and Science of the Russian Federation (state registration number: 121041300060-4).

Модификация метода LBFGS с экономичным одномерным поиском

Аникин Антон Сергеевич anikin@icc.ru
Иркутск, ИДСТУ СО РАН

На текущий момент времени имеется весьма развитый математический аппарат, применимый для решения задач выпуклой оптимизации. Но, для задач других классов, к сожалению, ситуация не столь благоприятна как с точки зрения теории, так и с точки зрения практики. При рассмотрении выпуклых задач мы, как правило, считаем, что обладаем определённой информацией о свойствах оптимизируемой функции, которая и позволяет применять оптимальные с точки зрения оценок методы. При решении реальных прикладных задач подобной информации, к сожалению, практически никогда нет. Например, априори неизвестно значение констант Липшица L, что приводит к отсутствию «простого» способа выполнения градиентного шага:

$$\alpha^k = \frac{1}{L}, \quad x^{k+1} = x^k - \alpha^k \nabla f(x^k),$$

Вместо этого приходится запускать процедуру одномерной минимизации, т.е. решать вспомогательную задачу оптимизации на каждой итерации градиентного метода:

$$\alpha^k = \arg\min_{\alpha \geqslant 0} f(x^k - \alpha \nabla f(x^k)). \tag{1}$$

Очевидно, что в общем случае это достаточно трудоёмкая процедура, эффективность которой радикальным образом влияет быстродействие оптимизационного метода в целом.

В работе предлагается и исследуется несколько вариантов экономичных алгоритмов одномерной минимизации: на основе параболической аппроксимации (с использованием и без использования значения градиента в текущей точке), «адаптивная» схема (с возможностью увеличения и уменьшения шага), «сжимающая» схема (с возможностью только уменьшения шага). Общей идеей предложенных подходов является максимально возможное уменьшение числа вычислений оптимизируемой функции при, естественно, сохранении свойства монотонности. Фактически, предлагается существенное снижение сложности итерации (и, соответственно, времени) ценой более грубого решения задачи одномерного поиска. Низкая стоимость итерации имеет важность как в современных задачах локальной (унимодальной) оптимизации большой и сверхбольшой размерности, так и в задачах глобального поиска, при решении которых, как правило, множество раз запускается процедура локального спуска.

Представлены результаты вычислительных экспериментов для ряда модельных и прикладных задач оптимизации: минимизация потенциалов Морса и Китинга, решение задачи PageRank, а также обучение свёрточных нейросетей. Полученные результаты продемонстрировали работоспособность и эффективность

предложенных подходов, при этом наилучшие результаты были достигнуты на модифицированном варианте широко известного метода LBFGS [1]. Представленная модификация метода включает в себя «правильное» масштабирование направления одномерного поиска, использование истории шагов одномерного поиска и ряд других изменений, направленных на создание «комфортных условий» для предложенных процедур одномерной минимизации. Это позволило на большинстве рассмотренных задач иметь среднюю сложность приближенного («грубого») решения задачи (1) порядка 2-3 вычислений функции, а в отдельных случаях и вовсе порядка 1, когда предложенная процедура одномерной минимизации почти всегда правильно «угадывает» размер шага, обеспечивающий релаксацию оптимизируемой функции. Очевидно, что качество (релаксация, $f(x^k) - f(x^{k+1})$) такого решения может быть ощутимо хуже, чем в случае «честной» минимизации, но результаты вычислительных экспериментов показали, что за один и тот же промежуток времени как правило выгоднее сделать несколько «простых» итераций, вместо одной «сложной».

Полученные результаты внушают осторожный оптимизм и позволяют надеяться на возможность эффективного применения предложенных подходов как для решения других актуальных задач оптимизации, так и для реализации других «экономичных» методов оптимизации.

[1] *Liu D., Nocedal J.* On the Limited Memory Method for Large Scale Optimization // Mathematical Programming B, 1989. Vol. 45. Pp. 503–528.

Modification of the LBFGS method with economical line-search

Anikin Anton anikin@icc.ru
Irkutsk, ISDCT SB RAS

At the current time, there is an advanced mathematical tools applicable to solving convex optimization problems. But, unfortunately, for problems of other classes, the situation is not so favorable both from the point of view of theory and from the point of view of practice. When considering convex problems, we, as a rule, believe that we have certain information about the properties of the optimized function, which allows us to apply estimates-optimal methods. Unfortunately, there is almost never such information when solving real applied problems. For example, the value of the Lipschitz constants L is unknown a priori, which leads to the absence of a "simple" way to perform a gradient step:

$$\alpha^k = \frac{1}{L},$$

$$x^{k+1} = x^k - \alpha^k \nabla f(x^k),$$

Instead, we have to run a line-search procedure, i.e. solve an auxiliary optimization problem at each iteration of the gradient method:

$$\alpha^k = \arg\min_{\alpha \geqslant 0} f(x^k - \alpha \nabla f(x^k)). \tag{1}$$

Obviously, in the general case, this is a rather time-consuming procedure, the effectiveness of which radically affects the speed of the optimization method as a whole.

Several variants of economical line-search algorithms are proposed and investigated: parabolic approximation-based (with using and without using the gradient value at the current point), "adaptive" scheme (with the possibility of increasing and decreasing the step), "reduction" scheme (with the possibility of only decreasing the step). The general idea of the proposed approaches is to reduce the number of computations of the optimized function as much as possible while, of course, maintaining the monotonicity property. In fact, a significant reduction in iteration complexity (and, accordingly, time) is proposed at the cost of a more rough solution to the line-search problem. The low cost of iteration is important both for modern large- and huge-scale local (unimodal) problems, and for global optimization problems, the solution of which, as a rule, requires multiple runs of the local descent procedure.

The results of computational experiments for a number of model and applied optimization problems are presented: minimizing the Morse and Keating potentials, solving the PageRank problem, as well as training some convolutional neural networks. The obtained results demonstrate the operability and effectiveness of the proposed approaches, while the best results were achieved with a modified version of the well-known LBFGS method [1]. The presented modification of the method includes "accurate" scaling of the line-search direction, using the history of line-search

steps and a number of other changes aimed at creating “comfortable conditions” for the proposed line-search procedures. All this made it possible to reduce the average complexity of the approximate (“rough”) solution of the problem (1) to 2-3 calculations of the function, and in some cases even to ≈ 1, when the proposed line-search procedure almost always correctly “guesses” the step size that provides relaxation of the optimized function. Obviously, the quality (relaxation, $f(x^k) - f(x^{k+1})$) of such solutions may be significantly worse than in the case of ¡¡honest¿¿ minimization, but the results of computational experiments have shown that it is usually more profitable to make several “cheap” (fast) iterations in the same period of time, instead of one “expensive” (slow).

The obtained results inspire conservative optimism and allow us to hope for the possibility of effective application of the proposed approaches both for solving other modern applied optimization problems and for implementing other “economical” optimization methods.

[1] *Liu D., Nocedal J.* On the Limited Memory Method for Large Scale Optimization // Mathematical Programming B, 1989. Vol. 45. Pp. 503–528.

Вычислительная сложность задачи цензурирования данных

Кутненко Ольга Андреевна[1,2]⋆ olga@math.nsc.ru
Плясунов Александр Владимирович[1,2] apljas@math.nsc.ru
[1]Новосибирск, Институт математики им. С. Л. Соболева
[2]Новосибирск, Новосибирский гос. университет

Развитие технологий в современном мире приводит к экспоненциальной скорости роста объема информации в самых разных областях, что дает, с одной стороны, новые возможности для решения различных прикладных задач, но, с другой стороны, повышает риск появления ошибок в анализируемых данных. Поэтому проблема цензурирования данных (Data filtering, Data cleaning) в настоящее время актуальна при решении самых разных задач.

Рассматривается задача очистки обучающей выборки, представленной объектами двух классов, от шумовых объектов только одного класса. Такие задачи возникают, в частности, при анализе биомедицинских данных, требующем полного сохранения данных одного из образов. Исключение из обучающей выборки неверно классифицированных объектов (или объектов-выбросов) осуществляется на основе анализа локального окружения объектов. Данный подход опирается на гипотезу локальной компактности. Количественная характеристика локальной компактности образа оценивается с помощью функции конкурентного сходства, успешно используемой в когнитивном анализе данных при решении различных прикладных задач.

Цензурирование объектов-выбросов с помощью функции конкурентного сходства. Для получения количественной оценки компактности образов в фиксированном признаковом пространстве используется FRiS-функция (Function of Rival Similarity)[1], с помощью которой формализуется представление о компактности как о «высоком» сходстве объектов одного образа друг с другом и «низком» сходстве с объектами других образов.

Для решения рассматриваемой задачи необходимо при сохранении объектов образа $\mathbf{B}$ найти множество $\mathbf{A}'$ удаленных объектов образа $\mathbf{A}$ или, соответственно, множество $\mathbf{A}^* = \mathbf{A} \setminus \mathbf{A}'$ оставшихся объектов, на котором достигается максимум FRiS-компактности образа $\mathbf{A}$:

$$H_{\mathbf{A}|\mathbf{B}}(\mathbf{A}^*) = \frac{1}{|\mathbf{A}|} \sum_{a \in \mathbf{A}^*} \frac{\tau(a, \mathbf{B}) - \tau(a, \mathbf{A}^*)}{\tau(a, \mathbf{B}) + \tau(a, \mathbf{A}^*)}.$$

В качестве метрики τ используется среднее расстояние от объекта до $k \geqslant 1$ ближайших объектов образа.

Вычислительная сложность задачи. Доказательство NP-трудности в сильном смысле задачи цензурирования данных выполнено сведением известной NP-полной задачи о вершинном покрытии графа к задаче выбора подмножества, на котором компактность образа максимальна.

Задача ВП (вершинное покрытие)[2]. Дан граф $G = (V, E)$ и положительное целое число $J \leqslant |V|$. Имеется ли в графе G вершинное покрытие не более чем из J элементов, то есть такое подмножество $V' \subseteq V$, что $|V'| \leqslant J$ и для каждого ребра $\{u, v\} \in E$ хотя бы одна из вершин u или v принадлежит V'?

В [3] доказана следующая

Теорема. Задача поиска наименьшего вершинного покрытия произвольного графа $G = (V, E)$ сводится к задаче выбора из некоторой искусственной выборки X_G множества объектов $\mathbf{A}^* \subseteq \mathbf{A}$, на котором достигается максимум функционала H. Причем выборка X_G строится по G за полиномиальное время и имеет полиномиальное количество объектов относительно $|V| + |E|$. И в $\mathbf{A}^*$ содержится не менее $(k + 1)$ объектов образа $\mathbf{A}$.

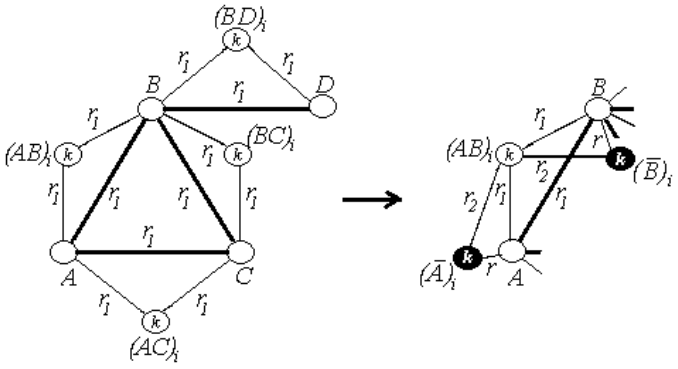


Рис. 1.

На рис. 1 приведено построение искусственной выборки X_G по графу G, состоящему из четырех вершин. Расстояния между объектами внутри групп черных и белых объектов равны r и r_2, соответственно. Для r и r_2 при любом $k \in \mathbb{N}$ и любом положительном $r_1 \in \mathbb{Q}$ выполняются следующие неравенства:

$$0 < r < r_1 < r_2 < 2r, \qquad \frac{r - r_1}{r + r_1} + \frac{r_2 - r_1}{2kr_2 - r_2 + r_1} > 0.$$

Так как $\mathbf{A}' = \mathbf{A} \setminus \mathbf{A}^*$, то из доказанной теоремы получим

Следствие. Задача поиска $\mathbf{A}' = \mathbf{A} \setminus \mathbf{A}^*$ — множества удаляемых объектов-выбросов образа $\mathbf{A}$ является NP-трудной.

Показана NP-трудность в сильном смысле экстремальной задачи поиска множества, на котором достигается согласно заданному критерию максимум оценки компактности образа. Что означает труднорешаемость соответствующей проблемы анализа данных и обосновывает применение различных эвристических алгоритмов для решения задач цензурировния данных, в которых в

качестве количественной оценки компактности образов используется функция конкурентного сходства [4, 5].

Работа выполнена в рамках государственного задания ИМ СО РАН (проекты № 0314–2019–0015, № 0314–2019–0014).

[1] *Zagoruiko N., Borisova I., Dyubanov V., Kutnenko O.* Methods of recognition based on the function of rival similarity // Pattern Recognition and Image Analisys, 2008. Vol. 18(1). Pp. 1–6.

[2] *М. Гэри, Д. Джонсон.* Вычислительные машины и труднорешаемые задачи. М: Мир, 1982. 416 с.

[3] *Кутненко О. А., Плясунов А. В.* NP-трудность некоторой задачи цензурирования данных // Дискретный анализ и исследование операций, 2021. Т. 28(2). С. 34–47.

[4] *Борисова И. А., Кутненко О. А.* Исправление диагностических ошибок в целевом признаке с помощью функции конкурентного сходства // Математическая биология и биоинформатика, 2018. Т. 13(1). С. 38–49.

[5] *Загоруйко Н. Г., Кутненко О. А.* Цензурирование обучающей выборки // Вестн. Том. гос. ун-та: Управление, вычислительная техника и информатика, 2013. № 22. С. 66—73.

Computational complexity of data cleaning problem

Kutnenko Olga[1,2]⋆ olga@math.nsc.ru
Plyasunov Alexander[1,2] apljas@math.nsc.ru
[1]Novosibirsk, Sobolev Institute of Mathematics
[2]Novosibirsk, Novosibirsk State University

In the modern world the development of technologies leads to the exponential rate of the information volume growth in the variety of areas, which on one hand provides new opportunities for solving various applied problems but on the other hand increases the risk of error appearance in the analyzed data. Therefore, the problem of data cleaning is currently relevant in solving a variety of problems.

The problem of cleaning the training sample, represented by the objects of two classes, from noise objects of only one class is considered. Problems of this type arise, in particular, in the analysis of the biomedical data which requires the complete preservation of the data of one of the images. The elimination of incorrectly classified objects (or outlier objects) from the training sample is carried out by the analysis of the local neighborhood of the objects. This approach is based on the hypothesis of local compactness. The local compactness of the image is evaluated quantitatively using the function of rival similarity which is successfully used in cognitive data analysis for solving various applied problems.

Cleaning outliers by the function of rival similarity. To obtain a quantitative estimate of the compactness of images in a fixed feature space, the FRiS-function (Function of Rival Similarity)[2], is used to formalize the idea of compactness as "high" similarity of the objects of one image and their "low" similarity to the objects of other images.

To solve the problem under consideration, we need, while preserving the objects of the image $\mathbf{B}$, to find the set $\mathbf{A}^{'}$ of deleted objects of the image $\mathbf{A}$ or, respectively, the set $\mathbf{A}^{*} = \mathbf{A} \setminus \mathbf{A}^{'}$ of the remaining objects, at which the maximum of FRiS-compactness of the image $\mathbf{A}$ is attained:

$$H_{\mathbf{A}|\mathbf{B}}(\mathbf{A}^{*}) = \frac{1}{|\mathbf{A}|} \sum_{a \in \mathbf{A}^{*}} \frac{\tau(a, \mathbf{B}) - \tau(a, \mathbf{A}^{*})}{\tau(a, \mathbf{B}) + \tau(a, \mathbf{A}^{*})}.$$

As the metric τ, we use the average distance to $k \geqslant 1$ nearest objects of the image.

Computational complexity of the problem. We carry out the proof of NP-hardness in the strong sense of the data censoring problem by reducing the well-known NP-complete vertex cover problem of a graph to the problem of selecting a subset on which the compactness of the image will attain the maximum.

Problem VC (vertex cover)[3]. A graph $G = (V, E)$ and a positive integer number $J \leqslant |V|$ are given. Is there a vertex cover of at most J elements in G, i.e., a subset $V^{'} \subseteq V$ such that $|V^{'}| \leqslant J$ and for each edge $\{u, v\} \in E$ at least one vertex u or v belongs $V^{'}$?

In [3] we proved the following

Theorem. The problem of finding the minimum vertex cover of an arbitrary graph $G = (V, E)$ is reduced to the problem of choosing some set of objects $\mathbf{A}^* \subseteq \mathbf{A}$ from some artificial sample X_G at which the maximum of the functional H is attained. In this case, the sample X_G is constructed from G in polynomial time and has polynomially many objects with respect to $|V| + |E|$, while $\mathbf{A}^*$ contains at least $(k + 1)$ objects of the image $\mathbf{A}$.

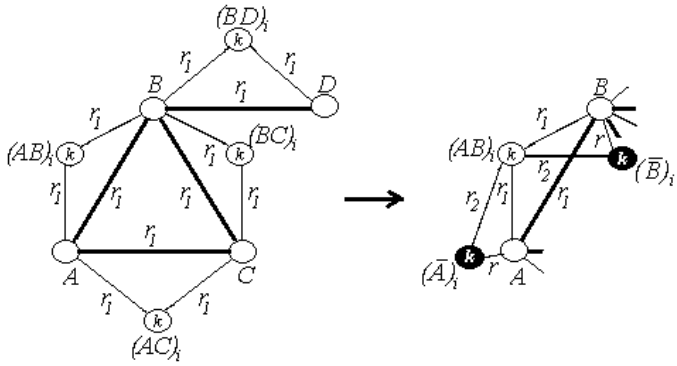


Fig. 1.

Figure 1 shows the construction of an artificial sample X_G over a graph G consisting of four vertices. The distances between objects within the groups of black and white objects are equal to r and r_2, respectively. For r and r_2, for any $k \in \mathbb{N}$ and any positive $r_1 \in \mathbb{Q}$, the following inequalities hold:

$$0 < r < r_1 < r_2 < 2r, \qquad \frac{r - r_1}{r + r_1} + \frac{r_2 - r_1}{2kr_2 - r_2 + r_1} > 0.$$

Since $\mathbf{A}' = \mathbf{A} \setminus \mathbf{A}^*$, the above theorem yields

Corollary. The problem of searching for the set $\mathbf{A}' = \mathbf{A} \setminus \mathbf{A}^*$ of the eliminated outliers of the image $\mathbf{A}$ is NP-hard.

We show NP-hardness in the strong sense of the extremal problem of searching for a set on which the maximum estimate of the compactness of the image is achieved according to a specified criterion. That means the intractability of the corresponding problem of data analysis and justifies the usage of various heuristic algorithms for solving data censoring problems, in which the function of rival similarity is used as a quantitative assessment of the compactness of images [5, 6].

The authors were supported by the State Task to the Sobolev Institute of Mathematics (projects nos. 0314–2019–0015 and 0314–2019–0014).

[1]

[2] *Zagoruiko N., Borisova I., Dyubanov V., Kutnenko O.* Methods of recognition based on the function of rival similarity // Pattern Recognition and Image Analisys, 2008. Vol. 18(1). Pp. 1–6.

[3] *Garey M., Johnson D.* Computers and Intractability: A Guide to the Theory of NP-Completeness. Mir. Moscow, 1982. 416 p.

[4] *Kutnenko O., Plyasunov A.* NP-hardness of some data cleaning problem // Diskretnyi analiz i issledovanie operatsii, 2021. Vol. 28(2). Pp. 34–47.

[5] *Borisova I., Kutnenko O.* The Problem of Correction Diagnostic Errors in the Target Attribute with the Function of Rival Similarity // J. Math. Biology and Bioinform, 2018. Vol. 13 (1). Pp. 38–49.

[6] *Zagoruiko N., Kutnenko O.* Censoring of the Training Sample // Vestnik Tomsk. Gos. Univer: Upravl, Vychisl. Tekhn. i Inform, 2013. Vol. 1(22). Pp. 66–73.

Достаточные условия полиномиальной сложности решения интервальных систем линейных алгебраических уравнений в задачах построения линейных зависимостей с интервальной неопределенностью данных

Ерохин Владимир Иванович[1]⋆ erohin_v_i@mail.ru
Кадочников Андрей Павлович[1] kado162@mail.ru
Сотников Сергей Владимирович[1] svsotnikov66@gmail.com

[1]Санкт-Петербург, Военно-космическая академия имени А.Ф. Можайского

Интервальные системы линейных алгебраических уравнений (ИСЛАУ) являются естественным инструментом создания моделей и алгоритмов обработки данных с интервальной неопределенностью. В общем случае поиск *слабых решений* ИСЛАУ является NP-трудной задачей [1], что сдерживает их широкое внедрение в практику моделирования и анализа данных. В то же время, как часто показывает решение практических (инженерных) задач построения линейных зависимостей по экспериментальным данным с интервальной неопределенностью, допустимое множество переопределенной ИСЛАУ оказывается 1) выпуклым многогранником, целиком лежащем в некотором ортанте n-мерного пространства и 2) с ростом числа экспериментов стягивающемся в точку, совпадающую с истинным вектором коэффициентов линейной модели. Свойство 2) является аналогом свойства состоятельности статистической модели, в то время как свойство 1) во-первых, гарантирует полиномиальную трудоемкость поиска слабых решений ИСЛАУ (с использованием методов линейного программирования), и во-вторых, является аналогом свойства статистической значимости коэффициентов статистической линейной модели. В докладе будут предложены неизвестные ранее легко проверяемые достаточные условия принадлежности допустимого множества конкретной ИСЛАУ множеству выпуклых многогранников, целиком лежащих в некотором ортанте, которые одновременно являются достаточными условиями полиномиальной сложности решения данной ИСЛАУ.

Пусть ИСЛАУ задана совокупностью условий

$$Ax = b,\ \underline{A} \leqslant A \leqslant \bar{A},\ \underline{b} \leqslant b \leqslant \bar{b}, \tag{1}$$

где $\underline{A}, \bar{A} \in \mathbb{R}^{m\times n}$ – заданные матрицы; $\underline{b}, \bar{b} \in \mathbb{R}^m$ – заданные векторы, такие что $\underline{A} \leqslant \bar{A}$, $\underline{b} \leqslant \bar{b}$; $A \in \mathbb{R}^{m\times n}$, $x \in \mathbb{R}^n$, $b \in \mathbb{R}^m$ – неизвестные (подлежащие определению) матрица и векторы, $\underline{A} \neq \bar{A}$, $\underline{b} \neq \bar{b}$, $m > n$. Заметим, что в большинстве прикладных исследований в центре внимания оказывается только *допустимое множество* ИСЛАУ, определяемое как $\mathbf{X} \triangleq \left\{ x \,\middle|\, \exists \underline{A} \leqslant A \leqslant \bar{A}, \underline{b} \leqslant b \leqslant \bar{b}, Ax = b \right\}$.

Эквивалентное (1) представление ИСЛАУ может быть записано с помощью *средней матрицы* $A_c = \frac{1}{2}(\underline{A} + \bar{A})$, *матрицы радиусов* $A_r = \frac{1}{2}(\bar{A} - \underline{A})$, *среднего вектора* $b_c = \frac{1}{2}(\underline{b} + \bar{b})$ и *вектора радиусов* $b_r = \frac{1}{2}(\bar{b} - \underline{b})$ [1].

ИСЛАУ представляются естестественной моделью построения линейных зависимостей по данным, обладающим интервальной неопределенностью (см., например [2, 3, 4]).

Пусть $\hat{x} = A_c^+ b_c$ – нормальное псевдорешение несовместной переопределенной системы $A_c x \cong b_c$, $\Delta b_c = b_c - A_c \hat{x}$ – её невязка с минимальной евклидовой нормой, A_c^+ – соответствующая псевдообратная матрица, и выполняются условия $A_c^1 \hat{x} < b_c^1$, $-A_c^2 \hat{x} < -b_c^2$, где с точностью до некоторой перестановки строк A_c и элементов b_c

$$A_c = \begin{bmatrix} A_c^1 \\ A_c^2 \end{bmatrix},\ b_c = \begin{bmatrix} b_c^1 \\ b_c^2 \end{bmatrix}.$$

Введем обозначения:

$$\tilde{A}_c \triangleq \begin{bmatrix} A_c^1 \\ -A_c^2 \end{bmatrix},\ \tilde{b}_c \triangleq \begin{bmatrix} b_c^1 \\ -b_c^2 \end{bmatrix},\ S \triangleq \operatorname{diag}(\operatorname{sign}(\hat{x})),$$

$$\widehat{\mathbf{X}} \triangleq \{x \,|(A_c - A_r S)x \leqslant b_c + b_r,\ (-A_c - A_r S)x \leqslant -b_c + b_r\},$$

$\mathbf{1}$ – n-мерный вектор, состоящий из единиц,

$\sigma_{\min}^{A_c}$ – минимальное сингулярное число матрицыA_c ,

$\sigma_{\max}^{A_r}$ – максимальное сингулярное число матрицыA_r,

$||\cdot||$ обозначает евклидову векторную норму.

Справедлива следующая

Теорема 1. *Пусть выполняются условия* $\operatorname{rank}(A_c) = n$, $\sigma_{\min}^{A_c} > \sigma_{\max}^{A_r}$, $S\hat{x} > 0$, $\min\limits_{j=1,\ldots,n} |\hat{x}_j| > \frac{1}{\sigma_{\min}^{A_c} - \sigma_{\max}^{A_r}} \left(\sigma_{\max}^{A_r} ||\hat{x}|| + \frac{||\Delta b_c||}{\sigma_{\min}^{A_c}} + ||b_r|| \right)$, *системы неравенств*

$$(A_c - A_r S)x \leqslant b_c + b_r, \tag{2}$$

$$(\tilde{A}_c + A_r S)x \leqslant \tilde{b}_c + b_r \tag{3}$$

совместны и существует скаляр $\delta > 0$ *такой, что система линейных неравенств* $Sx \geqslant \mathbf{1}\delta$, *является следствием как системы* (2) *так и системы* (3). *Тогда*

1. *Все* 2^n *систем линейных неравенств*

$$(\tilde{A}_c - A_r \tilde{S})x \leqslant \tilde{b}_c + b_r, \tag{4}$$

где $\tilde{S}$ *– диагональная матрица порядка* n *с элементами* ± 1 *на диагонали, совместны.*

2. *Существует скаляр* $\tilde{\delta}$ *такой, что система линейных неравенств* $Sx \geqslant \mathbf{1}\tilde{\delta}$, *является следствием любой системы линейных неравенств вида* (4).

3. *Множество* $\mathbf{X}$ *совпадает с множеством* $\widehat{\mathbf{X}}$ *и, в случае непустоты, представляет собой выпуклый многогранник, целиком лежащий строго внутри ортанта, определяемого знаками диагональных элементов матрицы* S, *т.е.,* $\forall x \in \mathbf{X} \Rightarrow Sx > 0$.

Несложно заметить, что условия теоремы 1 проверяются за полиномиальное время средствами вычислительной линейной алгебры и линейного программирования.

[1] *Фидлер М.*, *Недома Й.*, *Рамик Я.*, *Рон И.*, *Циммерман К.* Задачи линейной опимизации с неточными данными // М. Ижевск: НИЦ «Регулярная и хаотическая динамика». Институт компьютерных исследований, 2008. 288 с.

[2] *Вощинин А. П.*, *Боков А. Ф.*, *Сотиров Г. Р.* Метод анализа данных при интервальной нестатистической ошибке // Завод. лаб., 1990. Т. 56(7). С. 76–81.

[3] *Белов В. М.*, *Суханов В. А.*, *Лагуткина Е. В.*, Интервальный подход при решении задач кинетики простых химических реакций // Вычисл. технологии, 1997. Т. 2(1). С. 10–18.

[4] *Шарый С. П.* Задача восстановления зависимостей по данным с интервальной неопределенностью // Заводская лаборатория. Диагностика материалов, 2020. Т. 86(1). С. 62–74.

Sufficient conditions for the polynomial complexity of solving interval systems of linear algebraic equations in problems of constructing linear dependencies with interval uncertainty of data

Erokhin Vladimir[1]⋆ erohin_v_i@mail.ru
Kadochnikov Andrey[1] kado162@mail.ru
Sotnikov Sergey[1] svsotinkov66@gmail.com
[1]Saint-Petersburg, A.F. Mozhaysky Military-Space Academy

Interval systems of linear algebraic equations (ISLAE) are a natural tool for creating models and algorithms for processing data with interval uncertainty. In general, the search for *weak solutions* is a NP-hard problem [1], which hinders their widespread introduction into the practice of modeling and data analysis. At the same time, as the solution of practical (engineering) problems of constructing linear dependencies based on experimental data with interval uncertainty often shows, the permissible set of the redefined ISLAE turns out to be 1) a convex polyhedron lying entirely in some orthant of the n-dimensional space and 2) with an increase in the number of experiments, shrinking to a point coinciding with the true vector of the coefficients of the linear model. Property 2) is an analogue of the consistency property of the statistical model, while property 1) firstly, guarantees the polynomial complexity of finding weak solutions to the problem (using linear programming methods), and secondly, is an analogue of the statistical significance property of the coefficients of the statistical linear model. The report will propose previously unknown, easily verifiable sufficient conditions for the membership of an admissible set of a particular ISLAE to a set of convex polyhedra lying entirely in some orthant, which at the same time are sufficient conditions for the polynomial complexity of solving this ISLAE.

Let ISLAE be given by a set of conditions

$$Ax = b,\ \underline{A} \leqslant A \leqslant \bar{A},\ \underline{b} \leqslant b \leqslant \bar{b}, \tag{1}$$

where $\underline{A}, \bar{A} \in \mathbb{R}^{m\times n}$ are the given matrices; $\underline{b}, \bar{b} \in \mathbb{R}^m$ are given vectors such that $\underline{A} \leqslant \bar{A}$, $\underline{b} \leqslant \bar{b}$; $A \in \mathbb{R}^{m\times n}$, $x \in \mathbb{R}^n$, $b \in \mathbb{R}^m$ are unknown (to be determined) matrix and vectors, $\underline{A} \neq \bar{A}$, $\underline{b} \neq \bar{b}$, $m > n$. Note that in most applied research, the focus is only on the *allowable set* of the ISLAE, defined as $\mathbf{X} \triangleq \left\{ x \,\middle|\, \exists \underline{A} \leqslant A \leqslant \bar{A}, \underline{b} \leqslant b \leqslant \bar{b}, Ax = b \right\}$.

An equivalent (1) representation of ISLAE can be written using *middle matrix* $A_c = \frac{1}{2}(\underline{A} + \bar{A})$, *radius matrices* $A_r = \frac{1}{2}(\bar{A} - \underline{A})$, *middle vector* $b_c = \frac{1}{2}(\underline{b} + \bar{b})$ and *radius vectors* $b_r = \frac{1}{2}(\bar{b} - \underline{b})$ [1]. To ISLAE seem to be a natural model for constructing linear dependencies based on data with interval uncertainty (see, for example [2, 3, 4]).

Let $\hat{x} = A_c^+ b_c$ be a normal pseudo-solution of an incompatible redefined system $A_c x \cong b_c$, $\Delta b_c = b_c - A_c \hat{x}$ is its residual with the minimum Euclidean norm, A_c^+ is the corresponding pseudo-return matrix and the conditions $A_c^1 \hat{x} < b_c^1$, $-A_c^2 \hat{x} < -b_c^2$, where with accuracy up to some permutation of the rows of A_c and elements of b_c

$$A_c = \begin{bmatrix} A_c^1 \\ A_c^2 \end{bmatrix}, \; b_c = \begin{bmatrix} b_c^1 \\ b_c^2 \end{bmatrix}.$$

Let's introduce the notation:

$$\tilde{A}_c \triangleq \begin{bmatrix} A_c^1 \\ -A_c^2 \end{bmatrix}, \; \tilde{b}_c \triangleq \begin{bmatrix} b_c^1 \\ -b_c^2 \end{bmatrix}, \; S \triangleq \operatorname{diag}(\operatorname{sign}(\hat{x})),$$

$$\widehat{\mathbf{X}} \triangleq \{x \,|(A_c - A_r S)x \leqslant b_c + b_r, \; (-A_c - A_r S)x \leqslant -b_c + b_r\},$$

$\mathbf{1}$ is *ann*-dimensional vector consisting of units,

$\sigma_{\min}^{A_c}$ is the minimum singular number of the matrixA_c ,

$\sigma_{\max}^{A_r}$ is the maximum singular number of the matrixA_r,

$||\cdot||$ denotes the Euclidean vector norm.

The following is true

Theorem 1. *Let the conditions be fulfilled* $\operatorname{rank}(A_c) = n$, $\sigma_{\min}^{A_c} > \sigma_{\max}^{A_r}$, $S\hat{x} > 0$, $\min\limits_{j=1,\ldots,n} |\hat{x}_j| > \frac{1}{\sigma_{\min}^{A_c} - \sigma_{\max}^{A_r}} \left(\sigma_{\max}^{A_r} ||\hat{x}|| + \frac{||\Delta b_c||}{\sigma_{\min}^{A_c}} + ||b_r|| \right)$, *systems of inequalities*

$$(A_c - A_r S)x \leqslant b_c + b_r, \tag{2}$$

$$(\tilde{A}_c + A_r S)x \leqslant \tilde{b}_c + b_r \tag{3}$$

are compatible and there exists a scalar $\delta > 0$ *such that the system of linear inequalities* $Sx \geqslant \mathbf{1}\delta$ *is a consequence of both the system* (2) *and the system* (3). *Then*

1. *All* 2^n *systems of linear inequalities*

$$(\tilde{A}_c - A_r \tilde{S})x \leqslant \tilde{b}_c + b_r, \tag{4}$$

where $\tilde{S}$ *is a diagonal matrix of order* n *with elements* ± 1 *on the diagonal, are compatible.*

2. *There exists a scalar* $\tilde{\delta}$ *such that the system of linear inequalities* $Sx \geqslant \mathbf{1}\tilde{\delta}$ *is a consequence of any system of linear inequalities of the form* (4).

3. *The set* $\mathbf{X}$ *coincides with the set* $\widehat{\mathbf{X}}$ *and, in the case of non-emptiness, is a convex polyhedron lying entirely strictly inside the orthant defined by the signs of the diagonal elements of the matrix* S, *i.e.,* $\forall x \in \mathbf{X} \Rightarrow Sx > 0$.

It is easy to notice that the conditions of the theorem 1 are checked in polynomial time by means of computational linear algebra and linear programming.

[1] *Fiedler M.*, *Nedoma J.*, *Ramik J.*, *Rohn J.*, *Zimmerman K.* Linear optimization problems with inexact data // Springer, 2006. 224 p.

[2] *Voshchinin A. P.*, *Bokov A. F.*, *Sotirov G. R.* Method of data analysis in case of interval non-statistical error // Zavod. lab., 1990. Vol. 56(7). Pp. 76–81.

[3] *Belov V. M.*, *Sukhanov V. A.*, *Lagutkina E. V.*, Interval approach for the solution of kinetic problems of simple chemical reactions // Vychisl. tekhnologii, 1997. Vol. 2(1). Pp. 10–18.

[4] *Shary S.* Data fitting problem under interval uncertainty in data // Industrial Laboratory. Diagnostics of Materials, 2020. Vol. 86(1). Pp. 62–74.

Порождение целочисленных алгоритмов генетическими методами

Ваганов Сергей Евгеньевич[1] pro100-pioner@mail.ru
Хашин Сергей Иванович[1,2]⋆ khash2@mail.ru
[1]г.Иваново ИвГУ
[2]г.Иваново ГК «Цифра»

Предложен и разработан метод генерации программы на языке программирования «Форт», реализующей заданный алгоритм. Алгоритм получает на входе набор из нескольких целых четырехбайтовых чисел и возвращает некоторый другой набор. Алгоритм задается примерами своей работы (тестами). Например:

сумма квадратов трех целых чисел	Сортировка трех чисел	Факториал
1, 2, 3 -> 14	3, 2, 0 -> 0, 2, 3	0, -> 1
0, 2, 3 -> 13	-1, -2, 3 -> -2, -1, 3	1 -> 1
-1, 1, -1-> 3	2, 1, 1, -> 1, 1, 2	...
...	...	12 -> 479001600

Более подробное описание можно найти в [1].

Язык Форт. Все рассматриваемые функции будут работать со стеком данных. В стеке хранятся только 4-байтовые целые числа. Функции в Форте не имеют аргументов. Исходные данные они берут из стека и там же оставляют результаты своей работы.

Мы используем не весь язык «Форт», а лишь его небольшую часть: команды манипуляции со стеком (DUP DROP ...), арифметика: (+ - * ...), битовые операции, сравнения, а также константы, условные и безусловные переходы. Общее количество исходных команд — 32.

Ниже приведены примеры двух программ на языке Форт, реализующих сумму квадратов двух чисел и факториал:

```
: SUMSQ2 DUP * SWAP DUP * + ;
: FACTORIAL CONST 1 OVER -- -ROT * OVER IF -6 SWAP DROP ;
```

Поиск программы, реализующей алгоритм выполняется в несколько шагов.

Шаг 1. Полный перебор. Так как исходных команд 32, на одноядерном процессоре за несколько минут можно перебрать все 6-байтовые программы, их около миллиарда. На полный перебор всех N-байтовых команд требуется примерно 32^{N-7} часов процессорного времени.

Шаг 2. Вероятностный и марковский подход. На первом шаге, даже если не удалось найти искомую программу, получаем «частичные программы», то есть выполняющие не весь тест, а лишь его часть. На их основе формируем таблицу вероятностей появления каждой команды в программе (вероятностный подход)

и вероятности команд после известной предыдущей (марковский подход). Если же таких программ нет совсем, или их слишком мало, то наш метод неприменим к данному алгоритму.

По построенным таблицам можно генерировать программы выполняющие большую долю тестов. На первом шаге мы перебрали все программы длиной $\leqslant 6$. Теперь будем последовательно генерировать программы длины от 7 до 14 с помощью вероятностного и марковского метода в течении заданного промежутка времени. По его истечении перестроим таблицу частот для её лучшего соответствия решаемой задаче. В нашей системе этот процесс повторяется 8 раз.

Шаг 3. «Генетика». Если искомая программа не найдена, то среди имеющихся «частичных программ» найдем несколько наиболее часто встречающихся корректных подпоследовательностей длины $\geqslant 2$. Будем их называть «кандидатами в гены».

Язык Форт позволяет легко расширить свой словарь — список имеющихся команд («геном») с помощью любых заданных корректных подпрограмм.

Введем понятие «качества» словаря: общее количество выполненных тестовых элементов на один миллион сгенерированных программ в этом словаре.

Будем по-очереди добавлять в словарь одного из кандидатов. Если качество словаря не улучшилось, то данного кандидата отклоняем. Иначе переводим «кандидата» в «члены» словаря и повторяем процесс. При этом среди следующих, вновь появляющихся кандидатов оставляем только тех, которые ранее не встречались.

Результаты. Предложенным методом (не более одного часа на программу) удалось отыскать программы, реализующие следующие алгоритмы: сумму квадратов двух и трёх чисел, сортировка двух и трёх чисел, максимум/минимум двух и трёх чисел, GCD, факториал, биномиальные коэффициенты, числа Фибоначчи, вычисление значения многочлена первой и второй степени. Все булевы функции от 2, 3 и 4 аргументов, почти все рассмотренные булевы функции от 5 аргументов (все рассмотреть не удается, их 2^{32} штук). Десятичные и двоичные цифры числа (сумма, максимум и т.д.). Задача Коллатца ($3n + 1$), но лишь в два этапа (с подсказкой).

Наиболее сложные из построенных программ имеют длину $12, 13, 14$, это биномиальные коэффициенты, некоторые булевы функции и многочлены.

[1] *Khashin S., Vaganov S.* Genetic Algorithms Using Forth // Infromational Technoligies, 2020. Vol. 26(1). Pp. 3–8.

Generation of integer algorithms by genetic method

Khashin Sergei[1,2]⋆ khash2@mail.ru
Vaganov Sergei[1] pro100-pioner@mail.ru
[1]Ivanovo, IvSU
[2]Ivanovo, Zyfra int.

A method for generating a program in the "Fort" programming language that implements the given algorithm is proposed and developed. The algorithm takes as input a sequence of several four-byte integers and returns some other sequence. The algorithm is set by examples of its work (tests). For example:

the sum of the of three integers	Sorting three numbers	Factorial
1, 2, 3 → 14	3, 2, 0 → 0, 2, 3	0 → 1
0, 2, 3 → 13	-1, -2, 3 → -2, -1, 3	1 → 1
-1, 1, -1 → 3	2, 1, 1 → 1, 1, 2	...
...	...	12 → 479001600

A more detailed description can be found in [1].

Fort language. All considered functions will work with the data stack. Only 4-byte integers are stored on the stack. Functions in Fort have no arguments. They take the source data from the stack and leave the results of their work there.

Of course, we do not take the whole language ¡¡Fort¿¿, but only a small part of it: stack manipulation commands, arithmetic, bitwise, comparison, constants, conditional and unconditional jumps. Total number of source commands — 32.

Here are two examples of Fort programs that implement the sum of squares of two numbers and the factorial:

```
: SUMSQ2 DUP * SWAP DUP * + ;
: FACTORIAL CONST 1 OVER -- -ROT * OVER IF -6 SWAP DROP ;
```

The search for a program that implements the given algorithm is performed in several steps.

Step 1. "Brute force", that is, a complete bust. Since the source commands are 32, on a single-core processor in a few minutes you can iterate over all 6-byte programs, there are about a billion of them. It takes approximately 32^{N-7} hours of CPU time to fully iterate through all N-byte commands.

Step 2. Probabilistic and Markov approach. At the first step, even if the desired program could not be found, we get "partial programs", that is, not performing the entire test, but only part of it. After the first step, we heave several hundred of the best partial programs. Based on them, we can understand which of the available commands are more suitable for our task, and which are less. If there are no such programs at all, or there are too few of them, then our method is not applicable to this algorithm.

More precisely, we form a table of the probability of each command appearing in the program (probabilistic approach) and the probability of a command after the known previous one (Markov approach).

Having these tables, it is possible to generate programs better suited to our task. In the first step, we itgerate through all the programs with a length of $\leqslant 6$. Now we will consistently generate programs of length from 7 to 12 or even 14 using the probabilistic and/or Markov method for a given of time. After that, we rebuild the frequency table to better match the problem being solved. In our system, this is repeated for 8 time slices.

Step 3. "Genetics". If the desired program is not found after step 2, then inside the available "partial programs" we will find several most frequently encountered correct subsequences of lengths $\geqslant 3$. We will call them "the gene candidates".

The Fort language makes it easy to expand your dictionary — the list of available words ("genome") using any given correct subroutines. However, it is not necessary to add all "candidates" at once.

Let's introduce the concept of "quality" of the dictionary: the total number of executed test items per one million generated programs in this dictionary.

We will add one of the candidates to the dictionary one by one. If the quality of the dictionary has not increased, then this candidate is discarded. Otherwise, we converse the "candidate" into the "members" of the dictionary and repeat the process. At the same time, among the following candidates, we leave only those who have not met before.

Results. Using the proposed method (no more than one hour per program), we were able to find programs that implement the following algorithms. Sum of squares of two and three numbers, sorting of two and three numbers, maximum/minimum of two and three numbers, GCD, factorial, binomial coefficients, Fibonacci numbers, calculation of the value of a polynomial of the first and second degree. ALL Boolean functionsof 2, 3 and 4 arguments, almost all considered Boolean functions of 5 arguments (it is not possible to consider all of them, there are 2^{32} pieces). Decimal and binary digits of a number (sum, maximum, etc.). The Collatz problem $(3n+1)$, but only in two stages (with a hint).

The most complex of the constructed programs have a length of $12, 13, 14$, these are binomial coefficients, some Boolean functions and polynomials.

[1] *Khashin S., Vaganov S.* Genetic Algorithms Using Forth // Infromational Technoligies, 2020. Vol. 26(1). Pp. 3–8.

Построение диаграммы Вороного для сайтов-многоугольников на основе алгоритма заметания

Коптелов Дмитрий Андреевич[1]⋆ dimitar98@list.ru
Местецкий Леонид Моисеевич[1,2] mestlm@mail.ru
[1]Московский Государственный Университет имени М.В. Ломоносова
[2]Москва, Федеральный исследовательский центр «Информатика и управление» РАН

Диаграмма Вороного (ДВ) конечного множества простых непересекающихся многоугольников, называемых сайтами, есть разбиение плоскости на локусы – области, в которых расстояние от всех точек до данного сайта не превосходит расстояния до других сайтов. Набор многоугольников является универсальной моделью для многих прикладных задач в технике, геоинформатике, дизайне, компьютерном зрении и графике.

Построение ДВ многоугольников обычно осуществляется на основе редукции – сведения к задаче построения ДВ отрезков, для которой существуют эффективные $O(n \log n)$ алгоритмы для n сайтов-отрезков. Редукция включает предобработку – образование отрезков из сторон многоугольников, и постобработку – построение локусов многоугольников путём объединения локусов сторон каждого многоугольника. При таком подходе не учитываются два специфических свойства полученных сайтов-отрезков. Во-первых, все эти отрезки попарно соединяются в вершинах многоугольников. А во-вторых, по одну сторону каждого отрезка лежит внутренность многоугольника, которая заведомо входит в его локус. Использование этих свойств в алгоритме построения ДВ многоугольников является ресурсом для сокращения вычислений.

В данном докладе представлен алгоритм прямого построения ДВ для набора сайтов-многоугольников. Алгоритм основан на парадигме плоского заметания, позволяющей эффективно учесть эти особенности.

Решение выполняется на основе редукции. Предобработка состоит в построении множества сайтов-элементов из вершин и сторон многоугольников. У каждого элемента задана ориентация, определяющая, с какой стороны от него находится внутренность многоугольника. Предложен алгоритм построения ДВ множества ориентированных сайтов-элементов, основанный на парадигме плоского заметания. Постобработка состоит в выделении рёбер полученной ДВ, образованных центрами пустых кругов, касающихся разных многоугольников.

Вертикальная заметающая прямая движется слева направо, пересекая в процессе движения все сайты-многоугольники. Прямая разбивает плоскость на левую и правую полуплоскости. Каждая точка заметающей прямой имеет в левой полуплоскости касательный пустой круг, который также касается хотя бы одного сайта-элемента. Те круги, которые касаются сайтов-элементов с внешней стороны, называются максимальными внешними пустыми кругами. Отрезок заметающей прямой, все точки которого имеют максимальные внешние пустые круги, касающиеся одного и того же сайта-элемента, называется зоной

Рис. 1. Статус заметающей прямой

этого элемента. Зоны элементов могут иметь пересечения только в своих концевых точках. Отрезки заметающей прямой, которые лежат внутри сайтов-многоугольников, называются пустыми зонами. Таким образом, вся заметающая прямая разбивается на зоны. Структура данных, описывающая упорядоченное снизу вверх множество зон, называется статусом заметающей прямой. В примере на рис. 1 представлены два сайта-многоугольника. Цифрами обозначены сайты-элементы: вершины многоугольников – сайты 1, 4, 5, и стороны многоугольников – 2, 3, 6, 7. Номер зоны показывает, к какому из пронумерованных сайтов она относится. Перечень событий представляет собой упорядоченную последовательность положений заметающей прямой, в которых происходит порождение и уничтожение зон.

Повышение эффективности предлагаемого алгоритма плоского заметания по сравнению с общим алгоритмом Форчуна [1] достигается за счёт следующих принципиальных решений.

1. Алгоритм строит только те ребра ДВ сайтов-элементов, которые находятся с внешней стороны сайтов-многоугольников. Таким образом, использование концепции ориентированных сайтов-элементов позволило избежать построения той части ДВ отрезков, которая находится внутри сайтов-многоугольников.
2. Перечень событий в методе плоского заметания имеет особенность, которая определяется спецификой построенного множества сайтов-элементов: подавляющее большинство событий связано с так называемыми проходными вершинами многоугольников, в которых одна из инцидентных им сторон лежит позади, а другая – впереди заметающей прямой. Предлагаемый алгоритм выполняет обработку таких событий за константное время, а не за логарифмическое, как в общем алгоритме Форчуна [1].

Предложенный алгоритм реализован в полном объеме и прошёл проверку на большом количестве примеров. Высокая надёжность и эффективность алгоритма также подтверждается вычислительными экспериментами со сложными множествами многоугольников, полученными в результате полигональной ап-

проксимации бинарных изображений, включающих несколько тысяч связных компонент.

Следует отметить, что, несмотря на значительное время, прошедшее после публикации алгоритма Форчуна [1] в 1986 году, и на большое число последовавших публикаций, полномасштабная реализация этого алгоритма для произвольного множества сайтов-отрезков не была сделана. Разработанный нами алгоритм восполняет этот пробел для важного частного случая – множества сайтов-отрезков, образованных сторонами сайтов многоугольников.

Работа поддержана грантом РФФИ No. 20-01-00664.

[1] *Fortune S.* A sweepline algorithm for Voronoi diagrams // Algorithmica, 1987. Vol. 2. Pp. 153–174

Sweepline algorithm for Voronoi diagram of polygonal sites

Koptelov Dmitry[1]⋆ dimitar98@list.ru
Mestetskiy Leonid[1,2] mestlm@mail.ru
[1]Moscow, Lomonosov Moscow State University
[2]Moscow, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences

Voronoi Diagram (VD) of finite set of disjoint simple polygons, called sites, is a partition of plane into loci (for each site at the locus) – regions, consisting of points that are closer to a given site than to all other. Set of polygons is a universal model for many applications in engineering, geoinformatics, design, computer vision and graphics.

VD of polygons construction usually done with a reduction to task of constructing VD of segments, for which there are effective $O(n \log n)$ algorithms for n segments. Preprocessing – constructing segments from polygons' sides, and postprocessing – polygon's loci construction by merging the loci of the sides of each polygon are also included in reduction. This approach doesn't take into account two specific properties of the resulting segment sites. Firstly, all this segments are connected in pairs in the vertices of the polygons. Secondly, on the one side of each segment lies the interior of the polygon. The polygon is obviously included in its locus. Using this properties in the algorithm for VD construction is a resource to reduce computations.

The report proposes an algorithm for the direct construction of VD of polygonal sites. Algorithm is based on sweepline paradigm, allowing to effectively take into account these properties.

The solution is performed based on reduction. Preprocessing is an constructing of set of sites-elements from vertices and edges of polygons. Each site has an orientation such that the interior of the polygon lies to the left of it. Proposed algorithm constructs VD for set of oriented sites with sweepline paradigm. Postprocessing is a selecting of edges of this VD formed by the centers of empty circles touching different polygons.

The vertical sweepline moves from left to right, crossing all polygonal sites in the process of movement. The line splits the plane into left and right half-planes. Each point of the sweepline has a tangent empty circle in the left half-plane, which also touches at least one site-element. Those circles that touch the site-elements on the outside are called the maximum outer empty circles. A segment of the sweepline, all points of which have maximum outer empty circles touching the same site-element, is called the zone of this element. Zones of elements can only have intersections at their endpoints. The sweepline segments that lie inside polygonal sites are called empty zones. Thus, the entire sweepline is divided into zones. The data structure describing a set of zones ordered from bottom to top is called the sweepline status. The example in Fig. 1 shows two polygonal sites. The numbers indicate the sites-

elements: the vertices of the polygons are sites 1, 4, 5, and the sides of the polygons are 2, 3, 6, 7. The zone number indicates which of the numbered sites it belongs to. The list of events is an ordered sequence of positions of the sweepline in which the generation and destruction of zones occurs.

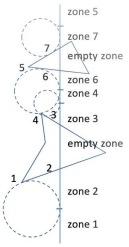


Fig. 1. Sweepline status

Improving the efficiency of the proposed sweepline algorithm in comparison with the general Fortune algorithm [1] is achieved due to the following fundamental solutions:

1. Algorithm constructs only such VD edges, which are on the outside of polygons. Concept of oriented sites allowed to avoid construction of VD edges located inside the polygons.
2. The list of events in sweepline algorithm has a special property: the majority of events are connected with "medium" polygon vertices, where one incident polygon side lies behind of the sweepline and the other in front of it. The proposed algorithm processes such events in constant time, and not in logarithmic time, as in the general Fortune algorithm [1].

The proposed algorithm is fully implemented and tested on a large number of examples. The high reliability and efficiency of the algorithm is also confirmed by computational experiments with complex sets of several thousand polygons.

It should be noted that, despite the considerable time that has passed since the publication of Fortune's algorithm [1] in 1986, a full-scale implementation of this algorithm for an arbitrary set of segment sites has not been made. The proposed algorithm fills this gap for an important special case - a set of sites formed by polygons.

This research is funded by RFBR, grant 20-01-00664.

[1] *Fortune S.* A sweepline algorithm for Voronoi diagrams // Algorithmica, 1987. Vol. 2. Pp. 153–174

Скелет многоугольной фигуры с выпуклым многоугольным структурирующим элементом: формализация и эффективный алгоритм построения

Ломов Никита Александрович[1]⋆ nikita_lomov@mail.ru
[1]Москва, ФИЦ ИУ РАН

Многие теоретические аспекты, связывающие операции математической морфологии и морфологические скелеты, справедливы не только для дисковых, но и для произвольных выпуклых (не обязательно строго) структурирующих элементов. Несмотря на это, в подавляющем большинстве случаев в морфологическом анализе оперируют скелетами, определёнными с помощью диска, что во многом обусловлено распространённостью и эффективностью соответствующих алгоритмов скелетизации.

Между тем известен ряд диаграмм Вороного, заданных нестандартными метриками, например, диаграмма в метрике Лагерра или диаграмма Аполлония [1], и можно считать, что каждая из таких метрик порождает свой скелет. Таким образом, возникает задача формального определения метрики, порождённой произвольным выпуклым элементом, в частности, многоугольником, и разработки эффективных алгоритмов скелетизации на основании данной метрики. Создание таких алгоритмов позволило бы обобщить процедуры морфологической обработки на базе скелетного представления [2] на широкий класс примитивов формы.

Рассмотрим выпуклую фигуру S, такую, что начало координат O — внутренняя точка фигуры, а также семейство фигур $\{S(A,r)\}$, полученных из S композицией гомотетии с коэффициентом $r \geqslant 0$ и параллельного переноса на вектор $\overrightarrow{OA}$. Точку A будем называть центром фигуры $S(A,r)$. Квазирасстояние $d_S(A,B)$ определяется как размер r такой фигуры $S(A,r)$, которая касается точки B: $d_S(A,B) = r$: $B \in \partial S(A,r)$.

При использовании стандартных определений скелета как множества центров вписанных элементов или множества точек, имеющих более одной ближайшей точки границы, в применении к многоугольному элементу возникают проблемы, вызванные возможным нарушением связности скелета или наличием в нём двумерной части [3]. Покажем, что от этих проблем можно уйти за счёт уточнения понятия более близкой точки.

Определение. Пусть $d_S(A,B) = d_S(A,C) = r$, и точки B и C лежат на одном и том же прямолинейном отрезке границы $S(A,r)$. Тогда из точек B и C ближней является та, которая ближе к середине этого отрезка в евклидовой метрике.

Фактически это определение задаёт отношение частичного порядка на множестве расстояний, снабжённых дополнительными параметрами — расстояниями до середин отрезков границы. При этом множество точек, имеющих несколь-

ко ближайших (с минимальным расстоянием) точек границы, будет иметь меру ноль, то есть скелет будет сопокупностью двумерных кривых.

Особый интерес представляет скелетизация многоугольной фигуры с помощью выпуклого многоугольника, так как такую задачу удобно рассматривать в терминах вычислительной геометрии. В этом случае считается, что граница фигуры состоит из сайтов двух типов — точек и сегментов — а скелет рассматривается как подмножество диаграммы Вороного этих сайтов, полученное исключением бисекторов между сайтами-точками и их соседними сайтами-сегментами.

Теорема. Пусть F— многоугольная фигура, а $\{S_n\}_{n=0}^{\infty}$ — последовательность выпуклых структурирующих элементов, полученных из выпуклого многоугольника S аппроксимацией его сторон дугами кривизны $\{\varkappa_n\}_{n=0}^{\infty}$, $\lim\limits_{n\to\infty} \varkappa_n = 0$. Пусть $C(G)$ — внутренняя (лежащая внутри F) часть ячейки Вороного сайта G в квазиметрике с уточнёнными расстояниями, порождённой S, а $C_i(G)$ — внутренняя часть ячейки Вороного в квазиметрике, порождённой S_n. Тогда $\{C_n(G)\}$ при $n \to \infty$ стремится к $C(G)$ в метрике Хаусдорфа: $\lim\limits_{n\to\infty} d_H(C_n(G), C(G)) = 0$.

Практическая реализация алгоритма использует парадигму заметающей прямой, представленную в работе [4]. В качестве структурирующего элемента используется многоугольник S, самая правая точка которого имеет координаты $(1, 0)$, а других точек с абсциссой 1 нет[1]. Тогда отображение, определённое как $*_S(A) = (x_A + r_S(A), y_A)$ (здесь $r_S(A) = \min\limits_{B\in\partial F}\{d_S(A, B)\}$), преобразует ячейки Вороного таким образом, что левой точкой ячейки становится либо сам сайт в случае сайта-точки, либо первый в лексикографическом порядке конец сайта-сегмента.

Предложенный алгоритм строит транформированную диаграмму Вороного $*_S(V)$. В основе алгоритма лежат две структуры данных — очередь событий-пересечений и статус заметающей прямой, содержащий бисекторы пар смежных в диаграмме сайтов, упорядоченные по высоте их точек пересечения с заметающей прямой. Также используется не изменяемая в ходе алгоритма очередь событий, связанных с появлением сайтов-точек, реализованная в виде простого списка. В зависимости от числа и направления исходящих бисекторов сайты-точки делятся на девять типов, каждый из которых определяет свой способ обработки события.

Разработанный алгоритм имеет сложность $O(mn \log n)$, где n — число сторон многоугольной фигуры, m — число сторон структурирующего элемента. Алгоритм реализован на языке C++ с использованием библиотеки CGAL для работы с геометрическими примитивами, скорость работы алгоритма соответствует теоретическим оценкам. Кроме того, так как бисекторы всех трёх типов — точка-точка, точка-сегмент и сегмент-сегмент — являются ломаными линиями,

[1]К выполнению этих условий можно прийти, применив к исходному многоугольнику аффинное преобразование M, а к полученному скелету — преобразование M^{-1}.

вычисления производятся в рациональной арифметике, обеспечивающей абсолютную точность.

Работа поддержана грантом РФФИ No. 20-01-00664.

[1] *Wormser C.* Generalized Voronoi Diagrams and Applications. PhD Thesis // Université Nice Sophia Antipolis, 2008. 122 p.

[2] *Ломов Н. А., Местецкий Л. М.* Площадь дискового покрытия — дескриптор формы изображения // Компьютерная оптика, 2016. Т. 40(2). С. 516–525.

[3] *Визильтер Ю. В., Сидякин С. В* Построение обобщенных скелетов многоугольных бинарных фигур с многоугольными выпуклыми структурирующими элементами // Интеллектуализация обработки информации: 9-я международная конференция. Сборник докладов, М.: Торус Пресс, 2012. С. 414–417.

[4] *Fortune S.* A Sweepline Algorithm for Voronoi Diagrams // Algorithmica, 1987. Vol. 2. Pp. 153–174.

Skeleton of a Polygonal Figure with a Convex Polygonal Structuring Element: Formalization and an Efficient Algorithm of Construction

Lomov Nikita[1]⋆ nikita_lomov@mail.ru

[1]Moscow, FRC CSC RAS

Many theoretical aspects connecting the operations of mathematical morphology and morphological skeletons are valid not only for disk, but also for arbitrary convex (not necessarily strictly) structuring elements. Despite this, in the overwhelming majority of cases in morphological analysis, skeletons, determined using a disk, are used, which is largely due to the prevalence and efficiency of the corresponding skeletonization algorithms.

Meanwhile, a number of Voronoi diagrams, given by non-standard metrics, are known, for example, a diagram in the Laguerre metric or an Apollonius diagram [1], and we can assume that each of these metrics induces its own skeleton. Thus, the problem of formally determining the metric generated by an arbitrary convex element, in particular, a polygon, and developing efficient skeletonization algorithms based on this metric, arises. The creation of such algorithms would allow generalizing the morphological processing procedures based on the skeletal representation [2] to a wide class of shape primitives.

Consider a convex figure S such that the origin O is an interior point of the figure, as well as a family of figures $\{S(A,r)\}$ obtained from S by a composition of homothety with coefficient $r \geqslant 0$ and parallel translation to the vector $\overrightarrow{OA}$. The point A will be called the center of the figure $S(A,r)$. The quasi-distance $d_S(A,B)$ is defined as the size r of such a figure $S(A,r)$ that touches the point B: $d_S(A,B) = r$: $B \in \partial S(A,r)$.

When using the standard definitions of a skeleton as a set of centers of inscribed elements or a set of points with more than one nearest boundary point, when applied to a polygonal element, problems, caused by a possible violation of the skeleton's connectivity or the presence of a two-dimensional part in it, arise [3]. Let us show that these problems can be avoided by clarifying the concept of a closer point.

Definition. Let $d_S(A,B) = d_S(A,C) = r$, and the points B and C lie on the same straight line segment of the boundary $S(A,r)$. Then, of the points B and C, the nearest one is the one that is closer to the midpoint of this segment in the Euclidean metric.

In fact, this definition specifies a partial ordering relation on the set of distances provided with additional parameters, that are distances to the midpoints of the boundary segments. In this case, the set of points that have several nearest (with the minimum distance) points of the boundary will have a measure of zero, that is, the skeleton will be a composition of two-dimensional curves.

Of particular interest is the skeletonization of a polygonal figure using a convex polygon, since it is convenient to approach such a problem in terms of computational

geometry. In this case, it is considered that the boundary of the figure consists of two types of sites—points and segments—and the skeleton is considered as a subset of the Voronoi diagram of these sites, obtained by excluding bisectors between point sites and their neighboring segment sites.

Theorem. Let F be a polygonal figure, and $\{S_n\}_{n=0}^{\infty}$ be a sequence of convex structuring elements obtained from the convex polygon S by approximating its sides by arcs of curvature $\{\varkappa_n\}_{n=0}^{\infty}$, $\lim\limits_{n\to\infty} \varkappa_n = 0$. Let $C(G)$ be the inner (lying inside F) part of the Voronoi cell of the site G in the quasimetric with clarified distances defined by S, and $C_i(G)$ be the inner part of the Voronoi cell in the quasimetric defined by S_n. Then $\{C_n(G)\}$ as $n \to \infty$ converges to $C(G)$ in the Hausdorff metric: $\lim\limits_{n\to\infty} d_H(C_n(G), C(G)) = 0$.

A practical implementation of the algorithm uses the sweeping line paradigm presented in [4]. Polygon S is used as a structuring element, the rightmost point of which has coordinates $(1, 0)$, and there are no other points with abscissa 1[1]. Then the mapping defined as $*_S(A) = (x_A + r_S(A), y_A)$ (here $r_S(A) = \min\limits_{B\in\partial F}\{d_S(A, B)\}$), transforms Voronoi cells in such a way that the leftmost point of the cell becomes either the site itself in the case of a point site, or the first endpoint of the site segment in lexicographic order.

The proposed algorithm builds the transformed Voronoi diagram $*_S(V)$. The algorithm is based on two data structures—a queue of intersection events and a sweeping line status containing bisectors of pairs of adjacent sites in the diagram, ordered by the height of their intersection points with the sweeping line. Also, a queue of events associated with the appearance of point sites, which is not changed during the algorithm, is implemented in the form of a simple list. Depending on the number and direction of outgoing bisectors, point sites are divided into nine types, each of which determines its own way of processing an event.

The developed algorithm has complexity $O(mn \log n)$, where n is the number of sides of the polygonal figure, and m is the number of sides of the structuring element. The algorithm is implemented in C++ using the CGAL library for working with geometric primitives, the speed of the algorithm corresponds to theoretical estimates. In addition, since bisectors of all three types—point-to-point, point-segment and segment-segment—are polygonal chains, the calculations are performed in rational arithmetic that ensures absolute accuracy.

This research is funded by RFBR, grant No. 20-01-00664.

[1] *Wormser C.* Generalized Voronoi Diagrams and Applications. PhD Thesis // Université Nice Sophia Antipolis, 2008. 122 p.

[2] *Lomov N., Mestetskiy L.* Area of the disk cover as an image shape descriptor // Computer Optics, 2016. Vol. 40(4). Pp. 516–525.

[1] These conditions can be met by applying the affine transformation M to the original polygon, and applying the transformation M^{-1} to the resulting skeleton.

[3] *Vizilter Yu., Sidyakin S.* Generalized skeletons of 2D polygonal figures with convex polygonal structuring elements // Proceedings of the 9th International Conference on Intelligent Data Processing, Moscow: Torus Press, 2012. Pp. 414–417.

[4] *Fortune S.* A Sweepline Algorithm for Voronoi Diagrams // Algorithmica, 1987. Vol. 2. Pp. 153–174.

Сложность вычисления Гамма-функции Эйлера

Карацуба Екатерина Анатольевна* ekaratsuba@gmail.com
Москва, ВЦ ФИЦ ИУ РАН

Памяти Константина Владимировича Рудакова, который всегда интересовался новыми научными результатами и открытиями, и был энтузиастом их применения на практике, и в частности внедрения и применения быстрых алгоритмов.

Одна из самых широко распространённых в анализе и математической физике высших трансцендентных функций Гамма-функция Эйлера определяется соотношением

$$\Gamma(z+1) = z\Gamma(z),\ Re\ z > 0.$$

При $z = x = n$ – натуральное число, $\Gamma(n+1) = n!$.

Задачу вычисления $\Gamma(z)$ для любого аргумента z с нужной точностью с наименьшими временными затратами можно сформулировать как задачу определения сложности вычисления этой функции.

Далее считаем, что числа записаны в двоичной системе счисления, знаки которой 0 и 1 называются битами. Запись знаков 0, 1, плюс, минус, скобка; сложение, вычитание и умножение двух битов назовём одной элементарной или битовой операцией. Вычислить (вещественную) функцию $y = f(x)$ в точке $x = x_0$ с точностью до n знаков, значит найти такое число A, что $|f(x_0) - A| \leqslant 2^{-n}$.

Количество битовых операций, достаточное для вычисления функции $f(x)$ в точке x_0 с точностью до n знаков посредством данного алгоритма, называется сложностью вычисления $f(x)$ в точке x_0 и обозначается $s_f(n) = s_{f,x_0}(n)$.

Первые задачи по оценке (битовой) сложности вычисления были сформулированы А.Н. Колмогоровым в 1950-х гг., первый быстрый алгоритм (умножения) был найден А.А. Карацубой в 1960 г. (см. [1]), что привело в дальнейшем к созданию серии алгоритмов быстрого вычисления различных функций (алгоритм умножения эквивалентен алгоритму вычисления функции $y = x^2$) со сложностью $O\left(n^{1+\varepsilon}\right)$ (вместо $O\left(n^{2+\varepsilon}\right)$ – лучшей сложности вычисления функций до эры быстрых алгоритмов).

Заметим, что задача уменьшения константы в знаке O в сложности вычисления (что приводит к практическому улучшению эффективности вычисления в частных случаях) не рассматривается в теории быстрых алгоритмов. В то же время основателей этого направления в математике очень интересовали оптимальные оценки сложности вычисления (которых до сих пор не существует ни для каких нетривиальных случаев). Поэтому иногда (см.[2]) под сложностью вычисления функции имелась в виду наилучшая (с наименьшим "ростом по n") сложность вычисления, обозначим её как $S_f(n)$ (например, в настоящий момент такой сложностью умножения $S_{x^2}(n)$ является "сложность М. Фюрера" [3]).

С помощью построенного автором метода БВЕ (см. [4]) можно доказать следующие утверждения относительно сложности вычисления Гамма-функции Эйлера (для простоты мы рассматриваем случай вещественного аргумента, для комплексного аргумента соответствующие теоремы доказываются отдельно для вещественной и мнимой частей аргумента)

Теорема 1.

$$S_{\Gamma(x_0)} = O\left((n^{1+\varepsilon}\right),\ x_0 = \frac{p}{q},$$

p, q –целые, $(p, q) = 1$.

Теорема 2.

$$S_{\Gamma(x_0)} = O\left((n^{1+\varepsilon}\right),\ x_0 = \alpha,$$

α – алгебраическое число, которое является корнем многочлена с целыми (заранее известными) коэффициентами.

Заметим, что алгоритм из второй теоремы предполагает гораздо большую константу "в O", чем алгоритм первой теоремы. Тем не менее оба алгоритма обеспечивают сложность вычисления, близкую к оптимальной, просты в применении и допускают распараллеливание.

Наряду с задачами по быстрым вычислениям, в которых основным растущим параметром является точность вычисления n, $n \to +\infty$, а аргумент считается фиксированным, встречаются задачи, в которых точность вычисления фиксирована, а растёт аргумент Гамма-функции Эйлера $\Gamma(x)$, $x \to +\infty$. В этом случае до 2000 г. для вычисления Гамма-функции часто использовалась формула Стирлинга (см. [5]). После доказательства гипотезы Рамануджана (см.[6]) для вычисления Гамма-функции при растущем аргументе уместно воспользоваться следующей удобной "интервальной" формулой

$$\sqrt{\pi}\left(\frac{x}{e}\right)^x \left(8x^3 + 4x^2 + x + \frac{1}{100}\right)^{\frac{1}{6}} < \Gamma(1+x) < \sqrt{\pi}\left(\frac{x}{e}\right)^x \left(8x^3 + 4x^2 + x + \frac{1}{30}\right)^{\frac{1}{6}}.$$"

Работа частично финансирована грантом РФФИ 19-07-00750.

[1] *Гриценко С. А., Карацуба Е. А., Королёв М. А., Резвякова И. С., Толев Д. И., Чанга М. Е.* Научные достижения Анатолия Алексеевича Карацубы // Совр. пробл. матем., 2012. Т. 16. С. 7–30.

[2] *Карацуба А. А.* Сложность вычислений // Тр. МИАН, 1995. С. 186–202.

[3] *Fürer M.* Faster Integer Multiplication // SIAM Journal on Computing, 2009. Vol. 39(3). Pp. 979–1005.

[4] *Карацуба Е. А.* Быстрые вычисления трансцендентных функций // Пробл. передачи информ., 1991. Т. 27(4). С. 76–99.

[5] *Temme N.* Special Functions. An Introduction to the Classical Functions of Mathematical Physics // J. Wiley and Sons, 1996.

[6] *Karatsuba E.* On the asymptotic representation of the Euler gamma function by Ramanujan // J. Comput. Appl. Math, 2001. Vol. 135(2). Pp. 225–240.

The Complexity of computation of the Euler Gamma function

Karatsuba Ekaterina⋆ ekaratsuba@gmail.com
CC FRC CSC RAS, Moscow, Russia

In memory of Konstantin Vladimirovich Rudakov, who was always interested in new scientific results and discoveries, and was an enthusiast of their application in practice, and in particular the implementation and application of fast algorithms.

One of the most widespread higher transcendental functions in analysis and mathematical physics, the Euler gamma function is defined by the relation

$$\Gamma(z+1) = z\Gamma(z),\ Re\ z > 0.$$

When $z = x = n$ is a positive integer, $\Gamma(n+1) = n!$.

The problem of calculating $\Gamma(z)$ for any argument z with the required accuracy with the least time costs can be formulated as the problem of determining the complexity of the calculation of this function.

Further, we assume that the numbers are written in a binary number system, the signs of which 0 and 1 are called bits. Writing symbols 0, 1, plus, minus, parenthesis; addition, subtraction and multiplication of two bits will be called one elementary or bit operation. Calculate the (real) function $y = f(x)$ at the point $x = x_0$ up to n signs, then find a number A such that
$|f(x_0) - A| \leqslant 2^{-n}$.

The number of bit operations sufficient to compute the function $f(x)$ at the point x_0 up to n digits using an algorithm is called the complexity of computing $f(x)$ at the point x_0 and is denoted by $s_f(n) = s_{f,x_0}(n)$.

The first problems of estimating the (bit) computational complexity were formulated by A.N. Kolmogorov in the 1950s, the first fast algorithm (of multiplication) was found by A.A. Karatsuba in 1960 (see [1]), which later led to the creation of a series of algorithms for fast computation of various functions (the multiplication algorithm is equivalent to the algorithm for computing the function $y = x^2$) with complexity $O\left(n^{1+\varepsilon}\right)$ (instead of $O\left(n^{2+\varepsilon}\right)$ - the best computational complexity of functions before the era of fast algorithms).

Note that the problem of reducing the constant in the sign of O in the computational complexity (which leads to a practical improvement in the computational efficiency in particular cases) is not considered in the theory of fast algorithms. At the same time, the founders of this direction in mathematics were very interested in optimal estimates of the computational complexity (which still do not exist for any non-trivial cases). Therefore, sometimes (see [2]), the complexity of calculation of a function meant the best (with the smallest "growth in n") complexity of the calculation; we denote it as $S_f(n)$ (for example, at the moment such complexity of multiplication $S_{x^2}(n)$ is the M. Fürer complexity [3]).

Using the FEE method constructed by the author (see [4]), one can prove the following statements regarding the complexity of computation of the Euler Gamma

function (for simplicity, we consider the case of a real argument; for a complex argument, the corresponding theorems are proved separately for the real and imaginary parts of the argument).

Theorem 1.

$$S_{\Gamma(x_0)} = O\left((n^{1+\varepsilon}\right),\ x_0 = \frac{p}{q},$$

p, q – integers, $(p, q) = 1$.

Theorem 2.

$$S_{\Gamma(x_0)} = O\left((n^{1+\varepsilon}\right),\ x_0 = \alpha,$$

α – algebraic number, which is a root of a polynomial with integer (known in advance) coefficients.

Note that the algorithm from the second theorem assumes a much larger constant "in O" than the algorithm of the first theorem. Nevertheless, both algorithms provide computational complexity close to optimal, are easy to use, and can be parallelized.

Along with the problems of fast computations, in which the main growing parameter is the computation accuracy n, $n \to +\infty$, and the argument is considered fixed, there are problems in which the computation accuracy is fixed, and the argument of the Euler Gamma function $\Gamma(x)$ grows $x \to +\infty$.

In this case, until 2000, the Stirling formula was often used to calculate the Gamma function (see [5]). After proving Ramanujan's conjecture (see [6]), to calculate the Gamma function for an increasing argument, it is appropriate to use the following convenient "interval" formula

$$\sqrt{\pi}\left(\frac{x}{e}\right)^x \left(8x^3 + 4x^2 + x + \frac{1}{100}\right)^{\frac{1}{6}} < \Gamma(1+x) < \sqrt{\pi}\left(\frac{x}{e}\right)^x \left(8x^3 + 4x^2 + x + \frac{1}{30}\right)^{\frac{1}{6}}."$$

The work was partially funded by the RFBR grant 19-07-00750.

[1] *Gritsenko S., Karatsuba E., Korolev M., Rezvyakova I., Tolev D., Changa M.* Scientific Achievements of Anatolii Alekseevich Karatsuba // Proc. Steklov Inst. Math, 2013. Vol. 2. Pp. 1–22.

[2] *Karatsuba A.* The complexity of computations // Proc. Steklov Inst. Math, 1995. Pp. 169–183.

[3] *Fürer M.* Faster Integer Multiplication // SIAM Journal on Computing, 2009. Vol. 39(3). Pp. 979–1005.

[4] *Karatsuba E.* Fast Evaluation of Transcendental Functions // Problems Inform. Transmission, 1991. Vol. 27(4). Pp. 339–360.

[5] *Temme N.* Special Functions. An Introduction to the Classical Functions of Mathematical Physics // J. Wiley and Sons, 1996.

[6] *Karatsuba E.* On the asymptotic representation of the Euler gamma function by Ramanujan // J. Comput. Appl. Math, 2001. Vol. 135(2). Pp. 225–240.

Постановка задачи формирования оптимального покрытия области неопределенности эталонами для систем оптической навигации

Гришин Владимир Александрович vgrishin@iki.rssi.ru
Москва, ИКИ РАН

В настоящее время существенно расширяется круг задач, в которых используется стыковка с беспилотными космическими аппаратами. Новые задачи связаны с обслуживанием космических аппаратов, пополнением запасов топлива/окислителя, заменой вышедших из строя компонентов космического аппарата, изменением орбиты и др.

Основной проблемой является стыковка с некооперируемыми космическими аппаратами. В этом случае алгоритмы работы оптико-электронных систем по сравнению с алгоритмами стыковки с кооперируемыми космическими аппаратами многократно усложняются. В случае стыковки с некооперируемыми космическими аппаратами используются методы распознавания и измерения с имеющейся на борту активного аппарата моделью пассивного аппарата (model-based object recognition).

В общем случае ракурс наблюдения пассивного аппарата может быть любым, а расстояние до него изменяться в очень широких пределах. Космический аппарат является сложным трехмерным объектом. При изменении ракурса наблюдения происходит топологическая перестройка проекции изображения космического аппарата на плоскости фотоприемных матриц. Для описания этой перестройки могут быть использованы так называемые аспектные графы [1], [2] . При этом количество топологически не эквивалентных изображений объекта может быть весьма велико, поскольку объект имеет очень сложную пространственную структуру. В принципе, количество эталонов, необходимых для распознавания/измерения должно соответствовать числу топологически не эквивалентных изображений объекта. Однако ограниченность разрешающей способности фотоприемных матриц требует увеличения количества эталонов. Большое количество эталонных изображений является крайне нежелательным, поскольку процесс их подготовки весьма трудоемок и должен обязательно выполняться под непосредственным управлением квалифицированного оператора. Оператор должен производить жесткий отбор информативных точек (interest points), используемых для распознавания и навигации. С другой стороны, большой объем эталонной информации требует большого объема бортового запоминающего устройства. Это также нежелательно.

В этих условиях возникает задача построения минимально необходимого покрытия области неопределенности эталонами, отвечающего следующим условиям:

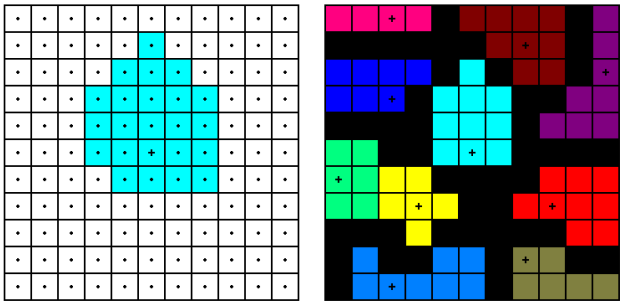
1. Покрытие должно обеспечивать для каждого ракурса наблюдение и распознание числа информативных точек, не меньше заданной величины.

2. Матрица ошибок измерения для каждого ракурса должна удовлетворять заданным требованиям по точности и корреляционным связям между ошибками измерения.

Поскольку построение аспектных графов является очень сложной вычислительной задачей для реальных космических аппаратов, то целесообразно покрыть всю область неопределенности частой сеткой, например, с шагом порядка 1 градуса. Сетка может быть неравномерной, например, как в случае покрытия сферы. Затем необходимо отобрать те эталоны, которые удовлетворяют указанным выше условиям.

Проблемой является то, что область применимости каждого эталона по ракурсам наблюдения определяется множеством точек, в которых удовлетворяются условия, указанные выше. Форма этих областей не является постоянной, зависит от многих факторов и индивидуальна для каждого конкретного эталона.

В левой части рисунка показан иллюстративный пример сетки для $11 \times 11 = 121$ ракурсов (это точки внутри ячеек). Эта сетка покрывает всю область неопределенности оптической навигационной системы. Для каждого ракурса строится эталон и определяется его область применимости в соответствии с перечисленными выше двумя условиями. Каждый эталон занимает связанное множество точек, соответствующих различным ракурсам, в которых выполнены приведенные выше условия. В левой части рисунка показан один такой эталон, ракурс которого показан крестиком. В правой части рисунка показано покрытие области неопределенности десятью эталонами. Области применимости эталонов пересекаются. Области пересечения залиты черным цветом. Ракурсы использованных для покрытия десяти эталонов отмечены крестиками.



Следует отметить, что прямой перебор вариантов всех возможных покрытий практически невозможен даже для $11 \times 11 = 121$ отсчетов углов ракурса. Число вариантов покрытия, которые необходимо рассмотреть и сравнить, равно следующему выражению:

$$N = C_{121}^{1} + C_{121}^{2} + C_{121}^{3} + \ldots + C_{121}^{120} + C_{121}^{121} \qquad (1)$$

Получить значение суммы этого конечного ряда можно через бином Ньютона степени 121 для случая, когда оба числа в скобках равны 1, а сама скобка соответственно равна 2:

$$N = (1+1)^{121} - 1 = (2)^{121} - 1 = 2.658 \times 10^{36} \tag{2}$$

Если принять, что на формирование одного покрытия и оценку его качества требуется всего 1 микросекунда (что очень быстро), тогда на полный перебор потребуется до 8.429×10^{22} лет. Таким образом, возникает задача построения оптимального или квазиоптимального покрытия области неопределенности с приемлемыми затратами вычислительных ресурсов.

[1] *Van Effelterre T.* Aspect graphs for visual recognition of three-dimensional objects // Perception, 1994. Vol. 23(5). Pp. 563–582.

[2] *Yang C., Marefat M., Johnson E.* Entity-based aspect graphs: Making viewer centered representations more efficient // Pattern Recognition Letters, 1998, Vol. 19(3). Pp. 265–277.

Formulation of the problem of generation the optimal coverage by reference images of the uncertainty region of optical navigation systems

Grishin Vladimir vgrishin@iki.rssi.ru
Moscow, IKI (Space Research Institute of Russian Academy of Sciences)

Currently, the range of tasks for which unmanned spacecraft docking is used is expanding significantly. New tasks include the maintenance of spacecraft, the replenishment of fuel/oxidiser reserves, the replacement of failed components of the spacecraft, and orbital changes.

The main problem is docking with non-cooperative spacecraft, where the algorithms for operating optoelectronic navigation systems are more complex. In such cases, model-based object recognition and measurement methods are used onboard the active vehicle.

Generally speaking, the aspect angle of a passive spacecraft observation can be anything, and the distance from it can cover a wide range. The spacecraft is a complex three-dimensional (3D) object. When the angles of observation are changed, a topological rearrangement of the projection of the spacecraft image on the plane of the photodetector matrices occurs. To describe this rearrangement, the so-called aspect graphs can be used [1], [2]. The number of topologically non-equivalent images of an object can be very large, since the object has a complex spatial 3D structure. In principle, the number of reference images required for recognition/measurement should correspond to the number of topologically non-equivalent images of an object. However, the limitation of the resolution of the photodetector matrices requires an increase in the number of reference images. A large number of reference images is highly undesirable since their preparation is a laborious process and must be carried out under the direct control of a qualified operator, who must make a rigorous selection of interest points used for recognition and navigation. On the other hand, a large amount of reference images requires a large amount of onboard storage. This is also undesirable.

Under these conditions, the problem of achieving the minimum required coverage of the uncertainty region by reference images arises. Such coverage should meet the following conditions:

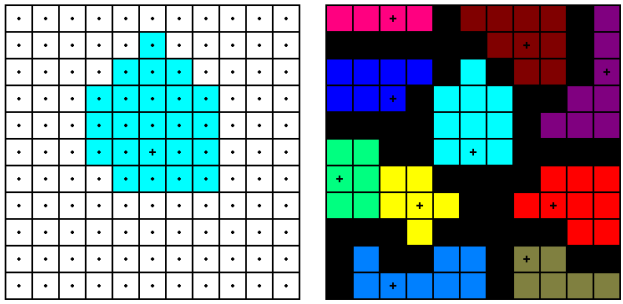
1. It must ensure, for each aspect angle, observation and recognition of the number of interest points not less than a given value.

2. The matrix of measurement errors for each aspect angle must satisfy the specified requirements for accuracy and correlations between measurement errors.

Since the formation of aspect graphs is a complex computational issue in the case of real spacecraft, it is advisable to cover the entire uncertainty region with a dense grid, for example with a step of the order of 1 angular degree. The grid can be uneven, for example when a sphere is covered. Then, a minimal set of reference images that meet the above conditions are selected.

The problem is that the area of applicability of each reference image in terms of observation angles is determined by the set of points at which the conditions indicated above are satisfied. The shapes of these areas are not consistent; they depend on many factors and vary according to each specific reference image.

The left section of figure shows an illustrative example of a grid for $11 \times 11 = 121$ aspects angles (these are points inside the cells). This grid covers the entire uncertainty region of optical navigation systems. For each aspect, a reference image is constructed, and its area of applicability is determined in accordance with the above two conditions. Each reference image occupies a connected set of points corresponding to different aspects in which the above conditions are met. The left section of figure shows one such reference image, the aspect angles of which are marked with a cross. The right section of figure shows the coverage of the uncertainty region with ten reference images. The areas of applicability of the reference images overlap. The intersection areas are filled in black. The aspect angles of these ten reference images are marked with crosses.



It should be noted that a direct estimation of all possible coverings is practically impossible, even for $11 \times 11 = 121$ aspects angles. The maximum number of coverage variants to be considered and compared is determined as follows:

$$N = C_{121}^{1} + C_{121}^{2} + C_{121}^{3} + \cdots + C_{121}^{120} + C_{121}^{121} \tag{1}$$

The value of the sum of this finite series can be obtained through the Newton binomial formula of degree 121 when both numbers in parentheses are equal to one, and the parenthesis itself is, respectively, equal to 2:

$$N = (1+1)^{121} - 1 = (2)^{121} - 1 = 2.658 \times 10^{36} \tag{2}$$

If we assume that it takes only 1 microsecond to form one coverage and evaluate its quality (which is very fast), then it will take up to 8.429×10^{22} years for a complete search of the optimal coverage. Thus, the problem arises of constructing an optimal or quasi-optimal coverage of the region of uncertainty with an acceptable computational cost.

[1] *Van Effelterre T.* Aspect graphs for visual recognition of three-dimensional objects // Perception, 1994. Vol. 23(5). Pp. 563–582.

[2] *Yang C., Marefat M., Johnson E.* Entity-based aspect graphs: Making viewer centered representations more efficient // Pattern Recognition Letters, 1998, Vol. 19(3). Pp. 265–277.

Методы 3D-реконструкции поверхности в задаче автономной навигации робота-марсохода

Бобков Александр Валентинович[1] Alexander.Bobkov@bmstu.ru
Дай Ифань[1]⋆ daiyifan1997@outlook.com
[1]Москва, МГТУ им. Н.Э.Баумана

Получение трехмерной информации об объектах по наблюдаемому двухмерному изображению является важной, интересной и так и до конца не решенной задачей в области компьютерного зрения. Особую актуальность задача трёхмерной реконструкции приобретает в приложениях, связанных с автономной навигацией роботов по пресеченной местности, например – для навигации робота-марсохода.

3D-реконструкция сцены по изображению широко применяется в задачах искусственного интеллекта, в робототехнике, в беспилотном вождении, в задачах визуальной навигации, 3D-печати и т. д. Данная работа обобщает и анализирует методы трехмерной реконструкции с использованием видеоизображений.

Все методы трехмерной реконструкции можно разделить на контактные и бесконтактные. К контактным методам в основном относятся: триггерное измерение, непрерывное измерение, координатно-измерительные машины и робототехнические руки. Контактные методы используют определенные инструменты и могут напрямую получать трехмерную информацию о сцене с высокой точностью. Они подходят для случаев, когда можно напрямую контактировать с объектами сцены и проводить их измерения. Однако во многих случаях контактные методы неприемлимы, поскольку могут привести к повреждению как поверхности объекта, так и измерительного инструмента.

Бесконтактный метод лишен этого недостатка. Он как правило использует методы анализа изображений для получения информации о сцене. Бесконтактный метод имеет более широкий спектр применения, хотя точность восстановления существенно ниже, нежели у контактного метода.

Бесконтактные методы можно разделить на активные и пассивные в зависимости от метода получения информации о глубине сцены. К активным относят радиодальномеры, лазерное сканирование, метод структурированного света. Информация о глубине объекта измеряется непосредственно, и может быть получена путем воздействия источника излучения (лазер, ультразвук, электромагнитные волны и т. Д.) на целевой объект и последующего анализа отраженного сигнала, после чего рассчиывается приближенная форма объекта.

Технология активной реконструкции хорошо отработана, имеет высокую точностью, однако оборудование, как правило, достаточно сложное, дорогостоящее и чувствительно к внешним воздействиям. Этих недостатков лишены пассивные методы, использующие видеоизображение с одной или нескольких камер. Пассивные методы получают последовательность изображений при по-

мощи камеры и восстанавливают трехмерную структурную модель сцены, основываясь на понимании двухмерного изображения.

Пассивные методы можно разделить на монокулярные и бинокулярные (с несколькими камерами).

Монокулярное зрение может использовать камеру для съемки одного или нескольких изображений с одной точки, или съемки нескольких изображений с нескольких точек. При использовании одного кадра информация о глубине сцены извлекается из двухмерных характеристик изображения (затенение, текстура, фокус, контур и т. Д.). Данная группа методов называется восстановлением формы по освещённости (Shape from shadow). При использовании нескольких кадров трёхмерная реконструкция выполняется путем сопоставления одних и тех же ключевых точек на разных изображениях для получения информации о их пространственных координатах. Поскольку получение нескольких кадров связано, как правило, с собственным движением камеры, метод получил название "восстановление формы по движению"(Shape from motion).

Метод бинокулярного зрения основывается на биологической аналогии со зрением человека. В этом методе используются две камеры для получения воспринимаемого изображения объекта с двух разных ракурсов. Далее используется метод триангуляции для преобразования бинокулярного дисбаланса соответствующих точек в информацию о глубине.

Алгоритмической основой современных методов трехмерной реконструкции являются методы машинного обучения. Основные направления - статистическое обучение, нейросети глубокого обучения и семантические методы.

Метод статистического обучения основан на поиске закономерностей по большой базе данных. Он вычисляет статистику признаков по каждой цели в базе данных, определяет функцию вероятности признаков, сравнивает с функцией признаков данных и принимает максимальную аналогичную глубину в качестве глубины цели реконструкции. Метод статистического обучения широко используется для реконструкции больших сцен, лица и тела человека, и может применяться в других областях, таких как системы поиска и распознавания видео.

Нейросетевой подход заключается в использовании нейронной сети с хорошей функциональной аппроксимационной способностью для нелинейных функций. Идея использования глубокого обучения - пропустить такие шаги, как извлечение и сопоставление признаков, определение параметров камеры, и смоделировать форму трехмерного объекта прямо из исходного изображения (зачастую - только из одного). Это, пожалуй, один из наиболее популярных подходов в области компьютерного зрения.

Семантические подходы описывают изображение на уровне высокоуровневого описания отдельных взаимодействующих объектов сцены. Использование такого описания позволяет повысить надежность детектирования отдельных объектов при их сложных взаимодействиях - например, при взаимном перекрытии,

попадании в тень и т.д. Обычно данные методы не являются самостоятельными, а являются надстройкой над другими методами, и позволяют устранить неоднозначности реконструкции и повысить ее точность.

Анализ существующих методов трехмерной реконструкции показывает, что ни один из них не может быть напрямую использован для задачи навигации автономного робота. Это заставляет искать как новые подходы, так и комбинировать существующие, для обеспечения желаемой производительности, точности и надёжности.

3D surface reconstruction in the task of autonomous navigation of a Martian rover

Bobkov Alexander[1] Alexander.Bobkov@bmstu.ru
Dai Yifan[1]⋆ Daiyifan1997@outlook.com
[1]Moscow, Bauman MSTU

Obtaining three-dimensional information about objects from the observed two-dimensional image is an important, interesting and still not completely solved problem in the field of computer vision. The problem of three-dimensional reconstruction is of particular relevance in applications related to autonomous navigation of robots over rough terrain, for example, for navigation of a Martian rover.

3D scene reconstruction from images is widely used in artificial intelligence, robotics, self-driving, SLAM (simultaneous localization and display), VR, AR, 3D printing, etc. This work summarizes and analyses the methods of three-dimensional reconstruction based on on the video image.

All three-dimensional reconstruction methods can be divided into contact and non-contact. Contact methods mainly include: trigger measurement, continuous measurement, coordinate measuring machines and robotic arms. Contact methods use specific tools and can directly obtain 3D information about a scene with high precision. They are suitable for cases where you can directly contact objects in the scene and measure them. However, in many cases, contact methods are unacceptable because they can damage both the surface of the object and the measuring tool.

The contactless method is free from this drawback. He typically uses image analysis techniques to obtain information about the scene. The non-contact method has a wider range of applications, although the reconstruction accuracy is significantly lower than that of the contact method.

Contactless methods can be divided into active and passive, depending on the method of obtaining information about the depth of the scene. Active include radio range finders, laser scanning, structured light method. Information about the depth of the object is measured directly, and can be obtained by exposing the target object to a radiation source (laser, ultrasound, electromagnetic waves, etc.) and then analysing the reflected signal, after which the approximate shape of the object is calculated.

The technology of active reconstruction is well developed, has high accuracy, but the equipment is usually quite complex, expensive and sensitive to external influences. Passive methods that use video images from one or several cameras are devoid of these drawbacks. Passive methods take a sequence of images with a camera and reconstruct a three-dimensional structural model of the scene based on an understanding of the two-dimensional image.

Passive methods can be divided into monocular and binocular (with multiple cameras).

Monocular vision can use the camera to capture one or more images from a single point, or capture multiple images from multiple points. In a single frame, scene depth information is extracted from the 2D characteristics of the image (shading, texture, focus, outline, etc.). This group of methods is referenced as Shape from Shadow. When using several frames, three-dimensional reconstruction is performed by comparing the same key points on different images to obtain information about their spatial coordinates. Since the acquisition of several frames is associated, as a rule, with the camera's own movement, the method is referenced as Shape from Motion.

The binocular vision method is based on a biological analogy with human vision. This method uses two cameras to capture a perceived image of an object from two different angles. Next, a triangulation method is used to convert the binocular imbalance of the corresponding points into depth information.

The algorithmic basis of modern three-dimensional reconstruction methods is machine learning methods. The main directions are statistical learning, deep learning neural networks and semantic methods.

The statistical learning method is based on searching for patterns in a large database. It calculates feature statistics for each target in the database, determines the feature probability function, compares with the data feature function, and takes the maximum similar depth as the depth of the reconstruction target. The statistical learning method is widely used to reconstruct large scenes, the face and body of a person, and can be applied in other areas such as video search and recognition systems.

The neural network approach is to use a neural network with good functional approximation ability for nonlinear functions. The idea behind using deep learning is to skip steps such as feature extraction and matching, determine camera parameters, and simulate the shape of a 3D object directly from the original image (often just one). This is perhaps one of the most popular approaches in computer vision.

Semantic approaches describe the image at the high-level description of individual interacting objects in the scene. The use of such a description makes it possible to increase the reliability of detecting individual objects during their complex interactions - for example, when they overlap, fall into the shadow, etc. Usually, these methods are not independent, but are a superstructure on top of other methods, and allow you to eliminate the ambiguities of the reconstruction and increase its accuracy.

Analysis of the existing methods of three-dimensional reconstruction shows that none of them can be directly used for the task of navigating an autonomous robot. This forces us to look for both new approaches and combine existing ones to ensure the desired performance, accuracy and reliability.

Методы компьютерного зрения на основе спектральной теории графов

Захаров Алексей Александрович[1]⋆ aa-zaharov@ya.ru
Жизняков Аркадий Львович[1] lvovich1975@mail.ru

[1]Муром, Муромский институт (филиал) ФГБОУ ВО «Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых»

Использование структурных подходов в области компьютерного зрения позволяет значительно повысить качество результатов. Основная идея заключается в описании отношений между частями сцены. Это позволяет рассматривать группу объектов как единое целое. Преимуществом структурных методов является возможность анализа сцен на основе малого количества составляющих и правил формирования графической модели.

При разработке представляемых методов компьютерного зрения использовалась спектральная теория графов. Спектральное представление основано на разложении матрицы графа в спектр. Спектральное представление при использовании в компьютерном зрении имеет следующие достоинства: является инвариантным к порядку вершин; описывает структуру графа в виде числовых значений, что удобно для вычислений; задача вычисления собственных значений хорошо изучена и решается за полиномиальное время.

Разработаны следующие методы компьютерного зрения на основе спектральной теории графов: методы нахождения соответствий на изображениях, методы сегментации изображений, метод выделения значимых областей изображения, методы распознавания объектов на изображениях на основе вложения графов в векторное пространство. Разработанные методы при исследовании на некоторых наборах данных превосходят существующие подходы компьютерного зрения по ряду показателей.

Работа выполнена при финансовой поддержке Министерства науки и высшего образования РФ (Госзадание ВлГУ ГБ-1187/20).

[1] *Захаров А. А., Титов Д. В., Жизняков А. Л., Титов В. С.* Метод визуального внимания на основе ранжирования вершин графа по разнородным признакам изображений // Компьютерная оптика, 2020. Т. 44(3). С. 427–435.

Computer vision methods based on spectral graph theory

Zakharov Alexey[1]⋆ aa-zaharov@ya.ru
Zhiznyakov Arkady[1] lvovich1975@mail.ru
[1]Murom, Murom Institute (branch), Vladimir State University named after Alexander and Nikolay Stoletovs

The use of structural approaches in the field of computer vision can significantly improve the quality of results. The main idea is to describe the relationship between the parts of the scene. This property describes a group of objects as a whole. The advantage of structural methods is the ability to analyze scenes based on a small number of components and rules for the formation of a graphic model.

Spectral graph theory was used in the development of the presented computer vision methods. Spectral representation is based on the decomposition of a graph matrix into a spectrum. Spectral representation when used in computer vision has the following advantages: is invariant to the order of vertices; describes the structure of the graph in the form of numerical values, which is convenient for calculations; the problem of calculating the eigenvalues is well studied and can be solved in polynomial time.

The following computer vision methods based on spectral graph theory have been developed: image matching methods, image segmentation methods, a method for selecting significant areas of the image, methods of object recognition in images based on the embedding of graphs in a vector space. The developed methods in the study on some data sets are superior to existing computer vision methods in a number of indicators.

This work was financially supported by the Ministry of Science and Higher Education of the Russian Federation (State task of VlSU GB-1187/20).

[1] *Zakharov A., Titov D., Zhiznyakov A., Titov V.* Visual attention method based on vertex ranking of graphs by heterogeneous image attributes // Computer Optics, 2020. Vol. 44(3). Pp. 427–435.

Уменьшение числа ложных срабатываний детектора для домофонов с био-идентификацией

Свитов Давид Вячеславович[1,2]⋆ d.svitov@expasoft.tech
Алямкин Сергей Анатольевич[1] s.alyamkin@expasoft.ru
[1]г. Новосибирск, ООО "Экспасофт"
[2]г. Новосибирск, Институт автоматики и электрометрии СО РАН

Свёрточные нейронные сети (СНН) широко применяются в задачах детектирования объектов на изображениях. Но детекторы всё ещё часто подвержены ложным срабатываниям. Обработка региона, не содержащего изображения человека, может привести к недетерминированному поведению системы. В данной работе предлагается подход к детектированию, позволяющий снизить число ложных срабатываний за счёт обработки только движущихся объектов - подходящих к домофону людей.

Предлагаемый подход заключается в модификации уже обученной на детекцию СНН и может быть применён для повышения точности имеющейся системы путём небольших изменений в ней. В данной работе предлагается использовать промежуточные карты признаков сети детектора, что не увеличивает вычислительную нагрузку и может быть применено для развёртывания на встраиваемых системах. Эффективность предлагаемого подхода была продемонстрирована на открытом наборе данных CDNet2014 pedestrian.

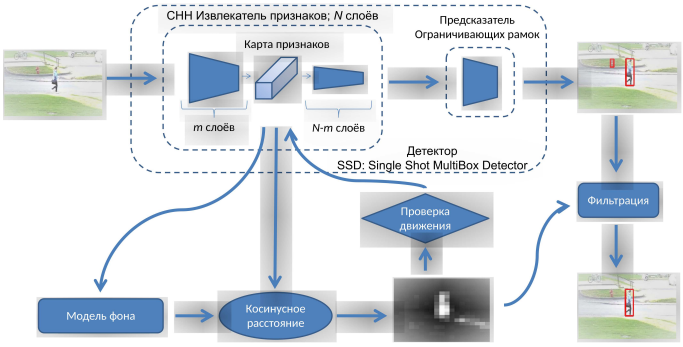


Рис. 1. Принцип работы фильтра ложных срабатываний детектора

Входное изображение обрабатывается первыми m слоями извлекателя признаков, входящим в состав модели детектирования объектов (Рис. 1). Полученные на этом этапе карты признаков используются для построения модели фона. Карта признаков для текущего кадра сравнивается с моделью фона по косинусной близости. Полученная после этого карта движений используется для проверки необходимости заканчивать выполнение детектора. Карта движений

также используется для фильтрации ограничивающих рамок, чтобы отбраковать ложные срабатывания, соответствующие статичным объектам.

Предложенный метод позволяет получить ускорение обработки потокового видео за счёт остановки обработки статичных кадров, а также снизить число ложных срабатываний детектора для задач, в которых необходимо детектировать движущиеся объекты.

Для сравнения предложенного детектора с базовым подходом, эксперименты проводились на архитектуре SSD [1] с извлекателем признаков MobileNetV2 [2] со входом 300x300. Выбор данной архитектуры был обусловлен её популярностью для встраиваемых систем. В экспериментах оценивались скорость и точность предложенного алгоритма. Скорость оценивалась как среднее время обработки кадра для всех видео в рассматриваемом датасете. Точность оценивалась как метрика mean average precision (MAP).

Была проведена симуляция случаев, когда детектор осуществляет ложное срабатывание на статичный объект, схожий с объектом целевого класса. Для этого в кадры исходного видео было добавлено статичное изображение объекта, напоминающее человека. Также видео, получаемые с камеры домофона, подвержены зашумлению из-за качества матрицы захватывающего устройства. Для того, чтобы приблизить набор данных к реальному случаю, каждый кадр модифицировался зашумлением. Исходные кадры умножались на Гауссовский шум с $\mu = 0.8$ и $\sigma = 0.2$, симулируя несовершенства устройства захвата видео.

Предложенный подход позволяет снизить среднее время обработки кадра. Для сравнения среднего времени обработки кадра были выполнены замеры времени на Intel Core i5-4210U CPU 1.70GHz × 4. Данный метод позволяет сократить среднее время на обработку кадра на 20% за счёт эффективной фильтрации кадров без движения, что достигается использованием для сравнения кадров карт признаков из промежуточного слоя сети.

Таблица 1. Сравнение MAP для SSD с извлекателем признаков MobileNetV2 и предложенного подхода на основе SSD+MobileNetV2.

Модель	MAP
SSD+MobileNetV2	0.326
Предложенный подход	0.548

Было проведено сравнение точности предложенного подхода на основе SSD+MobileNetV2 со стандартным использованием SSD+MobileNetV2 (Таблица 1).

[1] *Liu W. et al.* Ssd: Single shot multibox detector // European conference on computer vision, 2016. Pp. 21–37.

[2] *Sandler M. et al.* Mobilenetv2: Inverted residuals and linear bottlenecks // Proceedings of the IEEE conference on computer vision and pattern recognition, 2018. Pp. 4510–4520.

Reducing the number of false positive detections for intercoms with bio-identification

Svitov David[1,2]⋆ d.svitov@expasoft.tech
Alyamkin Sergey[1] s.alyamkin@expasoft.ru
[1]Novosibirsk, Expasoft LLC
[2]Novosibirsk, Institute of Automation and Electrometry of the SB RAS

Convolutional neural networks (CNNs) are widely used for object detection in images. But detectors are still prone to false detections. Processing a region that does not contain a human image may result in non-deterministic system behavior. In this paper, we propose an approach to object detection, which makes it possible to reduce the number of false-positive detections by processing only moving objects, like people approaching the intercom.

The proposed approach is modification of the CNN already trained for object detection task. This method can be used to improve the accuracy of an existing system by applying minor changes to the existing algorithm. It is proposed to use intermediate feature maps of the detector, which does not increase the computational load and can be used for deployment on embedded systems. The efficiency of the proposed approach was demonstrated on the open dataset "CDNet2014 pedestrian".

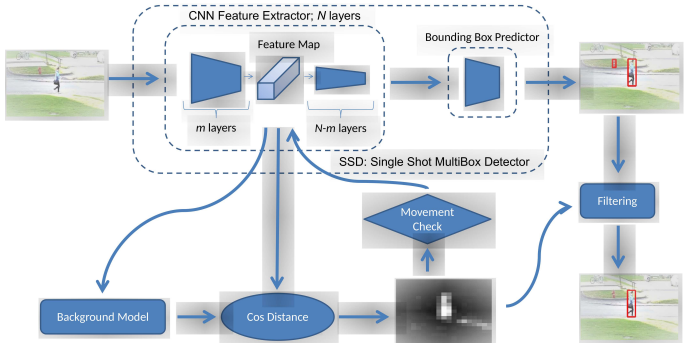


Fig. 1. Algorithm for filtering false positive detections

The input image is processed by the first m layers of the feature extractor, which is part of the object detection model (Fig. 1). The feature maps obtained at this stage are used to build a background model. The feature map for the current frame is compared to the background model by cosine similarity. The resulting motion map is used to check whether it is necessary to end the inference of the detector. The motion map is also used to filter the bounding boxes to filter out false-positives detections corresponding to static objects.

This method allows us to reduce the average processing time of a video frame and reduce the number of false positive detections for tasks in which it is necessary to detect moving objects.

To compare proposed approach with the basic approach, we conducted experiments on the SSD [1] architecture with the MobileNetV2 [2] feature extractor with an input of 300x300. The choice of this architecture was due to its popularity for embedded systems. In our experiments, we evaluated the speed and accuracy of the proposed algorithm. The speed was measured as the average frame processing time for all videos in the dataset. Precision was measured as a mean average precision (MAP) metric.

We have simulated cases where the detector performs a false positive detection on a static object which looks like the object of the target class. To do this we added a static image of an object resembling a human to the frames of the original video. Also, videos received from the intercom camera are subject to noise due to the quality of the capture device. In order to bring the data set closer to the real case, each frame was modified with noise. The original frames were multiplied by Gaussian noise with $\mu = 0.8$ and $\sigma = 0.2$, simulating artifacts of the video capture device.

The proposed approach reduces the average frame processing time. To compare the average frame processing time, we measured the time on the Intel Core i5-4210U CPU 1.70GHz $\times$ 4. This approach allows us to reduce the average time for frame processing by 20% due to effective filtering out frames without movement. This is achieved by using feature maps from the network intermediate layer to compare frames.

Table 1. Comparison of MAP for SSD with MobileNetV2 feature extractor and proposed approach based on SSD + MobileNetV2.

Model	MAP
SSD+MobileNetV2	0.326
Proposed approach	0.548

We compared the accuracy of an AmphibianDetector based on SSD + MobileNetV2 with the baseline solution based on SSD + MobileNetV2 (Table 1).

[1] *Liu W. et al.* Ssd: Single shot multibox detector // European conference on computer vision, 2016. Pp. 21–37.

[2] *Sandler M. et al.* Mobilenetv2: Inverted residuals and linear bottlenecks // Proceedings of the IEEE conference on computer vision and pattern recognition, 2018. Pp. 4510–4520.

Дистилляция моделей для распознавания лиц, обученных с применением softmax с отступами

Свитов Давид Вячеславович[1,2]⋆ d.svitov@expasoft.tech
Алямкин Сергей Анатольевич[1] s.alyamkin@expasoft.ru
[1]г. Новосибирск, ООО "Экспасофт"
[2]г. Новосибирск, Институт автоматики и электрометрии СО РАН

Использование сверточных нейронных сетей (CNN) в сочетании с softmax с отступом демонстрирует наилучшую точность в задаче распознавания лиц. Недавно были представлены облегченные модели нейронных сетей, обученные с помощью softmax с отступом, для задачи верификации по лицу на встраиваемых устройствах. В данной работе предлагается метод дистилляции, который позволяет получить большую точность, чем другие методы для задачи распознавания лиц на наборах данных LFW, AgeDB-30 и Megaface. Основная идея предлагаемого подхода заключается в использовании центров классов сети учителя для инициализации сети ученика. Затем сеть ученик обучается получать углы между центрами классов и векторами лиц, которые равны углам сети учителя.

Существует несколько вариантов softmax с отступом, используемых для обучения нейронных сетей для задач распознавания лиц: Cosface, Sphereface и ArcFace.

Для обучения легковесных нейронных сетей для распознавания лиц с помощью softmax с отступом, используются следующие методы дистилляции: Триплетная дистилляция [1], Дистилляция по углу [2] и Дистилляция на основе отступа [3].

Интуиция, лежащая в основе предлагаемого подхода, заключается в том, чтобы стягивать вектора для лиц и центры классов для сети ученика, когда они близки для сети учителя. Это позволяет передачи данных от учителя к ученику быть более эффективной, так как сеть ученик акцентирует внимание на примерах с более уверенными предсказаниями сети.

Пусть $x_{s_i} \in R^D$ обозначает вектор признаков сети ученика для экземпляра данных с номером i, $x_{t_i} \in R^D$ обозначает вектор признаков сети учителя для того же экземпляра. Обозначим матрицы весов последних слоёв сетей ученика и учителя соответственно $W_s \in R^{D \times n}$ и $W_t \in R^{D \times n}$. Столбец с индексом j соответствующий центру класса y_i обозначается $W_{s_j} \in R^D$ и $W_{t_j} \in R^D$ для сетей ученика и учителя.

Методы, основанные на добавлении отступа m в функцию $softmax$, нормализуют матрицу весов и вектор признаков на 1: $||W_j|| = 1$ и $||x_i|| = 1$. Такая нормализация позволяет рассматривать выход $logit$ слоя как косинус угла между вектором признаков и соответствующим центром класса: ${W_j}^T x_i = ||W_j|| \cdot$ $\cdot\, ||x_i|| cos(\theta_j) = cos(\theta_j)$. ArcFace, как метод позволяющий достичь наибольшую точность в задаче верификации лиц, рассматривается как частный случай обу-

чения с softmax с отступом:

$$L_{\text{ArcFace}} = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{s(cos(\theta_{y_i}+m))}}{e^{s(cos(\theta_{y_i}+m))}+\sum_{j=1,j\neq y_i}^{n} e^{s\,cos\theta_j}}. \quad (1)$$

В ArcFace отступ m имеет фиксированное значение 0.5. В данной работе предлагается дистиллировать знания из сети учителя, вычисляя значения отступа m для каждого экземпляра данных i. Предложенный метод дистилляции основан на двух ключевых идеях:

— Центры классов, найденные сетью учителем, используются для инициализации сети ученика: $W_s = W_t$. Так как центры классов обучаемые значения, сеть с большим числом параметров способна выучить лучшие положения классов на гиперсфере.
— Вычисляемые значения отступа m_i используются для дистилляции. Они контролируют значения между векторами x_{s_i} и соответствующими центрами классов W_{s_j}: большие значения m_i приводят к сильному стягиванию векторов x_{s_i} к центру классов. Предлагается вычислять значение m_i основываясь на информации от сети учителя.

В экспериментах использовалась сеть ResNet100 в качестве сети учителя. Легковесная архитектура с малым числом параметров MobileFaceNet использовалась как сеть ученик. Все методы дистилляции сравнивались по одному сценарию - знания передавались из обученной сети ResNet100 в MobileFaceNet.

Таблица 1. Точность верификации на датасетах LFW, AgeDB-30 и MegaFace

Архитектура	Обучение	LFW	AgeDB-30	MegaFace
ResNet100 (учитель)	ArcFace	99.76	98.21	98.35
MobileFaceNet (ученик)	ArcFace	99.51	96.13	90.62
MobileFaceNet	Триплетная по L2	99.56	96.23	89.10
MobileFaceNet	Триплетная по cos	99.55	95.60	86.52
MobileFaceNet	Дис. с отступом	99.41	96.01	90.77
MobileFaceNet	По углу	99.55	96.01	90.73
MobileFaceNet	Предложенный подход	**99.61**	**96.55**	**91.70**

Датасет MegaFace наиболее сложный из рассмотренных, он включает большее число людей и изображений, чем остальные. На датасете MegaFace сеть учитель достигает 98.35%. Сеть ученик, обученная по методу ArcFace, достигает 90.62%. Как показано в Таблице 1, предложенный метод демонстрирует наибольшую точность 91.70%.

[1] *Feng Y. et al.* Triplet distillation for deep face recognition // 2020 IEEE International Conference on Image Processing (ICIP), 2020. Pp. 808–812.

[2] *Duong C. N. et al.* Shrinkteanet: Million-scale lightweight face recognition via shrinking teacher-student networks // arXiv preprint arXiv:1905.10620, 2019.

[3] *Nekhaev D., Milyaev S., Laptev I.* Margin based knowledge distillation for mobile face recognition // Twelfth International Conference on Machine Vision (ICMV 2019), 2020. Vol. 11433.

Distillation for face recognition neural networks with margin-based softmax

Svitov David[1,2]⋆ d.svitov@expasoft.tech
Alyamkin Sergey[1] s.alyamkin@expasoft.ru
[1]Novosibirsk, Expasoft LLC
[2]Novosibirsk, Institute of Automation and Electrometry of the SB RAS

The usage of convolutional neural networks (CNNs) in conjunction with the margin-based softmax approach demonstrates the state-of-the-art performance for the face recognition problem. Recently, lightweight neural network models trained with the margin-based softmax have been introduced for the face identification task for edge devices. In this paper, we propose a distillation method for lightweight neural network architectures that outperforms other known methods for the face recognition task on LFW, AgeDB-30 and Megaface datasets. The idea of the proposed method is to use class centers from the teacher network for the student network. Then the student network is trained to get the same angles between the class centers and face embeddings predicted by the teacher network.

There are several variations of the margin-based softmax used for training of neural networks for the face recognition problem. They include Cosface, Sphereface and ArcFace approaches.

For training lightweight face recognition neural networks with the margin-based softmax, the following distillation methods are used: Triplet distillation[1], Angular distillation [2] and Margin Based Knowledge Distillation[3].

The intuition behind the proposed method is to pull feature vectors and class center closer to each other for the student network when these vectors are close for the teacher network. It allows the knowledge transfer from the teacher to the student to be more efficient, because the student network focuses on samples with more confident predictions while paying less attention to samples with the low confidence.

Let $x_{s_i} \in R^D$ denote the feature vector of student network for the sample with number i, $x_{t_i} \in R^D$ denotes the feature vector of teacher network for the same sample. We will denote the weight matrices of the last layer of student and teacher networks respectively by $W_s \in R^{D \times n}$ and $W_t \in R^{D \times n}$. The column with the index j corresponding to the center of the class y_i will be denoted by $W_{s_j} \in R^D$ and $W_{t_j} \in R^D$ for the student and teacher networks.

Methods based on adding the margin m to the $softmax$ function normalize the weight matrix and sample vectors by 1: $||W_j|| = 1$ and $||x_i|| = 1$. This normalization allows considering the output of the *logit* layer as the cosine of the angles between the sample vectors and corresponding class centers: ${W_j}^T x_i = ||W_j|| \cdot ||x_i|| cos(\theta_j) = = cos(\theta_j)$. We will consider ArcFace as a special case of the margin-based softmax

approach since it gives the best performance among the margin-based methods:

$$L_{\text{ArcFace}} = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{s(cos(\theta_{y_i}+m))}}{e^{s(cos(\theta_{y_i}+m))} + \sum_{j=1,j\neq y_i}^{n} e^{s\ cos\theta_j}}. \quad (1)$$

In ArcFace, the margin m is fixed at 0.5. We propose to distil the knowledge from the teacher network by calculating the margin values m for each sample i. The proposed distillation method contains two key ideas:

— Class centers found by the teacher network are used for the student network: $W_s = W_t$. Since the class centers are learning values, a deeper network is able to learn more optimal position of classes on the hypersphere.

— The calculated margin values m_i are used for distillation. They explicitly control the distance between vectors x_{s_i} and corresponding class centers W_{s_j}: larger m_i leads to the stronger attraction of the vector x_{s_i} to the class center. It is proposed to calculate m_i based on the information from the teacher.

In our experemnts, the ResNet100 architecture was chosen as a teacher network. The novel lightweight architecture called MobileFaceNet was used as the student network. All distillation methods were compared in the same scenario where the knowledge was transferred from the trained ResNet100 to MobileFaceNet.

Table 1. Verification Accuracy at LFW, AgeDB-30 and MegaFace

Architecture	Training	LFW	AgeDB-30	MegaFace
ResNet100 (teacher)	ArcFace	99.76	98.21	98.35
MobileFaceNet (student)	ArcFace	99.51	96.13	90.62
MobileFaceNet	Triplet dist. L2	99.56	96.23	89.10
MobileFaceNet	Triplet dist. cos	99.55	95.60	86.52
MobileFaceNet	Margin based	99.41	96.01	90.77
MobileFaceNet	Angular dist.	99.55	96.01	90.73
MobileFaceNet	Our	**99.61**	**96.55**	**91.70**

MegaFace dataset is most challenging, which includes a much larger number of people and images then other datasets. On the MegaFace dataset the teacher network reaches 98.35%. The student network trained with ArcFace reaches 90.62%. As shown in Table 1, our method demonstrates the best accuracy of 91.70%.

[1] *Feng Y. et al.* Triplet distillation for deep face recognition // 2020 IEEE International Conference on Image Processing (ICIP), 2020. Pp. 808–812.

[2] *Duong C. N. et al.* Shrinkteanet: Million-scale lightweight face recognition via shrinking teacher-student networks // arXiv preprint arXiv:1905.10620, 2019.

[3] *Nekhaev D., Milyaev S., Laptev I.* Margin based knowledge distillation for mobile face recognition // Twelfth International Conference on Machine Vision (ICMV 2019), 2020. Vol. 11433.

Идентификация человека в реальном времени с помощью конструктора приложений MATLAB на основе YOLO v2 и VGG 16

Бобков Александр Валентинович[1] alexander.bobkov@bmstu.ru
Хтет Аунг[1]⋆ happyland27057@gmail.com
[1]Москва, МГТУ им. Н.Э.Баумана

Задача распознавания лиц - одна из наиболее интересных тем в области компьютерного зрения - это способность распознавать или идентифицировать личность человека, анализируя различные черты лиц. Система распознавания лиц обеспечивает огромные преимущества по сравнению с другими решениями биометрической безопасности, такими как распознавание радужной оболочки глаза и отпечатков пальцев. Распознавание лиц сегодня имеет множество методов в своем применении. Например, метод главных компонент, Гистограмма направленности совместно с машиной опорных векторов (HOG + SVM), сети глубокого обучения и так далее.

В системах распознавания лиц есть две главные части, которые являются обнаружением лиц и идентификацией лиц. Для первого этапа разработан метод YOLO v2 [1], а для второго этапа разработан VGG16 в предлагаемой нами системе распознавания лиц [2].

Для обнаружения лиц использовалась сеть YOLOv2 на основе предобученной сети ResNet-18.

Выход алгоритма обнаружения объектов YOLOv2 – отклик размера S*S, где S – количество ячейеек сетки. Каждая ячейка собержит пять параметров (x, y, w, h) и Pr(obj), где x,y – координаты центра ограничивающей рамки, w,h – её ширина и высота, Pr(obj) - вероятность нахождения объекта внутри рамки. Показатель достоверности отражает вероятность включения в модель целевого объекта и точность блока обнаружения предсказания. Показатель C(obj) достоверности определяется так:

C(obj) = Pr (obj) * IoU (Pred,Gtruth)

Если искомый объект отсутствует в ячейке, то Pr(obj) будет равен нулю, а доверительный балл должен быть равен нулю: C(obj) = 0.

IoU – это величина перекрытия найденной ограничивающей рамки и рамки из обучающей выборки, то есть отношение их пересечения и объединения:

IoU (Pred,Gtruth)= (перекрывающаяся область предсказанной рамки и рамки обучающей выборки)/(вся область предсказанной рамки и рамки обучающей выборки)

YOLO v2 использует суммарную квадратическую ошибку в качестве функции потерь. Метод пытается оптимизировать следующие многосоставные потери: потери локализации (ошибка опредаления положения объекта), потери доверия (ошибка определения вероятности обнаружения) и потери классификации (ошибка определения класса объекта).

Для распознавания лиц использовалась предварительно обученная модель сети VGG16. Модель VGG16 имеет большое количество гиперпараметров. Размер входного изображения первого слоя составляет 224x224 с кодированием RGB. Изображение пропускается через последовательность сверточных слоев, в которых использовался фильтр размером 3x3 с шагом 1, и всегда используется один и тот же слой заполнения и максимального объединения фильтра 2x2 с шагом 2. Расположение слоев в этой архитектуре выглядит следующим образом: сверточные слои, слои ReLU и слои максимального пула. В конце модели есть 2 полносвязных слоя, за которыми следует слой softmax для вывода. Эта сеть VGG16 является довольно большой сетью и имеет около 138 миллионов обучаемых параметров. При предъявлении сети изображения лица, на выходе сети появляется его описание в виде вектора признаков. При этом одинаковые лица будут иметь схожие признаки, а разные - соответственно, несхожие, даже при наличии мешающих факторов - изменения ракурса, освещенности и т.д. Это позволяет распознавать лица, заранее неизвестные сети, путём сравнения вектора признака, сгенерированного сетью, с ранее заданным образцом.

Для реализации системы, способной работать в режиме реального времени был использован тулбокс App designer среды MATLAB. Эксперименты показали, что система обнаружения лиц на базе YOLOv2 обладает не только высокой точностью, но и позволяет обнаруженивать человеческие лица на видео в режиме реального времени, с высокой скоростью обнаружения. Именно поэтому алгоритм назван “YOLO (You Only Look Once)” – «посмотри на изображение только один раз». Для этапа распознавания лиц мы использовали предварительно обученную сетевую модель VGG16, перенеся её в Матлаб. Для 105 разных людей с общим количеством изображений 5250 мы получили точность более 99.5 процентов. Это довольно хороший результат, достаточный для работы многих приложений захвата и распознавания изображения с видеокамеры в реальном времени.

[1] *Redmon J., Farhadi A.* YOLO9000: Better, Faster, Stronger // Institute of Electrical and Electronics Engineers (IEEE), 2017.

[2] *Simonyan K., Zisserman A.* Very Deep Convolutional Networks for Large-Scale Image Recognition // ICLR, 2015.

Identification of person in real-time using the MATLAB application designer based on YOLOv2 and VGG 16

Bobkov Alexander[1] alexander.bobkov@bmstu.ru
Aung Htet[1]⋆ happyland27057@gmail.com
[1]Moscow, Bauman MSTU

The task of face recognition is one of the most interesting topics in the field of computer vision is the ability to recognize or identify a person's personality by analyzing various facial features. The facial recognition system provides huge advantages over other biometric security solutions, such as iris and fingerprint recognition. Facial recognition today has many methods in its application. For example, Principal Component Analysis (PCA), Histogram of Gradient + Support Vector Machine (HOG + SVM), Deep Learning, and so on. Methodology: There are two main parts in face recognition systems, which are face detection and face identification. For the first stage, the YOLO v2 method was developed [1], and for the second stage, VGG 16 was developed in our proposed face recognition system [2].

Face detection is performed by the YOLOv2 algorithm based on ResNet-18. The output of the YOLOv2 object detection algorithm is a response of size S*S, where S is the number of grid cells. Each cell will contain five parameters (x, y, w, h) and Pr(obj), where x,y are the coordinates of the center of the bounding box, w,h are its width and height, Pr(obj) is the probability of box which exit an object. The confidence score reflects how likely the box contains an object (objectness) and how accurate is the boundary box.. The confidence score C(obj) is defined as:

C(obj) = Pr (obj) * IoU (Pred,Gtruth)

If the desired object is not in the cell, then Pr(obj) will be zero, and the confidence score should be zero: C(obj) = 0.

IoU is the amount of overlap between the predicted bounding box and the groundtruth box, that is, the ratio of their intersection and union: IoU (Pred,Gtruth)= (overlapping area of the predicted box and the groundtruth box)/(the entire area of the predicted box and the groundtruth box)

YOLO uses sum-squared error between the predictions and the ground truth to calculate loss. The loss function composes of the localization loss, the confidence loss and the classification loss.

Face recognition performed by using a pre-trained VG 16 model. The model of VGG16 has a large number of hyper-parameters. The input image size of the first layer is 224x224 with RGB. The image is passed through a stack of convolutional layers, which were used a 3x3 filter size with a stride 1 and always use the same padding and max-pooling layer of a 2x2 filter with stride 2. The arrangement of the layers in this architecture is as follows convolutional layers, ReLU layers, and max pool layers. The end of the model has 2 fully connected layers followed by a softmax for output. This VGG16 network is a pretty large network and it has about 138 million trainable parameters. When a face image is passed to the network, its

description appears at the output of the network in the form of a vector of features. The same faces have similar vectors, and different ones - respectively, dissimilar, even in the presence of interfering factors - changes in perspective, illumination, etc. This makes it possible to recognize faces unknown to network by means of a feature object generated by the network with a previously specified pattern.

Face detection by using the YOLOv2 algorithm has shown not only high accuracy, but also the ability to detect human faces in real-time with high detection speed. That is why the algorithm is called "YOLO (You Only Look Once)". And also for the facial recognition part, we have developed a pre-trained VGG 16 network model. For 105 different persons with a total of 5250 images, we got an accuracy of more than 96 percent. This is a pretty good result for the proposed system. And then we applied both results using the App designer in MATLAB for intended to use in real-time.

[1] *Redmon J., Farhadi A.* YOLO9000: Better, Faster, Stronger // Institute of Electrical and Electronics Engineers (IEEE), 2017.

[2] *Simonyan K., Zisserman A.* Very Deep Convolutional Networks for Large-Scale Image Recognition // ICLR, 2015.

Поиск заимствованных изображений в больших коллекциях научных документов

Бахтеев Олег Юрьевич[1,2,3] bakhteev@ap-team.ru
Горленко Татьяна Александровна [1] gorlenko@ap-team.ru
Каприелова Мариам Семеновна[1] kaprielova@ap-team.ru
Кильдяков Александр Сергеевич[1] kildyakov@ap-team.ru
Огальцов Александр Владимирович [1] ogaltsov@ap-team.ru
Финогеев Евгений Леонидович[1]⋆ finogeev@ap-team.ru
Чехович Юрий Викторович[1,3] chehovich@ap-team.ru

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

Мы рассматриваем задачу информационного поиска заимствованных изображений в больших коллекциях научных документов. Актуальность задачи обусловлена наличием прецедентов заимствования изображений из других источников в области медицины и биологии. Так в [1] показано, что в научных статьях области по биологии и медицине обнаруживается до 4 % проблемных изображений. При этом до настоящего момента не существовало автоматических средств, обеспечивающих поиск заимствованных изображений.

Наряду с этим разработаны и широко используются системы, которые способны обнаруживать некорректные заимствования в текстах [2]. В связи с этим, в данной работе представлен подход, который способен решать задачу поиска заимствованных изображений в больших коллекциях научных документов.

Предполагается, что заимствованное изображение сформировано из какого-то оригинального изображения находящегося в референтной коллекции. При этом к исходному изображению могли быть применены преобразования различных типов (изменение масштаба, компрессия, поворот, зеркальное отображение, перевод в серый цвет, выделение одного канала и т.д.). Пример оригинально изображения и сформированного из него заимствования представлены на рисунке 1.

(а) (б)

Рис. 1. (а) - оригинальное изображение; (б) - заимствование, полученное изменением цветов и зеркалированием относительно горизонтальной оси.

Поиск заимствований включает четыре этапа:

1) выделение изображений из документа;
2) фильтрация изображений;
3) поиск кандидатов в коллекции изображений из научных документов;
4) точное сопоставление кандидатов.

На первом этапе из документа извлекаются все изображения. Для этого каждая страница документа обрабатывается методами классического компьютерного зрения, которые выделяют вставленные на страницу изображения. Это позволяет извлекать изображения из страницы независимо от способа формирования документа.

На втором этапе из всех изображений документа исключаются графики, диаграммы, схемы. Это делается, чтобы избежать большого количества ложных срабатываний, так как изображения такого типа легко будут опознаваться как похожие с любым входящим изображением-графиком, потому что структуры изображений графиков и диаграмм зачастую очень похожи, однако природа может быть совершенно различной. Изображения, которые остаются после второго этапа считаются подходящими для поиска. В дальнейшем мы планируем разработать независимое решение, предназначенное для обработки графиков и диаграмм.

На третьем этапе происходит поиск кандидатов в коллекции изображений из научных документов. В результате поиска для каждого подходящего изображения формируется фиксированный набор кандидатов из коллекции. Особенностью данного этапа является то, что необходимо производить поиск в коллекции, которая может содержать миллионы изображений. При этом, очевидно, что для выполнения поиска за разумное время сопоставление с каждым изображением осуществляться не может.

На четвертом этапе происходит точное сопоставление кандидатов с подходящим изображением. Для сопоставления кандидатов вычисляется функция схожести между подходящим изображением и каждым кандидатом. На основании значений функции схожести определяется, является ли данное подходящее изображение заимствованием или нет.

Предложенный подход испытывался на коллекции с размером 1 млн изображений, полученных из документов, взятых из каталога журналов открытого доступа DOAJ [3]. На основании проведенных экспериментов данный подход показал свою эффективность и работоспособность, а также возможность применения на коллекциях большего объема.

Работа выполнена при поддержке Федерального государственного бюджетного учреждения «Фонд содействия развитию малых форм предприятий в научно-технической сфере» (Фонд содействия инновациям), проект № 63449.

[1] *Shen H.* Meet this super-spotter of duplicated images in science papers // Nature, 2020. Vol. 581. Pp. 132–136.

[2] *Журавлев Ю. И. и др.* Система распознавания интеллектуальных заимствований «Антиплагиат» // Математические методы распознавания образов, 2005. Т. 12(1). С. 329–332.

[3] Directory of Open Access Journals, URL: https://doaj.org

Image reuse detection in large-scale document scientific collection

Bakhteev Oleg[1,2,3] bakhteev@ap-team.ru
Gorlenko Tatiana[1] gorlenko@ap-team.ru
Kaprielova Mariam[1] kaprielova@ap-team.ru
Kildyakov Aleksandr[1] kildyakov@ap-team.ru
Ogaltsov Aleksandr[1] ogaltsov@ap-team.ru
Finogeev Evgeny[1] ⋆ finogeev@ap-team.ru
Chekhovich Yury[1,3] chehovich@ap-team.ru
[1]Moscow, Russia, Antiplagiat Company
[2]Moscow, Moscow Institute of Physics and Technolog
[3]Dorodnicyn Computing Centre, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences

We consider the problem of information retrieval of reused images in extensive collections of scientific documents. The problem's relevance is due to the presence of precedents for reusing images from other sources in the field of medicine and biology. Thus, in [1], it is shown that up to 4 % of misconducted images are found in scientific articles on biology and medicine. At the same time, until now, there were no automatic means to search for reused images.

Along with this, developed and widely used systems that can detect reused texts [2]. In this regard, this paper presents an approach that can solve the problem of finding reused images in extensive collections of scientific documents.

It is assumed that the reused image is formed from some original image in the reference collection. At the same time, various types of transformations (scaling, compression, rotation, mirroring, grayscaling, selection of one channel, etc.) could have been applied to the original image. An example of the original image and the reused one formed from it is shown in Figure 1.

(a) (b)

Fig. 1. (a) - original; (b) - reused image, obtained using colour change and mirroring relative to the horizontal axis

Searching for reused images involves four steps:

1) image selection from the document;
2) filtering the extracted images;
3) search of candidates in the collection of scientific documents;
4) exact match of candidates.

The first step is to extract all the images from the document. In order to get all the images, each page of the document is processed using the methods of classical computer vision, which highlight the images inserted on the page. This approach makes it possible to extract images from the page regardless of how the document is generated.

At the second stage, graphs, diagrams, schemes are excluded from all images of the document. This is done in order to avoid a large number of false positives since images of this type will be easily recognized as similar to any incoming graph because the structure of the images of graphs and diagrams are often very similar. At the same time, nature can be completely different. Images that remain after the second stage are considered suitable for searching. In the future, we plan to develop an independent solution for processing graphs and charts.

The third step is to search for a fixed set of candidates for each suitable image in the collection of scientific documents (index). The special feature of this stage is that it is necessary to search the index, which can contain millions of images. In this case, it is obvious that a comparison with each image can not be carried out to perform a search in a reasonable time.

The fourth step is to match candidates with the right image accurately. A similarity function is calculated between the matching image and each candidate to compare the candidates. Based on the values of the similarity function, it is determined whether a given matching image is reused or not.

The proposed approach was evaluated on collections with 1 million images from DOAJ [3] documents. DOAJ is the list of open access journals. Based on the conducted experiments, our approach has shown its efficiency and usability on larger collections.

This work was supported by FASIE project 63449.

[1] *Shen H.* Meet this super-spotter of duplicated images in science papers // Nature, 2020. Vol. 581. Pp. 132–136.

[2] *Zhuravlev Yu. et al.* The system of intellectual reuse detection Antiplagiat // Matematicheskie metody raspoznavanija obrazov, 2005. Vol. 12(1). Pp. 329–332.

[3] Directory of Open Access Journals, URL: https://doaj.org

Клонирование и конверсия произвольного голоса с использованием генеративных потоков

Обухов Дмитрий Сергеевич[1,2] bstodin@gmail.com
[1]Новосибирск, Новосибирский Государственный Технический Университет
[2]Новосибирск, Dasha.AI

В настоящее время задача синтеза речи стремительно расширяет сферу своего использования и уже находит применение в области медицины, в умных колонках, голосовых ассистентах и других окружающих человека умных устройствах, а также в различных задачах бизнеса. Одним из актуальных направлений развития синтеза речи сегодня является синтез голосом произвольного человека. Умение генерировать речь с заданным голосом является необходимым требованием для ряда задач, например, для построения диалоговых систем.

Современные подходы на основе глубокого обучения позволили эффективно и качественно формировать естественную речь голосом одного заданного диктора, представленного в наборе данных обучения. Предложенные недавно техники позволяют учитывать несколько дикторов при обучении, однако множество голосов, которыми формируется речь, по-прежнему остается ограниченным. Построение системы клонирования произвольного голоса становится следующим вызовом в области формирования речевых сигналов.

Задача клонирования голоса подразумевает использование заданного образца речи человека для синтеза таким же голосом речевого сигнала с произвольным заданным текстом. Важной отличительной чертой клонирования голоса от обычного синтеза речи является то, что обученная модель может синтезировать речь голосами даже тех спикеров, которые не были представлены в наборе данных обучения. Задача конверсии голоса заключается в преобразовании аудио сигнала с голосом исходного спикера в аудио сигнал с тем же лингвистическим содержанием, т.е. произнесенным текстом, но с произношением голосом целевого спикера. В совокупности задачи клонирования и конверсии голоса обеспечивают полный набор возможностей по преобразованию голоса речи - как для случая, когда исходная речь имеет текстовое представление, так и для случая, когда исходная речь задана в виде аудио сигнала.

Модели с использованием генеративных потоков недавно показали впечатляющие результаты в области синтеза речи, позволяя формировать разнообразные произнесения заданного текста. Предложенный в данной работе подход на основе потоковых генеративных моделей позволяет выполнять задачу клонирования голоса, за счет использования полученных из внешней системы вещественных векторов фиксированной размерности, содержащих информацию о спикере, т.н. эмбеддингов спикера. Кроме того, за счет своих архитектурных возможностей генеративные потоки позволяют одновременно с этим решать задачу конверсии голоса. Ещё один вклад настоящей работы заключается в предложенном способе учета переменной во времени информации о спикере, с целью повыше-

ния качества формируемого речевого сигнала. Благодаря этой технике система синтезирует более естественную речь голосом, похожим на заданный целевой голос, как в задаче клонирования голоса, так и в задаче конверсии голоса.

Voice cloning and any-to-any voice conversion using generative flows

Obukhov Dmitry[1,2] bstodin@gmail.com
[1]Novosibirsk, Novosibirsk State Technical University
[2]Novosibirsk, Dasha.AI

Currently, the speech synthesis task is rapidly expanding the scope of its use and is already finding application in the field of medicine, in smart speakers, voice assistants, and other human smart devices, as well as in various business tasks. One actual trend in speech synthesis research today is synthesis by the voice of an arbitrary person. The ability to generate speech with a given voice is a prerequisite for a number of tasks, for example, building dialog systems.

Modern deep learning approaches have made it possible to effectively and efficiently form natural speech with the voice of one given speaker presented in the training dataset. Recently proposed techniques make it possible to take into account several speakers, but the set of voices that form speech is still limited. Building a voice cloning system becomes the next challenge in the field of speech generation.

Voice cloning task implies using a given voice recording to synthesize the speech signal of an arbitrary given text with the same voice. An important distinguishing feature of voice cloning from conventional speech synthesis is that the trained model can synthesize speech with the voices of even those speakers that were not represented in the training dataset. The task of voice conversion is to convert the audio signal from the original speaker's voice into an audio signal in the same linguistic content, i.e. spoken text, but with the pronunciation of the target speaker's voice. Together, the tasks of cloning and converting voice provide a complete set of voice-to-speech conversion capabilities - both for the case when the original speech has a textual representation, and for the case when the original speech is specified as an audio signal.

Generative flow models have recently shown impressive results in the field of speech synthesis, allowing for the formation of various pronunciations of a given text. The proposed approach based on generative flow models allows one to perform the voice cloning task by using fixed-size vectors obtained from the external system containing information about the speaker, the so-called speaker embeddings. In addition, due to specific architecture, generative flows allow resolving the problem of voice conversion at the same time. Another contribution of this work is the proposed method for taking into account the time-dependent speaker information in order to improve the quality of the generated speech signal. With this technique, the system synthesizes more natural speech with a voice similar to the target voice, both in the voice cloning task and the voice conversion task.

Модели восстановления пропущенных данных во временных рядах концентрации углекислого газа

Алешновский Валентин Сергеевич[1]⋆ alesh-valentin@yandex.ru
Безрукова Александра Владимировна[1] aleksandra_bezrukova@mail.ru
Зюзина Нина Александровна[1] zjuzina.na15@physics.msu.ru
Газарян Варвара Арамовна[1,3] varvaragazaryan@yandex.ru
Курбатова Юлия Александровна[2] kurbatova.j@gmail.com
Чуличков Алексей Иванович[1] achulichkov@gmail.com
Шапкина Наталья Евгеньевна[1,4] neshapkina@mail.ru

[1]Москва, МГУ имени М.В.Ломоносова, физический факультет
[2]Москва, ИПЭЭ им. А.Н. Северцова РАН
[3]Москва, Финансовый университет при правительстве РФ
[4]Москва, ИТПЭ РАН

Для анализа временного ряда, определения его природы, прогнозирования (предсказания будущих значений временного ряда по настоящим и прошлым значениям), а также для управления процессом, порождающим данный ряд, необходимо построить математическую модель ряда динамики и интерпретировать результаты моделирования [1] [2].

В работе измерительных приборов иногда происходили сбои, в результате которых для некоторых моментов времени показания отсутствуют. Полноценно работать с таким рядом не представляется возможным, поскольку наличие временных промежутков с отсутствующими показаниями может плохо сказаться на построении математической модели и последующем анализе ряда. В связи с чем авторами были предложены два подхода к восстановлению пропущенных данных, а также проведено сравнение полученных результатов [3].

В ходе анализа зависимостей между измеренными значениями концентраций CO_2 на разных высотах над землей была выявлена корреляция между этими значениями концентрации на высотах 50 и 30 метров, которая была использована для восстановления пропущенных значений концентрации CO_2 на высоте 50 метров. Таким образом была построена “Модель-1”:

$$X_t^{50} = MX_t^{50} + \operatorname{cov}\left(X_t^{50}, X_t^{30}\right)\left(X_t^{30} - MX_t^{30}\right)/DX_t^{30},$$
$$DX_t^{50} = DX_t^{50} - \operatorname{cov}^2\left(X_t^{50}, X_t^{30}\right)/DX_t^{30},$$

где X_t^{30}, X_t^{50} - концентрации CO_2 на высоте 30 и 50 метров соответственно в момент времени t, M и D – знаки математического ожидания и дисперсии, $\operatorname{cov}\left(X_t^{50}, X_t^{30}\right)$ - ковариация между X_t^{30} и X_t^{50}. Математические ожидания, дисперсии и ковариация рассчитывались по выборке значений временных рядов в предыдущие моменты времени.

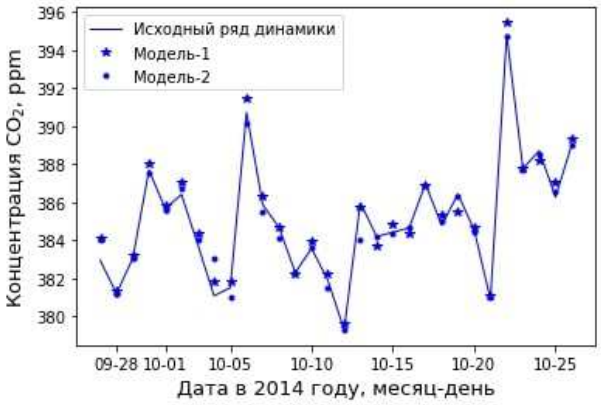
Во втором подходе для восстановления пропусков использовался алгоритм на основе модели авторегрессии - $ARIMA(p, d, q)$, где d - порядок интегрируемости, p - порядок авторегрессии и q - порядок скользящего среднего [5] [6].

Авторегрессионным называется процесс, в котором значение временного ряда находится в линейной зависимости от предыдущих значений временного ряда. Таким образом, “Модель-2” имеет вид:

$$\Delta^d X_t = c + \sum_{i=1}^{p} a_i \Delta^d X_{t-i} + \sum_{i=1}^{q} b_i \varepsilon_{t-i} + \varepsilon_t,$$

где ε_t - стационарный временной ряд, $c, a_i, i = 1, \ldots, p; b_j, j = 1, \ldots, q$ - параметры модели, Δ^d - оператор разности временного ряда порядка d. Параметры модели подбираются из условий максимального согласия результатов наблюдения с моделью.

На графике представлены исходный временной ряд концентрации углекислого газа за один месяц в 2014 году и оценки концентрации, вычисленные на основе моделей.



Для сравнения используются следующие критерии: модельная оценка среднеквадратичной погрешности оценивания (Погр. наил. с.к. оценивания) [4], среднеквадратическая ошибка (Mean Squared Error или MSE) как средний квадрат разности ряда и его оценки.

Период восстановления	Модель-1 неделя	Модель-2 неделя	Модель-1 2 мес	Модель-2 2 мес
Погр. наил. в с.к.оценивания	0,149	2,779	0,160	0,025
MSE	0,160	2,012	0,581	0,048

Таким образом, опираясь на значения погрешностей, можно сделать вывод о том, что “Модель-1” справляется лучше с краткосрочными пропусками, находящимися на небольшом расстоянии от начала отсчета. В свою очередь, “Модель-

2" показала себя лучше в восстановлении более длинных пропусков, когда доступна предыстория измерений, на которую и опирается метод при построении модели ряда и оценки концентрации.

Работа поддержана грантом РФФИ №19-29-09044

[1] *Kurbatova J., Tatarinov F., Molchanov A. et al.* Environ // Res. Lett, 2013.

[2] *Тимохина А. В.* Динамика концентрации атмосферного диоксида углерода над среднетаежными экосистемами Приенисейской Сибири (по данным измерений на обсерватории "ZOTTO") // Красноярск, 2017.

[3] *Вуколов Э.А.* Основы статистического анализа // М.: Форум, 2008. 464 с.

[4] *Пытьев Ю. П., Шишмарев И. А.* Теория вероятностей, математическая статистика и элементы теории возможностей для физиков, 2010.

[5] *Box G., Jenkins G.* Time series analysis: Forecasting and control. // San Francisco: Holden–Day, 1970.

[6] *Hocke K., Kämpfer N.* Gap filling and noise reduction of unevenly sampled data by means of the Lomb-Scargle periodogram // Atmos. Chem. Phys, 2009. No. 9. Pp. 4197–4206.

Recovery models of missing data in time series of carbon dioxide concentration

Aleshnovskiy Valentin[1]⋆ alesh-valentin@yandex.ru
Bezrukova Alexandra[1] aleksandra_bezrukova@mail.ru
Zyuzina Nina[1] zjuzina.na15@physics.msu.ru
Gazaryan Varvara[1,3] varvaragazaryan@yandex.ru
Kurbatova Julia[2] kurbatova.j@gmail.com
Chulichkov Alexey[1] achulichkov@gmail.com
Shapkina Natalya[1,4] neshapkina@mail.ru

[1]Moscow, Lomonosov Moscow State University, Faculty of Physics
[2]Moscow, IPEE named after A.N. Severtsov of the Russian Academy of Sciences
[3]Moscow, Financial University under the Government of the Russian Federation
[4]Moscow, ITPE RAS

It is necessary to create a mathematical model of the dynamics series and to interpret the results of modelling in order to analyse a time series and to determine its nature, to predict (predict future values of the time series by present and past values), as well as to control the process generating this series [1] [2].

While measuring data sometimes there are instrument failure, as a result of which for some moments of time there are no readings. Such series cannot be used for modelling because the presence of time intervals with missing readings can badly affect the construction of the mathematical model and the subsequent analysis of the series. In this connection, the authors have offered two approaches to restoration of the missed data, and also have carried out comparison of the received results [3].

During the analysis of dependencies between rows at neigh boring heights a correlation between values of CO_2 concentration at 50 and 30 meters was revealed, which was used to restore missed values of CO_2 concentration at an altitude of 50 meters. Thus, "Model-1" are constructed:

$$X_t^{50} = MX_t^{50} + \operatorname{cov}\left(X_t^{50}, X_t^{30}\right)\left(X_t^{30} - MX_t^{30}\right)/DX_t^{30},$$
$$DX_t^{50} = DX_t^{50} - \operatorname{cov}^2\left(X_t^{50}, X_t^{30}\right)/DX_t^{30},$$

where X_t^{30}, X_t^{50} - concentrations of CO_2 at an altitude of 30 and 50 meters, respectively, at time t, M and D - signs of expectation and dispersion, $\operatorname{cov}\left(X_t^{50}, X_t^{30}\right)$ - covariance between X_t^{30} and X_t^{50}. Mathematical expectations, dispersions and covariance were calculated from a sample of time series values at previous time points.

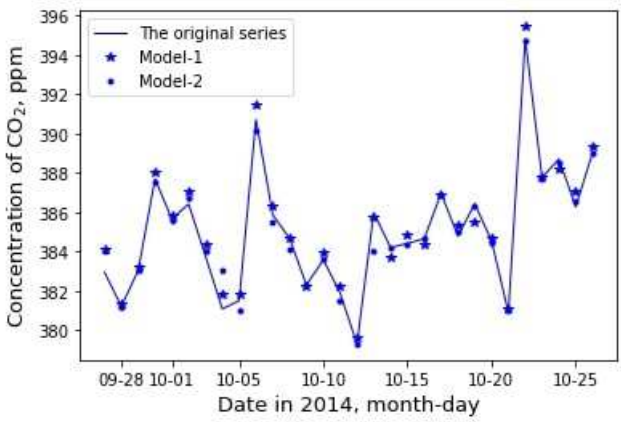
In the second approach, an algorithm based on an autoregression model was used to recover the gaps namely $ARIMA(p, d, q)$, where d is the order of integrability, p is the order of autoregression and q is the order of moving average [5] [6]. Autoregression is a process in which the values of the time series are in linear dependence

on the previous values. Thus, "Model-2" has the form:

$$\Delta^d X_t = c + \sum_{i=1}^{p} a_i \Delta^d X_{t-i} + \sum_{i=1}^{q} b_i \varepsilon_{t-i} + \varepsilon_t,$$

where ε_t is a stationary time series, $c, a_i, i = 1, \ldots, p; b_j, j = 1, \ldots, q$ - model parameters, Δ^d - the time series difference operator of order d. The model parameters are chosen from the conditions of maximum agreement of the observation results with the model.

The graph shows the original time series of the carbon dioxide concentration for the month in 2014 and the concentration estimates calculated using both models:



For comparison, the following criteria are used: model estimation of the mean square error (Model est. of the MSE) [4], Mean Squared Error (MSE) as the average square of the difference of the series and its estimates.

Recovery period	Model-1 week	Model-2 week	Model-1 2 months	Model-2 2 months
Model est. of the MSE	0,149	2,779	0,160	0,025
MSE	0,160	2,012	0,581	0,048

Thus, based on the error values, we can conclude that "Model-1" copes better with short-term passes, located at a short distance from the start of the reference. In turn, "Model-2" showed itself better in the recovery of longer skips, when the prehistory of measurements is available, as the method relies on these data when constructing the series model and estimating the concentration.

This research is funded by RFBR, grant 19-29-09044.

[1] *Kurbatova J., Tatarinov F., Molchanov A. et al.* Environ // Res. Lett., 2013.

[2] *Timokhina A.* Dynamics of atmospheric carbon dioxide concentration over the Middle Taiga ecosystems of Yenisei Siberia (according to measurements at the ZOTTO observatory) // Krasnoyarsk, 2017.

[3] *Vukolov E.* Fundamentals of statistical analysis // Moscow: Forum, 2008. 464 p.

[4] *Pytyev Yu., Shishmarev I.* Probability theory, mathematical statistics and elements of the theory of possibilities for physicists, 2010.

[5] *Box G., Jenkins G.* Time series analysis: Forecasting and control. // San Francisco: Holden-Day, 1970.

[6] *Hocke K., Kämpfer N.* Gap filling and noise reduction of unevenly sampled data by means of the Lomb-Scargle periodogram // Atmos. Chem. Phys, 2009. No. 9. Pp. 4197–4206.

Визуальная одометрия по изображениям опорной поверхности с малыми межкадро-выми поворотами

Минаев Евгений Юрьевич[1,2] e.minaev@gmail.com
Жердева Лариса Анатольевна[2]⋆ lara.zherdeva.taskina@gmail.com
Фурсов Владимир Алексеевич[1,2] fursov@ssau.ru
[1]Самара, Самарский университет
[2]Самара, Институт систем обработки изображений - филиал ФНИЦ «Кристаллография и фотоника» РАН

Системы визуальной одометрии, в которых последовательность видеокадров форми-руется с использованием камеры, направленной перпендикулярно вниз на поверхность в последние годы завоевывает все большую популярность [1, 2]. Возможно, это связано с тем, что не всегда есть надежные ориентиры в окружающих сценах, в то время как опорная поверхность наблюдается непрерывно. Например, в работе [3] приведен пример удачной реализации системы, основанной на корреляционном методе.

В работе [4] авторы настоящей работы предложили технологию визуальной одомет-рии по последовательности изображений опорной поверхности, регистрируемых БПЛА с малой высоты. Технология включает три этапа: определение сдвига и поворота с использование корреляции фрагментов (1), уточнение параметров сдвига и поворота методом оптического потока (2) и коррекция ошибок оценивания траекторий, связанных с неравномерностью движения и флуктуациями высоты (3). Предполагалось, что камера установлена в гироподвесе и сохраняет ориентацию относительно глобальной системы координат в течение всего полета.

Если камера жестко закреплена на корпусе движущегося аппарата, задача существенно усложняется. Для оценивания траектории необходимо определять не только межкадровый сдвиг, но и параметры поворота. В общем случае для этого необходимо строить последовательные оценки матриц сдвига и поворота по соответствующим точкам соседних кадров. Для этого кадры должны иметь значительные перекрытия, что легко достигается увеличением частоты съемки. При этом относительный поворот изображений кадров обычно оказывается небольшим.

С учетом указанной специфической особенности, в настоящей работе мы строим трехэтапную технологию визуальной одометрии, основанную на идеях, описанных в работе [4]. В данном случае важным отличием является то, что последовательность изображений опорной поверхности формируется камерой, жестко связанной с корпусом движущегося аппарата.

Рассматривается задача автономной навигации аппарата по последовательности изображений поверхности. Идея состоит в том, что для определения межкадровых поворотов на изображении задаются два фрагмента, разнесенные на большое расстояние (2*L), при котором фрагменты «видны» на обоих видах.

Размеры фрагментов задаются таким образом, чтобы сдвиг одноименных пикселей, соответствующих фрагментов на двух видах не превышал половины межпиксельного расстояния при максимальном межкадровом повороте. С учетом сказанного алгоритм строится в виде следующей последовательности шагов:

1. Задавая центры областей поиска в точках второго кадра определяем относительные целопиксельные координаты сдвига центров исходных фрагментов, заданных в соответствующих точках, методом экстремальной корреляции.

2. Уточняем оценки координат путем добавления малых приращений, вычисленных методом оптического потока. Для повышения точности кадры предварительно совмещаются так, чтобы относитель-ный сдвиг в любом направлении не превышал межпиксельного расстояния.

3. В предположении, что аппарат является жесткой конструкцией, вычисляем коор-динаты центральных точек и определяем общий сдвиг кадров.

4. Определяется угол межкадрового поворота, а текущее направление движения аппарата в произвольной точке траектории является суммой углов всех предшествующих данной точке пар кадров.

5. Далее для каждой точки траектории определяются приращения координат, где знаки приращений определяются текущими значениями угла.

6. Заключительный этап технологии состоит в коррекции полученных оценок коор-динат траектории. Цель этого этапа уменьшение ошибок, связанных с отклонениями калибровочного коэффициента от среднего значения вследствие неравномерности движений и малых вариаций расстояния от камеры до опорной поверхности. Алгоритм коррекции состоит в подсчете текущей «цены» пикселя на малых отрезках траектории и коррекции оценок координат в конце каждого такого отрезка на величину, определяемую степенью отклонения локального калибровочного коэффициента от среднего значения.

В заключении, получены результаты ошибок оценивания траектории (в см.), полученные при следующих значениях параметров алгоритмов: размеры фрагментов – 9×9; расстояние от центра кадра до фрагмента – L=100 (пикселей), соответственно расстояние между фрагментами 2*L=200 (пикселей); среднее значение калибровочного коэффициента K = 0.0782, общее число точек на траектории – 2111, длина участка для вычисления локальных калибровочных коэффициентов - 25 точек. При этом были получены относительные СКО: по координате X – 0.0647, по координате Y – 0.0491, по евклидовой метрике – 0.0569. Выявлено, что при отсутствии коррекции оценок ошибка одометрии возрастает при быстром возрастании текущего угла и существенно снижается в результате коррекции оценок траектории. Относительные ошибки в конечной точке траектории составили менее 0.02 и 0.04 для координат X и Y соответственно. Тестовые данные для приведенного примера можно найти на странице [4,5].

[1] *Muller P.* Flowdometry: An optical flow and deep learning based approach to visual odometry // In 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), 2017. Pp. 624–631.

[2] *Gonzalez R.* Improving robot mobility by combining downward-looking and frontal cameras // Robotics, 2016. Pp. 25–28.

[3] *Nourani-Vatani N.* Correlation-based visual odometry for ground vehicles // Journal of Field Robotics, 2011. Pp. 742–768.

[4] *Zherdeva L.* Synthetic dataset for navigation tasks of autonomous systems and ground robots // 2021 International Conference on Information Technology and Nanotechnology, 2021. Pp. 1–4.

[5] *Online resource* https://github.com/by-LZ-for/Synthetic-Dataset-Robot // Synthetic Dataset for Visual Odometry, 2021.

Visual odometry on surface image sequences with small inter-frame rotations

Minaev Evgeniy[1] e.minaev@gmail.com
Zherdeva Larisa[1]⋆ lara.zherdeva.taskina@gmail.com
Fursov Vladimir[1,2] fursov@ssau.ru
[1]Samara, Samara National Research University
[2]Samara, Image Processing Systems Institute of RAS – Branch of the FSRC ≪Crystallography and Photonics≫

Visual odometry systems use a frame sequence that formed by a camera directed perpen-dicularly downward on an underlying surface, have been gaining in popularity in recent years [1, 2]. Perhaps this is due to the lack of reliable reference points in the surrounding scenes for the visual system, while the surface monitored continuously. For example, in paper [3], an example of a successful implementation of a system based on the correlation method is given. In paper [4], the authors of this work proposed a visual odometry technology based on the image sequence of an underlying surface recorded by UAVs from a low altitude. Our technology includes three stages: determination of shift and rotation using fragment correlation (1), refinement of the shift and rotation parameters by the optical flow method (2) and correction of trajectory estimation errors associated with motion unevenness and height fluctuations (3). In the first work assumed that UAV's camera installed in a gyro suspension, thereby maintaining the camera orientation relative to the global coordinate system during entire flight. However, if the camera fixed rigidly to the moving vehicle body, the main problem becomes much more complicated. Thus, it is necessary to determine not only the interframe shift, but also the rotation parameters to estimate correctly the UAV's trajectory. In the general case, it is necessary to construct sequential estimates of shift and rotation matrices from the corresponding points in adjacent frames. Frames should have significant overlap to correspond the condition, which easily achieved by increasing the frequency of shooting. In this case, the relative rotation of frame images is usually small. Taking into account the described specific feature, in this work we build a threestage visual odometry technology based on article's ideas in [4]. In current paper case, an important dif-ference is that the image sequence of the underlying surface formed by a camera rigidly fixed to UAV's body.

The paper consider the problem of autonomous vehicle navigation while move along an image sequence of the underlying ground surface. Images are generated by the camera pointing perpendicularly downward. Since the distance from the camera to the surface is small, we neglect projective distortions when determining interframe shifts and rotations. In this case, for each frame of the sequence, we set a local coordinate system. For definiteness let us assume that the rectangular coordinate system is associated with the first image (by registration time), the point

(origin of the coordinate system) is at the center of frame, but the positive direction of the Ox axis coincides with the direction of vehicle movement.

The idea of the technique is that two fragments are set to determine interframe rotations in the image. These fragments are separated by a large distance (2 * L), at which the fragments are visible in both views. The fragment sizes are set so that the shift of same pixels of fragments did not exceed half the interpixel distance at maximum interframe rotation. Thus, the algorithm is constructed in the following sequence of steps:

1. By specifying the centers of the search areas at points for second frame, we determine the relative integer-pixel coordinates of the center shifts in the original fragments, given at points using the extreme correlation method.

2. Next, we refine the coordinate estimates by adding small increments calculated by the optical flow method. Frames are pre-aligned to improve accuracy so that the relative shift in any direction does not exceed the inter-pixel distance.

3. Assuming that the vehicle is a rigid structure, we calculate the coordinates of central points and determine the overall shift of frames.

4. The angle of interframe rotation, in accordance with the accepted designations, and the current direction of vehicle movement at an arbitrary point of the trajectory is the angle sum of all frame pairs preceding this point.

5. Further, for each trajectory point the coordinate increments are determinedm where the increment signs are determined by the current angle values.

6. The final stage of the technology is to correct the obtained estimates of trajectory coordi-nates. The purpose of this step is to reduce errors associated with deviations of calibration factor from the mean. This deviations occur due to uneven movement and small variations in the distance from camera to the underlying surface. The correction algorithm consists in calculating the current "price" of a pixel on small trajectory segments and correcting the coordinate estimates at the end of each such segment by the value - determined the degree of the local calibration coefficient deviation from the average value.

In conclusion, we got the trajectory estimation errors (in cm), obtained for following values of the algorithm parame-ters: fragment sizes – 9×9; the distance from the frame center to fragment is L=100 (pixels), the distance between the fragments is 2*L=200 (pixels) respectively; the average value of the calibration coefficient K = 0.0782, the total number of points on the trajectory is 2111, the sec-tion length for calculating local calibration coefficients is 25 points. In this case, the relative standard deviations were obtained: along the X coordinate - 0.0647, along the Y coordinate - 0.0491, and according to the Euclidean metric - 0.0569. It revealed that in the absence of estimate correction, the odometry error increas-es with a rapid change of the current angle, and is significantly reduced as a result of the correction of the trajectory estimates. The relative errors at the end point of the trajecto-ry were less than 0.02 and 0.04 for X and Y coordinates, respectively. Test data for the given example can be found on site [4,5].

[1] *Muller P.* Flowdometry: An optical flow and deep learning based approach to visual odometry // In 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), 2017. Pp. 624–631.

[2] *Gonzalez R.* Improving robot mobility by combining downward-looking and frontal cameras // Robotics, 2016. Pp. 25–28.

[3] *Nourani-Vatani N.* Correlation-based visual odometry for ground vehicles // Journal of Field Robotics, 2011. Pp. 742–768.

[4] *Zherdeva L.* Synthetic dataset for navigation tasks of autonomous systems and ground robots // 2021 International Conference on Information Technology and Nanotechnology, 2021. Pp. 1–4.

[5] *Online resource* https://github.com/by-LZ-for/Synthetic-Dataset-Robot // Synthetic Dataset for Visual Odometry, 2021.

Интерпретируемое распознавание изображений с помощью логических решающих функций

Бериков Владимир Борисович[1,2] berikov@math.nsc.ru
Козинец Роман Максимович[2]⋆ r.kozinets@g.nsu.ru
[1]Новосибирск, ИМ СО РАН
[2]Новосибирск, НГУ

Интерпретируемость предсказательных моделей глубокого обучения важна, особенно в тех случаях, когда речь идет о применениях в областях, где большое значение имеют вопросы этики, таких как юриспруденция, медицина и финансы; а также в критически важных приложениях, где необходимо проверить правильность рассуждений модели. Сверточные нейронные сети достигли высокого качества распознавания в различных задачах. Однако, помимо способности к распознаванию, интерпретируемость предсказаний по-прежнему является серьезной проблемой для нейронных сетей. Отсутствие интерпретируемости не позволяет использовать это семейство моделей "черного ящика" в приложениях, для которых необходимо знать логику прогнозирования, чтобы подтвердить процесс принятия решений. В некоторых областях(например, в медицине) помимо непосредственно предсказания, требуется знать, что является причинами предсказания, какие признаки повлияли на результат прогнозирования. Это может помочь специалистам, которые используют систему ИИ в своей работе, лучше понять механизм прогнозирования нейронной сети, и на основе этих знаний они могут принять свое собственное решение. Целью данной работы является повышение интерпретируемости классификации изображений при сохранении высокой точности.

В работе предлагается новый метод интерпретируемого распознавания изображений. Основная идея заключается в использовании комбинации сверточной нейронной сети и дерева решений по сходству [1]. Модель состоит из трех компонентов: полносверточная сеть, слой визуальных слов или слой образцов, "мягкое" дерево решений. Полносверточная сеть извлекает признаковую карту из изображения. Слой образцов, представленный в [2], содержит их векторные представления. Образцы являются специфичными частями изображения, наличие которых частично или полностью определяет категорию объекта. Слой образцов определяет образцы, имеющие наибольшее сходство с частями изображения. "Мягкое" дерево решений, предложенное в [3], применяется в качестве классификатора, используя в качестве признаков меры сходства с образцами предыдущего слоя.

Разработан алгоритм и соответствующее программное обеспечение. Тестирование на двух публичных наборах данных Animal-10 и RSNA Intracranial Hemorrhage Detection показало конкурентное качество распознавания для предложенного метода по сравнению с классическими сверточными сетями; при этом метод предоставляет возможность интерпретации предсказания.

Работа поддержана грантом РФФИ No. 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. Vol. 1368.

[2] *Chen C., Li O., Barnett A., Su J., Rudin C.* This looks like that: deep learning for interpretable image recognition // NeurIPS, 2019.

[3] *Frosst N., Hinton G.* Distilling a neural network into a soft decision tree // arXiv preprint arXiv:1711.09784, 2017.

Interpretable image recognition using logical decision functions

Berikov Vladimir[1,2] berikov@math.nsc.ru
Kozinets Roman[2]⋆ r.kozinets@g.nsu.ru
[1]Novosibirsk, IM SB RAS
[2]Novosibirsk, NSU

The interpretability of predictive deep learning models is important, especially when it comes to applications in areas where ethics issues are of great importance, such as law, medicine and finance; as well as in critical applications where it is necessary to verify the correctness of the model's reasoning. Convolutional neural networks have achieved high recognition quality in various tasks. However, apart from the ability to recognize, the interpretability of predictions is still a serious problem for neural networks. The lack of interpretability does not allow using this family of "black box" models in applications for which it is necessary to know the logic of the decision to confirm the decision-making process. In some areas (for example, in medicine), in addition to prediction itself, it is required to know what are the causes of prediction, which signs influenced the prediction result. This can help specialists who use the AI system in their work to better understand the mechanism of neural network prediction and based on this knowledge they can make their own decision. This work aims to increase the interpretability of image classification while maintaining high accuracy.

The paper proposes a new method of interpreted image recognition. The main idea is to use a combination of a convolutional neural network and a similarity decision tree [1]. The model consists of three components: a full-convoluted network, a layer of visual words or a layer of patterns, and a "soft" decision tree. The fully-convolution network extracts the feature map from the image. The pattern layer represented in [2] contains vector representations of patterns. Patterns are specific parts of the image, the presence of which partially or completely determines the category of the object. The pattern layer defines the patterns that have the greatest similarity to parts of the image. The "soft" decision tree proposed in [3] is used as a classifier, using similarity measures with the patterns of the previous layer as features.

The algorithm and the corresponding software have been implemented. Testing on two public datasets Animal-10 and RSNA Intracranial Hemorrhage Detection showed a competitive recognition quality for the proposed method compared to classical convolutional networks; at the same time, the method provides the possibility of interpreting the prediction.

This research is funded by RFBR, grant 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. Vol. 1368.

[2] *Chen C., Li O., Barnett A., Su J., Rudin C.* This looks like that: deep learning for interpretable image recognition // NeurIPS, 2019.

[3] *Frosst N., Hinton G.* Distilling a neural network into a soft decision tree // arXiv preprint arXiv:1711.09784, 2017.

Экономная модель трансформера для акустического моделирования речи

Чучупал Владимир Яковлевич[1] v.chuchupal@mail.ru

[1]Москва, Федеральный исследовательский центр «Информатика и управление» Российской академии наук

Разработка вычислительно экономных нейросетевых методов и моделей актуальна. Переход к вычислениям с пониженной разрядностью, использование методов структурной обрезки сети, линейных групповых преобразований, оптимизации вычислений софтмакса позволяет иногда на порядки уменьшить объём вычислительных ресурсов [1]. Существенную экономию также можно получить используя физические ограничения, налагаемые конкретным приложением. В докладе предложен пошаговый алгоритм вычисления внимания в модели трансформера для распознавания речи. Свойства речевого сигнала позволяют организовать вычисления для реализации потоковой обработки (в том числе распознавание в реальном масштабе времени) и при этом имеют оценку вычислительной сложности порядка $O(n)$ (для вычисления само–внимания, n – размер слоя) по сравнению с $O\left(n^2\right)$ в стандартной версии алгоритма.

Пусть $X_0^T = \{x_t \,|t \in [0,T]\}$ – входная последовательность из векторизованных параметров x_t речевого сигнала размерности d, Q_0^T, K_0^T, V_0^T – аналогичные последовательности т.н. запросов q_t, ключей k_t и значений v_t, которые соответственно имеют размерности d_q, d_k, d_v и получаются из X_0^T линейными преобразованиями с матрицами W_q, W_k, и W_v: $W_q \in R^{d_q \times d}$, $W_k \in R^{d_k \times d}$, $W_v \in R^{d_v \times d}$, т. е. $q_t = W_q \cdot x_t, \quad k_t = W_k \cdot x_t, \quad v_t = W_v \cdot x_t$.

Используя свойство фактической независимости параметров речевого сигнала, разнесённых по времени, можно аппроксимировать результат операции внимания на всей фразе конкатенацией частичных результатов на коротких перекрывающихся сегментах, в качестве которых могут использоваться порции данных, последовательно поступающие при потоковом вводе.

Принципиальная (без учёта пропусков и нормирования слоёв, применения маскирующих матриц – эти операции остаются без изменений) предлагаемая пошаговая реализация алгоритма вычисления внимания на одном слое акустической модели (выход записывается в переменную *out*) имеет вид на следующем Рис.1:

Рис.1. Алгоритм пошагового вычисления внимания.

В алгоритме Рис.1 переменные c, l и r обозначают, соответственно, длины сегментов вычисления значений внимания, левого и правого контекстов. Переменная t обозначает текущий индекс. Само–внимание вычисляется для значений из сегмента $[t{+}l : t{+}l{+}c]$. Переменные q, k, v – подпоследовательности запросов, ключей и значений, y – полученное в результате значение само–внимания.

```
t = 0
out = []
while t < T:
    q = Q[t+l:t+l+c]
    v = V[t+l:t+l+c]
    k = K[t:t+l+c+r]
    A = Att(q,k)
    y = Mean(A,v)
    y = ReLU(y)
    t += c
out = concat(out,y)
```

Остальные функции определяются как:

$$\begin{aligned}
\beta &= 1/\sqrt{d} \\
Att(q,k) &= \sigma(\beta \cdot q^{'} \cdot k) \\
\sigma(x_0^T)[j] &= \frac{\exp(x_j)}{\sum_{i=0}^{T}\exp(x_i)} \\
Mean(A,v) &= A \cdot v \\
ReLU(x) &= if \quad x >= 0 \quad then \quad x \quad else \quad 0
\end{aligned}$$

Алгоритм требует порядка $O(n)$ операций и может использоваться как для обучении модели, так и в режиме распознавания. Пошаговая реализация также существенно уменьшает объем используемой памяти: для последовательности длиной n (предполагая $c = l = r$) использует матрицы размером порядка $O\left(l^2\right)$, а стандартный - порядка $O\left(n^2\right)$, обычно $l << n$.

В режиме потокового распознавания оператор *while* должен ожидать поступления новой порции данных, после чего обновить значение t и подпоследовательности q, v и k.

Помимо экономной работы и возможности потоковой обработки, пошаговый алгоритм не приводит к существенным изменениям в матричной модели вычислений и позволяет почти так же эффективно как и стандартный метод использовать графические ускорители.

Численные эксперименты выполнялись на данных речевого корпуса TeCoRus [2], который содержит произнесение случайных наборов цифр от более чем ста дикторов. Алфавит содержит морфы: буквы, пары букв, символы пунктуации, начала и конца фраз, границы слов. Слова на выходе получаются конкатенацией морфов до появления символов конца слова. Параметрическое представление: 24–мерные мел-спектральные признаки, вычисляемые на окне 20 мс., с шагом 10 мс. Эти 24–мерные вектора последовательно объединялись по три и затем прореживались в отношении 2:1. Кодер и декодер состояли из двух слоёв, внешняя языковая модель не использовалась. Декодер на каждом шаге выбирал наиболее вероятный символ. Функция потерь – кросс–энтропия. При выборе

размера пакетов от 200 и выше обучение сходилось за 60–70 эпох. Для алгоритма проверялись несколько вариантов выбора параметров длин сегментов c, l, r. При выборе $c = l = r >= 1$ сек., точность (пословная ошибка 12%) на независимой тестовой выборке практически не снижается по сравнению со стандартным алгоритмом.

Замена стандартного подхода на пошаговый позволила проводить эксперименты с моделью трансформера на обычном ноутбуке.

[1] *Чучупал В. Я.* Акустическое и языковое моделирование в сквозных системах распознавания речи. // Цифровая обработка сигналов, Москва: Российское НТО радиотехники, электроники и связи им. А.С.Попова, 2020. Т. 4. С. 34–43.

[2] *Чучупал В. Я., Маковкин К. А., Чичагов А. В., Кузнецов В. Б., Огарышев В. Ф.* Речевой корпус данных TeKoPyc. // Роспатент, свидетельство о регистрации 200562020, Москва, 2005.

A computationally-effective transformer model for acoustic speech modeling

Chuchupal Vladimir[1] v.chuchupal@mail.ru

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

The development of computationaly effective neural network models is very important. The transition to low precision quantized computations, the use of methods of structural network pruning, group linear transformations, optimization of softmax computations can reduce the amount of computing resources by orders of magnitude [1]. Significant savings can also be obtained by using the physical constraints imposed by a particular application.

The report proposes a step-by-step attention computation algorithm for transformer-based speech recognition. The use of properties of the speech signal allow the encoder to operate in the streaming mode nearly to real-time.. At the same time have the computational complexity of the algorithm is $O(n)$ (n – layer size) compared to $O\left(n^2\right)$ in the vanilla version.

Let $X_0^T = \{x_t \,|t \in [0,T]\}$ be an input sequence of vectorized parameters x_t of a speech signal of dimension d, Q_0^T, K_0^T, V_0^T – similar sequences of the queries q_t, keys k_t and values v_t, which respectively have dimensions d_q, d_k, d_v and are obtained from X_0^T by linear projection with matrices W_q, W_k, and W_v: $W_q \in R^{d_q \times d}$, $W_k \in R^{d_k \times d}$, $W_v \in R^{d_v \times d}$, i. e. $q_t = W_q \cdot x_t, \quad k_t = W_k \cdot x_t, \quad v_t = W_v \cdot x_t$.

Using the independence property of the parameters which are sufficiently spaced in time, it is possible to approximate the result of the attention performed for the entire speech utterance by concatenating the partial results of attention performed on short sequential overlapping segments of speech.

The principal (without taking into account the skipping and normalizing layers, as well as operations with masking matrixes) the proposed step-by-step implementation of calculating self–attention on one layer of the acoustic model (the output is written to the variable *out*) is as follows Fig. 1.:

```
t = 0
out = []
while t < T:
    q = Q[t+l:t+l+c]
    v = V[t+l:t+l+c]
    k = K[t:t+l+c+r]
    A = Att(q,k)
    y = Mean(A,v)
    y = ReLU(y)
    t += c
out = concat(out,y)
```

Fig 1. Step–by–step self–attention computation in speech encoder.

In Fig. 1 t denotes the current index, self-attention is calculated for vectors with indexes in $[t+l : t+l+c]$, where c, l and r denote, respectively, the segment lengths for the values of attention, left and right contexts. q, k, v are subsequences of requests, keys and values, y is the self-attention value for the current segment. The rest of the functions in Fig.1 are defined as:

$$\beta = 1/\sqrt{d}$$
$$Att(q,k) = \sigma(\beta \cdot q^{'} \cdot k)$$
$$\sigma(x_0^T)[j] = \frac{\exp(x_j)}{\sum_{i=0}^{T} \exp(x_i)}$$
$$Mean(A,v) = A \cdot v$$
$$ReLU(x) = if \quad x >= 0 \quad then \quad x \quad else \quad 0$$

The algorithm Fig.1 can be used both for training the model and for the inference. It consumes the order of $O(n)$ operations. Step-by-step implementation can also significantly reduce the amount of required memory: for a sequence of length n, the vanilla algorithm uses matrices (attention values and masking) with a size of order $O\left(n^2\right)$, while a step-by-step implementation requires of order $O\left(l^2\right)$, where $l << n$.

To perform in a streaming mode, the *while* operator should wait for a new piece of data to arrive, and then updates the value of t and subsequences q, v, and k.

Importantly, the step-by-step algorithm does not lead to significant changes in the matrix computation model, that is, it allows using graphics accelerators almost as effectively as the vanilla method.

Numerical experiments were performed on TeCoRus [2] speech corpus, that contain pronunciation of random strings of numbers from more than a hundred speakers. The alphabet is of morphs: letters, pairs of letters, punctuation marks, symbols of beginnig and ending of sentences, word ending morps. The words in the recognized text are obtained by concatenating morphs until the end of word is encountered. As

a parametric representation, 24-dimensional mel-spectral features were used, calculated on a 20 ms window, with a 10 ms step. These 24-dimensional vectors were sequentially stacked by three at a time and then downsampled in a 2:1 ratio. Both the encoder and decoder consisted of two layers, no external language model was used. Decoder simply took the most probable symbol at each step. As a loss function the cross-entropy was used. When choosing batches of size 200 and above, training converged in 60–70 epochs.

Several choices of the segment lengths c, l, r have been tested and upon values $c = l = r >= 1$ sec. the observed accuracy (WER – 12%) is practically not degraded compared to the standard approach. Note that replacing the standard approach with a step-by-step approach made it possible to carry out all experiments with the transformer model and end-to-end recognition system on a regular laptop with a graphics accelerator.

[1] *Chuchupal V.* Acoustic and Language Modeling in End-to-End Speech Recognition Systems. // Digital Signal Processing, Moscow: The Russian Scientific & Technical A.S. Popov Society for Radio Engineering, Electronics & Communications, 2020. Vol. 4. Pp. 34–43.

[2] *Chuchupal V., Makovkin K., Chichagov A., Kouznetsov V., Ogaryshev V.* Speech data corpus TeCoRus. // Rospatent, registration certificate 200562020, Moscow, 2005.

NIGHT-HAZE: набор данных для оценки алгоритмов удаления тумана с изображений, полученных в темное время суток

Филин Андрей Игоревич[1]⋆ adnrewifilin@gmail.com
Копылов Андрей Валериевич[1] And.Kopylov@gmail.com
Середин Олег Сергеевич[1] oseredin@yandex.ru
Грачева Инесса Александровна[1] gia1509@mail.ru
Сурков Егор Эдуардович[1] eg-su@mail.ru
Спицын Данила Александрович[1] danila.spitsyn@bk.ru
Давыдкин Дмитрий Русланович[1] davydkin.dmitrii@bk.ru
Костинский Александр Николаевич[1] sasa-konst@rambler.ru
[1]Тула, Тульский государственный университет

Важность решения задачи устранения тумана для компьютерного зрения и обработки изображений привела к разработке большого количества методов ее решения, однако как сравнительная оценка их эффективности, так и задача обучения интеллектуальных алгоритмов наталкивается на ряд трудностей, связанных с недостатком реальных данных, в которых были бы сформированы пары соответствующих реальных изображений сцен с туманом и без него. Подавляющее большинство изображений в общедоступных наборах данных являются либо художественными фотоснимками пейзажей с туманом, что делает невозможным использование количественных мер качества, основанных на сравнении с эталоном, или синтезированы на основе некоторой модели изображения [1, 2]. Тем не менее, для проверки адекватности таких моделей все равно требуются реальные пары изображений с туманом и без. Нам удалось найти несколько баз изображений такого рода [3, 4, 5]. Но данные базы, во-первых, содержат малое количество изображений (суммарно 130), во-вторых, в них отсутствую изображения, снятые при малой освещенности и наличии локализованных источников света, что является критичным для современных методов устранения тумана. Присутствие локализованных источников света часто нарушает допущения, принятые при оценке атмосферного света, что приводит к переэкспонированию участков и потере деталей на восстановленном изображении. Для решения проблемы построения универсального метода удаления тумана на изображении, способного работать без дополнительной перенастройки как в дневное, так и в ночное время суток, необходимы соответствующие наборы данных. В связи с этим, нами было принято решение о сборе собственного набора данных, соответствующего данным критериям. В данной работе проведен анализ известных наборов данных, способов их сбора, а также составлен и осуществлен собственный план сбора данных. Было подготовлено 2 сцены - с объектами более простой текстуры и формы (гладкие, прямоугольные и круглые объекты), и более сложной (объекты с мелкими деталями, выступающими частями и точечными источниками света). Сцена представляла собой несколько

объектов различной формы, расположенных на столе перед камерой. В кадре также присутствовали различные устройства для возможности последующей калибровки изображений, а именно: мишень SpyderLensCal, таблица для калибровки цветов SpyderCheckr, Datacolor SpyderCube для определения баланса серого, тестовая таблица по ISO 12233). Для каждой сцены было сделано по 16 кадров: с варьированием по 4 степеням освещенности и 4 степеням насыщенности туманом. Степень освещенности изменялась путем изменения количества включенных ламп освещения. При минимальной степени освещенности был включен 1 переносной фонарь. Туман создавался с помощью генератора тумана JINWEIGE FM900-C, производительностью 100 $м^3$/мин. После размещения объектов на сцене и подготовки съемочной аппаратуры, делалось 4 снимка с разной степенью освещенности. Снимки были сделаны на камеру Canon 2000d в двух форматах: raw и jpg в разрешении 6000x4000 и глубиной цвета 24 bit. Дополнительно, для каждой сцены были получены по две карты глубины – с использованием Microsoft Kinnect v2 и Intel RealSence d435i. На полученном наборе проведены эксперименты с использованием современных методов удаления тумана [6, 7, 8, 9, 10]. Эксперименты показали неоднозначные результаты: среднее значение PSNR по всем методам на собранном наборе дает лучший результат, чем в среднем на наборах i-haze и o-haze (19.16 против 15.17); среднее значение SSIM примерно совпадает (0.66 и 0.67 соответственно). Возможно, данное улучшение связано с присутствием в собранном наборе изображений, полученных в условиях низкой освещенности и малой глубины сцены, что требует коррекции используемых метрик и проведения дальнейших исследований.

Работа выполнена при поддержке РФФИ, гранты No. 20-07-00441, 20-07-00055.

[1] *Ancuti C., Ancuti C., DeVleeschouwer C.* D-hazy: A dataset to evaluate quantitatively dehazing algorithms //2016 IEEE international conference on image processing (ICIP). – IEEE, 2016. — С. 2226-2230.

[2] *Li B. et al.* Benchmarking single-image dehazing and beyond //IEEE Transactions on Image Processing. – 2018. — Т. 28. – №. 1. – С. 492-505.

[3] *Ancuti C. et al.* I-HAZE: a dehazing benchmark with real hazy and haze-free indoor images //International Conference on Advanced Concepts for Intelligent Vision Systems. – Springer, Cham, 2018. — С. 620-631.

[4] *Ancuti C. O., Ancuti C., Timofte R.* NH-HAZE: An image dehazing benchmark with non-homogeneous hazy and haze-free images //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. – 2020. — С. 444-445.

[5] *Ancuti C. et al.* O-haze: a dehazing benchmark with real hazy and haze-free outdoor images //Proceedings of the IEEE conference on computer vision and pattern recognition workshops. – 2018. — С. 754-762.

[6] *Berman D. et al.* Non-local image dehazing //Proceedings of the IEEE conference on computer vision and pattern recognition. – 2016. — С. 1674-1682.

[7] *Dhara S. K. et al.* Color cast dependent image dehazing via adaptive airlight refinement and non-linear color balancing //IEEE Transactions on Circuits and Systems for Video Technology. – 2020. – Т. 31. – №. 5. — С. 2076-2081.

[8] *He K., Sun J., Tang X.* Single image haze removal using dark channel prior //IEEE transactions on pattern analysis and machine intelligence. – 2010. – Т. 33. – №. 12. — С. 2341-2353.

[9] *Qin X. et al.* FFA-Net: Feature fusion attention network for single image dehazing //Proceedings of the AAAI Conference on Artificial Intelligence. – 2020. – Т. 34. – №. 07. — С. 11908-11915.

[10] *Zhu Q., Mai J., Shao L.* A fast single image haze removal algorithm using color attenuation prior //IEEE transactions on image processing. – 2015. – Т. 24. – №. 11. — С. 3522-3533.

NIGHT-HAZE: A dehazing benchmark with real hazy and haze-free low-light indoor images

Filin Andrei[1]⋆ adnrewifilin@gmail.com
Kopylov Andrei[1] And.Kopylov@gmail.com
Seredin Oleg[1] oseredin@yandex.ru
Gracheva Inessa[1] gia1509@mail.ru
Surkov Egor[1] eg-su@mail.ru
Spitsyn Danila[1] danila.spitsyn@bk.ru
Davydkin Dmitrii[1] davydkin.dmitrii@bk.ru
Kostinskii Aleksandr[1] sasa-konst@rambler.ru
[1]Tula, Tula State University

Foggy scenes are common in image processing problems. The presence of fog restricts visibility and degrades image quality. The importance of haze removal for computer vision and image processing has led to the development of a fairly large number of dehazing methods. However, comparative effectiveness assessment of dehazing methods as well as training of intelligent algorithms encounter difficulties associated with a lack of real data, consisting of pairs of real hazy images and haze-free (ground truth) images. The vast majority of images from publicly available datasets contain either artwork photos of foggy landscapes, that's why it is impossible to use quantitative quality measures based on comparison with a ground truth image, or images synthesized based on some model [1, 2]. However, it should be noted that there are still required pairs of the real hazy and ground truth images for models verification. We managed to find several databases of images of this kind [3, 4, 5]. But firstly, these databases contain a small number of images (130 images in total), and, secondly, they do not contain images taken at low-light conditions with the presence of point-light sources, which is critical for modern dehazing methods. The presence of localized light sources often violates the assessing atmospheric light assumptions, resulting in overexposure of image regions and loss of details in the reconstructed image. To solve the problem of constructing a universal haze removal method, that can work both day and nighttime without additional fine-tuning, appropriate datasets are required. In this regard, we decided to collect our dataset that meets these criteria. In this paper, we have analyzed the known datasets, their gathering methods, and also compiled and implemented our own data acquisition plan. Two scenes were prepared - with the objects of simpler texture and shape (smooth, rectangular, and round objects), and more complex (the objects with small details, protruding parts, and point light sources). The scenes consisted of several objects of various shapes placed on the table in front of the camera. The frame also contained various devices for the possibility of image post-calibration: SpyderLensCal Focus Tool, SpyderCheckr color calibration table, SpyderCube for gray balance determination, ISO 12233 test chart) 16 frames were taken for each scene: with varying 4 degrees of illumination and 4 degrees of fog saturation. The degree of illumination

was changed by changing the number of lighting lamps on. At the minimum illumination, 1 portable lamp was switched on. The haze was produced by a JINWEIGE FM900-C haze machine with a capacity of 100 m^3 / min. After placing objects on the scene and preparing the shooting equipment, 4 pictures were taken with different illumination. Then the haze generator was turned on, and every 5 minutes 4 similar pictures were taken. The resulting dataset consists of 32 photos of scenes in various lighting conditions, visibility (haze intensity), with the presence/absence of point light sources, with varying complexity of objects shapes of the scene. The pictures were taken with a Canon 2000d camera in two formats: raw and jpg with 6000x4000 resolution and 24-bit color depth. In addition to photographs, two depth maps were obtained for each scene - using Microsoft Kinnect v2 and Intel RealSence d435i. The resulting dataset were used to perform experiments on several state-of-the-art haze removal methods [6, 7, 8, 9, 10]. The experiments showed ambiguous results: the average PSNR for all methods on the collected set is better than the average over the i-haze and o-haze sets (19.16 versus 15.17); the average SSIM is near matched (0.66 and 0.67, respectively). Perhaps this improvement is due to the presence in the collected dataset images, captured in low light conditions and shallow scene depth, which requires correction of the used metrics and further research.

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

[1] *Ancuti C., Ancuti C., DeVleeschouwer C.* D-hazy: A dataset to evaluate quantitatively dehazing algorithms // 2016 IEEE international conference on image processing (ICIP), 2016. Pp. 2226–2230.

[2] *Li B. et al.* Benchmarking single-image dehazing and beyond // IEEE Transactions on Image Processing, 2018. Vol. 28(1). Pp. 492–505.

[3] *Ancuti C. et al.* I-HAZE: a dehazing benchmark with real hazy and haze-free indoor images // International Conference on Advanced Concepts for Intelligent Vision Systems, 2018. Pp. 620–631.

[4] *Ancuti C. O., Ancuti C., Timofte R.* NH-HAZE: An image dehazing benchmark with non-homogeneous hazy and haze-free images // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020. Pp. 444–445.

[5] *Ancuti C. et al.* O-haze: a dehazing benchmark with real hazy and haze-free outdoor images // Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2018. Pp. 754–762.

[6] *Berman D. et al.* Non-local image dehazing // Proceedings of the IEEE conference on computer vision and pattern recognition, 2016. Pp. 1674–1682.

[7] *Dhara S. K. et al.* Color cast dependent image dehazing via adaptive airlight refinement and non-linear color balancing // IEEE Transactions on Circuits and Systems for Video Technology, 2020. Vol. 31(5). Pp. 2076–2081.

[8] *He K., Sun J., Tang X.* Single image haze removal using dark channel prior // IEEE transactions on pattern analysis and machine intelligence, 2010. Vol. 33(12). Pp. 2341–2353.

[9] *Qin X. et al.* FFA-Net: Feature fusion attention network for single image dehazing // Proceedings of the AAAI Conference on Artificial Intelligence, 2020. Vol. 34(7). Pp. 11908–11915.

[10] *Zhu Q., Mai J., Shao L.* A fast single image haze removal algorithm using color attenuation prior // IEEE transactions on image processing, 2015. Vol. 24(11). Pp. 3522–3533.

Сравнительный анализ алгоритмов в задаче сегментации срезов красочного слоя картин

Князев Денис Викторович[1]⋆ denis.denis-knyazev2018@yandex.ru
Мурашов Дмитрий Михайлович[1] d_murashov@mail.ru
[1]Москва, ФИЦ ИУ РАН

Работа посвящена сегментации изображений сложной структуры. Под изображениями сложной структуры понимаются изображения, на которых зафиксированы объекты с нечеткими границами и значительно отличающиеся по размерам, степени однородности, возможно присутствие текстурных областей.

Исследование проводилось на изображениях срезов красочного слоя картин, зафиксированных камерой, установленной на микроскоп. На изображениях необходимо выделить слои краски и частицы пигментов. В известных работах для сегментации микросрезов используются методы кластеризации и метод на основе авторегрессионной модели и модели гауссовой смеси, применяемых на нескольких уровнях разрешения исследуемого изображения (Kaspar,R., et al., 2005; Zitova, B. et al., 2010; M. Haindl et al., 2009). Эти методы демонстрируют невысокий процент правильно выделенных объектов, и авторы указывают, что для применения методов сегментации и коррекции результатов необходим опытный оператор. Поэтому для сегментации изображений сложной структуры, какими являются изображения микросрезов красочных слоев картин, возникает необходимость разработки новых алгоритмов. Обеспечить приемлемое качество сегментации представляется возможным за счет комбинирования нескольких алгоритмов.

В представляемой работе анализировались результаты трех наиболее эффективных, судя по литературным источникам, алгоритмов сегментации изображений. Производилась визуальная и количественная оценка качества сегментации. Для количественной оценки качества применялась вариация информации, так как она является одной из стандартных метрик, используемых в работах по сегментации. С помощью указанной мер оценивалась величина информационного различия между оригинальными и сегментированными изображениями. Первый алгоритм, описанный в статье [1] (Arbelaez и др.), основан на преобразовании результата работы контурного детектора (алгоритм gPb, globalized probability of boundary), в иерархическое дерево областей. Второй алгоритм – graph cutting (Felzenszwalb и др.)[2]. Он основан на представлении изображения в виде графовой модели. Ребра в таком графе соединяют соседние пиксели изображения. Затем при попарном сравнении областей выполняется выделение или слияние областей на основе весовых значений на ребрах. Третьим алгоритмом является алгоритм сегментации, основанный на применении Марковских случайных полей [3] (Kato и др.). Модель марковского поля строится на основе фрагментов изображения, для которых вычисляются параметры распределе-

ния (математическое ожидание и дисперсия). Затем составляется вероятностная карта изображения, для которой минимизируется энергетическая функция.

В эксперименте входными данными являются цветные изображения микросрезов красочного слоя картин глубиной 24 бита в формате JPEG в разрешении 150 пикселей на дюйм, зафиксированные цифровой CCD камерой AxioCam ICc 3,3 Мп, установленной на микроскопе Carl Zeiss Axio Imager 1. Для сегментации на каждом изображении была выделена прямоугольная область, содержащая образец красочного слоя. В таблице 1 представлены значения меры качества, полученной в результате применения трех описанных выше алгоритмов. Первый алгоритм (Arbelaez и др.,[1]) хорошо выделил границы слоев, но плохо выделил частицы пигментов. Второй алгоритм(graph cutting) значительно лучше выделяет частицы пигментов, однако границы слоев получились размытыми, что говорит о низкой эффективности данного метода для выделения границ объектов. Алгоритм сегментации на основе марковских полей оказался наиболее чувствительным при распознавании частиц пигментов, однако также были выделены ложные сегменты. Границы слоев этим алгоритмом выделяются хуже, чем алгоритмом контурного детектора (Arbelaez и др. [1]). Из таблицы 1 следует, что все алгоритмы при сегментации исследуемых изображений показали схожие значения вариации информации: Arbelaez (среднее значение 0.6175 и СКО 0.1349), graph cutting (Felzenszwalb и др. среднее значение 0.6085; СКО 0.1339) и Markov Random Fields (среднее значение 0.6372; СКО 0.143), однако первые два алгоритма обеспечили в среднем большее информационное сходство между сегментированными и оригинальными изображениями. Визуальная оценка показала, что третий алгоритм выделил много ложных сегментов, что оказалось причиной увеличения значений вариации информации.

Таким образом, проведенный анализ показал, что для сегментации изображений сложной структуры представляется целесообразным создание алгоритма, который бы комбинировал несколько алгоритмов, эффективно выделяющих отдельные компоненты структуры. Для построения комбинированного алгоритма возможно применять алгоритм контурного детектора (Arbelaez и др.) и алгоритм на основе графовой модели (Felzenszwalb и др.).

Таблица 1–Значения меры качества сегментации изображений

Идентификатор изображения (размер)	Arbelaez и др.	graph cutting (Felzenszwalb и др)	Markov Random Fields
Image1(280×223)	0.69380	0.66264	0.75076
Image2(300×179)	0.45974	0.54968	0.73122
Image3(300×147)	0.56790	0.55871	0.40364
Image4(300×177)	0.69989	0.54708	0.64118
Image5(300×328)	0.65579	0.58446	0.66151
Image6(150×223)	0.28594	0.25463	0.30839
Image7(543×449)	0.67961	0.72531	0.72881
Image8(437×921)	0.72462	0.70273	0.74450
Image9(1461×887)	0.58521	0.59735	0.58399
Image10(2080×1540)	0.82651	0.82092	0.85813
Image11(893×801)	0.65389	0.67918	0.66495
Image12(553×463)	0.54684	0.56954	0.54744
Image13(2080×1540)	0.64813	0.65842	0.65977

[1] *Arbelaez P. Maire M. Fowlkes C. Malik J.* Contour Detection and Hierarchical Image Segmentation // IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010. Vol. 33(5). Pp. 898–916.

[2] *Felzenszwalb P. F. Huttenlocher D. P.* Efficient Graph-Based Image Segmentation // International Journal of Computer Vision, 2005. Vol. 59(2). Pp. 167–181.

[3] *Kato Z. Zerubia J.* Markov Random Fields in Image Segmentation // Foundations and Trends in Signal Processing, 2011. Vol. 5(1).

Experimental research of algorithms for segmenting paint layer cross-sections of paintings

Knyazev Denis[1]⋆ denis.denis-knyazev2018@yandex.ru
Murashov Dmitry[1] d_murashov@mail.ru
[1]Moscow, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences

The work is devoted to the segmentation of images of a complex structure. Images of a complex structure are defined as images with fuzzy regions significantly different in size and uniformity. Also, the presence of textural areas is possible.

The study was carried out on images of paint layer cross-sections of paintings recorded by a camera mounted on a microscope. It is necessary to segment paint layers and pigment particles in the images. In well-known works, clustering methods and a method based on an autoregressive model and a Gaussian mixture model at several resolution levels of the image under the study are used (see Kaspar, R., et al., 2005; Zitova, B. et al., 2010; M. Haindl et al., 2009). These methods demonstrate a low percentage of correctly selected objects, and the authors point out that an experienced operator is needed to apply segmentation methods and correct the results. Therefore, it is necessary to develop new algorithms for segmenting images of a complex structure, such as microscopic images of paint layer cross-sections of paintings. It is possible to obtain acceptable segmentation quality by combining several algorithms.

In this work, the results of the three most effective, according to the literature, image segmentation algorithms were analyzed. Visual and quantitative assessment of segmentation quality was performed. To estimate the quality, variation of information was used, since it is one of the standard metrics used in segmentation tasks. Using this measure, we estimated the magnitude of the information difference between the original and segmented images. The first algorithm described in the article [1] (Arbelaez et al.) is based on the transformation of the result of the contour detector (the gPb algorithm, globalized probability of boundary) into a hierarchical tree of regions. The second algorithm is graph cutting (Felzenszwalb et al.)[2]. It is based on the representation of an image in the form of a graph model. Edges in such a graph connect adjacent pixels of the image. Then, when comparing the regions in pairs, the regions are selected or merged based on the weight values on the edges. The third algorithm is a segmentation algorithm based on the Markov random fields model[3] (Kato et al.). The Markov field model is constructed based on image fragments for which distribution parameters (mathematical expectation and variance) are calculated. Then a probabilistic image map is compiled, for which the energy function is minimized.

In the experiment, the input data are color images of microscopic paint layer cross-sections of paintings with a depth of 24 bits in JPEG format at a resolution of 150 pixels per inch, recorded by the AxioCam ICC 3.3 Mp digital CCD camera

mounted on a Carl Zeiss Axio Imager 1 microscope. For segmentation, a rectangular area containing a sample of the cross-section was selected in each image. Table 1 shows the values of the quality measure obtained as a result of applying the three algorithms described above. The first algorithm (Arbelaez et al.,[1]) detected the boundaries of the layers well, but poorly segmented the pigment particles. The second algorithm (graph cutting) detects pigment particles much better, but the boundaries of the layers turned out to be blurred, which indicates the low efficiency of this method for detecting the boundaries of objects. The segmentation algorithm based on Markov random fields turned out to be the most sensitive when recognizing pigment particles, but false segments were also identified. The boundaries of paint layers are distinguished worse by this algorithm than by the contour detector (Arbelaez et al. [1]). It follows from Table 1 that all segmentation algorithms applied to the studied images showed similar values of variation of information: Arbelaez (mean value 0.6175 and RMS 0.1349), graph cutting (Felzenszwalb et al. mean value 0.6085; RMS 0.1339), and Markov Random Fields (mean value 0.6372; RMS 0.143), however, the first two algorithms provided on average greater information similarity between segmented and original images. The visual evaluation showed that the third algorithm identified many false segments, which turned out to be the reason for the increase in the values of variation of information.

Thus, the analysis showed that for segmentation of images of a complex structure, it seems appropriate to create an algorithm that would combine several algorithms that effectively detect specific components of the structure. To build a combined algorithm, it is possible to use the contour detector algorithm (Arbelaez et al.) and the graph cutting algorithm (Felzenszwalb et al.).

Table 1– Vaslues of segmentation quality measure

Image identifier (size)	Arbelaez et al.	Graph cutting (Felzenszwalb et al.)	MRF
Image1(280×223)	0.69380	0.66264	0.75076
Image2(300×179)	0.45974	0.54968	0.73122
Image3(300×147)	0.56790	0.55871	0.40364
Image4(300×177)	0.69989	0.54708	0.64118
Image5(300×328)	0.65579	0.58446	0.66151
Image6(150×223)	0.28594	0.25463	0.30839
Image7(543×449)	0.67961	0.72531	0.72881
Image8(437×921)	0.72462	0.70273	0.74450
Image9(1461×887)	0.58521	0.59735	0.58399
Image10(2080×1540)	0.82651	0.82092	0.85813
Image11(893×801)	0.65389	0.67918	0.66495
Image12(553 ×463)	0.54684	0.56954	0.54744
Image13(2080×1540)	0.64813	0.65842	0.65977

[1] *Arbelaez P. Maire M. Fowlkes C. Malik J.* Contour Detection and Hierarchical Image Segmentation // IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010. Vol. 33(5). Pp. 898–916.

[2] *Felzenszwalb P. F. Huttenlocher D. P.* Efficient Graph-Based Image Segmentation // International Journal of Computer Vision, 2005. Vol. 59(2). Pp. 167–181.

[3] *Kato Z. Zerubia J.* Markov Random Fields in Image Segmentation // Foundations and Trends in Signal Processing, 2011. Vol. 5(1).

Использование нейронных сетей для выявления аномалий на снимках полученных на сканерах персонального досмотра

Карандашев Яков Михайлович[1,2] karandashev@niisi.ras.ru
Марков Александр Сергеевич[1]⋆ to.asmarkov@gmail.com
Котляров Евгений Юрьевич[1] tyztot@gmail.com
[1]Москва, Российский университет дружбы народов
[2]Москва, НИИСИ РАН

В данной работе рассматривается решение задачи выявления аномалий на рентгеновских снимках полученных сканерами персонального досмотра. На объектах, требующих повышенного контроля безопасности, часто используются сканеры персонального досмотра (СПД). Они позволяют быстро сделать снимок человека в рентгеновском диапазоне излучения, на котором оператор СПД может увидеть все объекты на теле человека и визуально подтвердить или опровергнуть наличие запрещённых среди них. Процесс обладает рядом существенных недостатков, связанных с человеческим фактором: для качественного анализа снимка требуется существенное время и повышенное внимание, что приводит к быстрой утомляемости оператора СПД и может негативно сказаться на качестве анализа снимков. В данный процесс можно внести существенную долю автоматизции, сделав его более дешёвым для организации и более комфортным для человека. Для решения поставленной задачи использовались глубокие нейронные сети. В этой работе представлены способы предобработки данных и применение сети U-2-Net, а также анализ результатов.

Требуется разработать решение, ставящее в соответствие каждому рентгеновскому снимку булевую маску, истинные значения которой соответствуют пикселям, на которых присутствуют инородные объекты, такие как телефоны, оружие, металлические предметы и прочее. Эти объекты в дальнейшем будем называть аномалиями.

Обучение проводилось на наборе данных, состоящем из оригинальных 16-битных снимков с аппаратов СПД в формате tiff. У этого набора было два существенных недостатка: не совпадающие распределения интенсивности пикселей среди снимков, сделанных на различных СПД, а также невозможность визуально наблюдать все аномалии на снимке, не прибегая к его модификации. Для исправления этих недостатков, все снимки были подвергнуты последовательности преобразований.

Сначала ко всем снимкам был применён фильтр резкости с ядром [[0,-1, 0], [-1, 5,-1], [0,-1,0]]. Затем были проведены трансформации гистограммы снимков, а также корректировки наиболее интенсивных пикселей относительно значений средней интенсивности.

После обработки набора данных, аномалии стали визуально различимыми, что позволило использовать сторонние сервисы (Yandex Toloka) для создания карт аномалий.

Рис. 1. Внешний вид изображений после преобразований(вверху) и гистограмма распределения цвета(внизу)

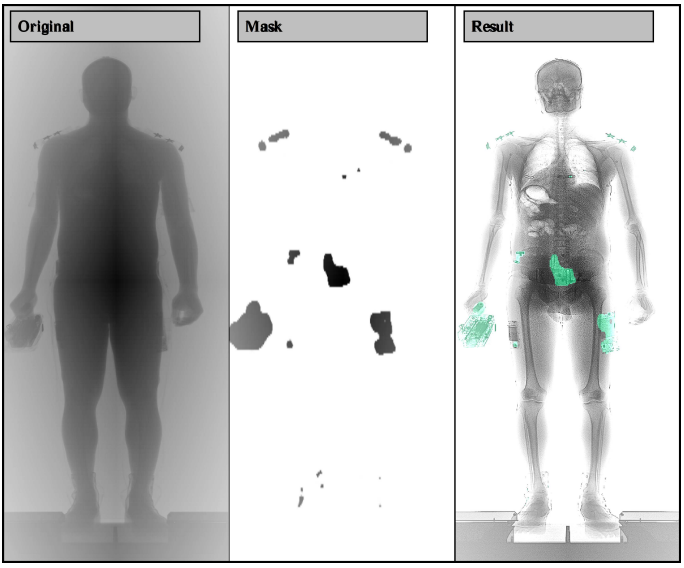


Рис. 2. Оригинальное изображение(слева), полученные маски(по центру), наложение маски на обработанное изображение(справа)

Для решения задачи выявления аномалий использовалась сеть U-2-Net[1], которая была представлена в 2020 году. Эта модель построена на базе архитектуры U-Net[2]. На вход модели подавались одноканальные снимки с указаной выше предобработкой. В процессе аугментации для увеличения вариативности данных мы использовали операцию случайной обрезки изображения(cropping), что позволило сгенерировать большое количество изображений содержащих различные части тела. Модель отображает снимок в матрицу, характеризующую вероятности нахождения аномалий в соответствующих участках снимка. Далее матрица вероятностей приводится к бинарному виду. Таким образом, на выходе имеем булеву маску. В качестве функции потерь была использована IoU (intersection over union)[3].

Качество сегментации обученной модели позволяет распознавать аномалии большого и среднего размера. На объектах сравнительно меньшего размера качество сегментации существенно снижается. Несмотря на это, модель может быть использована на промышленных объектах, в качестве средства автоматизации работы СПД для поиска объектов среднего размера, таких как оружие, телефоны, слитки металлов и прочее, значительно увеличивая скорость работы оператора.

[1] *Xuebin Q., Zichen Z., Chenyang H., Masood D., Zaiane M., Osmar R., Jagersand M.* U2-Net: Going deeper with nested U-structure for salient object detection // Pattern Recognition, 2020.

[2] *Ronneberger O., Fischer P., Brox T.* U-Net: Convolutional Networks for Biomedical Image Segmentation // Pattern Recognition, 2021.

[3] *Zheng Z., Wang P., Liu W., Jinze L., Rongguang Y., Dongwei N.* Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression // Pattern Recognition, 2019.

Anomaly Detection on full body scanner images with neural networks

Karandashev Iakov[1,2] karandashev@niisi.ras.ru
Markov Alexander[1]⋆ to.asmarkov@gmail.com
Kotlyarov Evgeniy[1] tyztot@gmail.com
[1]Moscow, Peoples Friendship University of Russia (RUDN University)
[2]Moscow, Scientific Research Institute for System Analysis of RAS

In this paper we consider the problem of detecting anomalies in X-ray images made by full body scanners. Full body scanners (FBS) are often used at places that require increased security control. They make it possible to quickly take a picture of a visitor in the X-ray spectrum, in which the FBS operator can see all the objects on a person's body and visually confirm or deny the presence of prohibited items among them. This process has a number of significant disadvantages related to the human factor: a qualitative research of the image requires considerable time and focused attention. This leads to rapid tiredness of the FBS operator and can have a negative impact on the quality of the image research. This process can be partly automated, making it cheaper for the organization and more comfortable for the visitors. Deep neural networks have been used to solve this problem. This paper presents ways to preprocess data and the use of U-2-Net, as well as an analysis of the results.

We need to develop a solution that matches each X-ray image with a boolean mask, the true values of which correspond to the pixels with foreign objects, such as phones, weapons, metal objects, and so on. These objects will be further referred to as anomalies.

The model was trained on a dataset consisting of the original 16-bit FBS images in tiff format. This dataset had two significant drawbacks: inconsistent pixel intensity distributions among the images taken at different FBSs, and the impossibility to visually observe all the anomalies in the image without modifying it. To correct these drawbacks, all the images were transformed with a sequence of operations.

First, a sharpening filter was applied to all the images with the kernel [[0,-1, 0], [-1, 5,-1], [0,-1,0]]. Then histogram equalization operations were added and alignment of the most intense pixels of the image were adjusted with respect to the average.

After processing the dataset, the anomalies became clearly visible, which allowed us to use third-party services (Yandex Toloka) to create anomaly maps, so we move from unsupervised task to supervised. The U-2-Net[1] network, which was introduced in 2020, was used to solve our problem. This model is based on the U-Net[2] architecture. The input to the model was single-channel images with the above-mentioned pre-processing. During the augmentation, to increase the variability of the data, we used the operation of random cropping, which allowed us to generate a large number of images containing different body parts. The model maps the image into a matrix that characterizes the probabilities of anomaly existing in the corre-

Fig. 1. Image after transformations

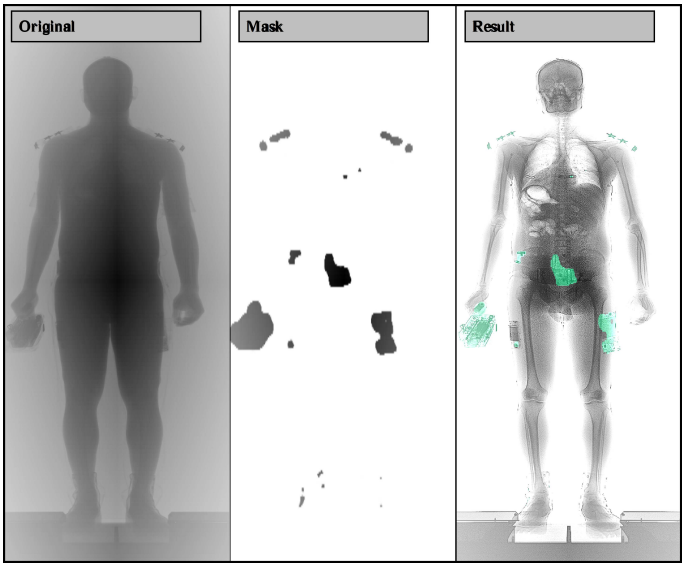


Fig. 2. Original image (left), obtained masks (center), masking over the processed image (right)

sponding image parts. Then the matrix of probabilities is converted to the binary, so the output is a boolean mask. IoU (intersection over union)[3] was used as the loss function.

As a result of this work, a model was trained that finds foreign objects on X-ray images. The segmentation quality of the model allows us to detect the large and medium-sized anomalies. On relatively smaller objects, such as needles, the quality of segmentation is significantly reduced. Despite this, the model can be used at industrial purposes, as an FBS automation tool to search for medium-sized objects, such as weapons, telephones, metal ingots etc, significantly increasing the speed of the operator work.

[1] *Xuebin Q., Zichen Z., Chenyang H., Masood D., Zaiane M., Osmar R., Jagersand M.* U2-Net: Going deeper with nested U-structure for salient object detection // Pattern Recognition, 2020.

[2] *Ronneberger O., Fischer P., Brox T.* U-Net: Convolutional Networks for Biomedical Image Segmentation // Pattern Recognition, 2021.

[3] *Zheng Z., Wang P., Liu W., Jinze L., Rongguang Y., Dongwei N.* Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression // Pattern Recognition, 2019.

Детектирование подделок в мобильных системах распознавания по лицу при помощи стереокамеры

Ефимов Юрий Сергеевич[1]⋆ yuri.efimov@phystech.edu
Матвеев Иван Алексеевич[2] matveev@ccas.ru
[1]Долгопрудный, Московский физико-технический институт (национальный исследовательский университет)
[2]Москва, ФИЦ ИУ РАН

Применимость биометрических систем идентификации на практике во многом определяется их возможностями по обнаружению предъявленных им подделок или спуфинг-атак. Изображение лица - одна из наиболее распространённых и в то же время наиболее уязвимых для атак модальностей, применяемых в современных мобильных устройствах. Некоторые из смартфонов, представленных на рынке, оборудованы парой фронтальных камер, формирующей стереопару с малым расстоянием между сенсорами или стереобазисом. Использование стереоизображений позволяет повысить предоставляемый системой уровень безопасности против подделывания биометрического признака.

Предложен метод [1] обнаружения спуфинг-атак в системах распознавания по лицу на мобильных устройствах. Метод способен работать в реальном времени на устройствах с ограниченными вычислительными возможностями, различными условиями регистрации и с использованием штатной стереокамеры, особенностью которой является малый стереобазис. Метод основан на сверточной нейронной сети и многоцелевом обучении, предложена специальная функция потерь.

Рассмотрены следующие типы подделок: высококачественные распечатанные изображения лиц, цифровые фотографии и видеопоследовательности, показываемые на экранах высокого разрешения. Предлагаемый подход достигает высоких показателей точности детектирования подделок, сравнимых с описанными в современной литературе аналогичными подходами, в том числе на данных открытой базы стереоизображений. От известных аналогов предлагаемый метод отличается малым временем выполнения на современных мобильных процессорах, поэтому может быть применен для детектирования подделок в биометрических системах с малыми вычислительными ресурсами.

Работа поддержана грантами РФФИ No. 19-31-90171, 19-07-01231.

[1] *Ефимов Ю. С., Матвеев И. А.* Детектирование подделок в мобильных системах распознавания по лицу при помощи стереокамеры // Известия РАН. Теория и системы управления, 2021.

Stereo Face Liveness Detection for Mobile Biometric Applications

Efimov Iurii[1]⋆ yuri.efimov@phystech.edu
Matveev Ivan[2] matveev@ccas.ru
[1]Dolgoprudny, MIPT
[2]Moscow, FRC CSC RAS

Practical applications of biometric systems require high capabilities of spoofing attack detection. Face image is one of the most commonly used and at the same time most vulnerable biometric traits for counterfeiting. Some modern smartphones with face unlock capability are also equipped with a pair of frontal camera, which form a stereo pair with a small stereo baseline. Stereo information provides additional depth cues necesary to improve system liveness detection accuracy.

A method [1] of face liveness detection is proposed in this work. It is based on applying a convolutional neural network, trained in multi-task learning paradigm using a special loss function. The proposed approach is capable of real-time inference on mobile devices with highly limited computational resources and equipped with a small baseline stereo camera. The method also shows high robustness under various environmental conditions, which is specific and significant for smartphone interaction.

The following typoes of face spoof samples are considered: high quality face printouts, digital photos and video, displayed on high resolution screens. Testing of the method is performed on different datasets, both open and manually collected, and reveals its high classification accuracy.

This research is funded by RFBR, grants 19-31-90171, 19-07-01231.

[1] *Efimov I., Matveev I.* Stereo Face Liveness Detection for Mobile Biometric Applications // Pattern Recognition and Image Analysis, 2021.

Детекция устаревших регионов в карте для лидарной локализации при помощи синтетичских модификаций лидарных облаков

Иванов Дмитрий Александрович[1,2]⋆ ivanov.dmitriy.99@gmail.com
Ольховников Сергей Юрьевич[2] olkhovnikov@yandex-team.ru
[1]Санкт-Петербург, НИУ ИТМО
[2]Москва, Яндекс Беспилотные технологии

Задача локализации заключается в определении положения и позы робота относительно какой-то карты по данным сенсоров. Одним из наиболее популярных сенсоров для локализации является лидар. При использовании лидарной локализации необходимо сопоставить облако точек, полученное во время проезда с картой для определения позиции, из которой это облако точек было получено.

Существуют алгоритмы, которые одновременно выполняют создание карты и локализацию по этой карте. Но в отрасли беспилотных автомобилей локализуются по заранее подготовленной HD-карте, дающей дополнительные гарантии безопасности [1].

Но использование карты несёт в себе другие проблемы. То насколько карта соответствует действительности, напрямую влияет на точность решения задачи локализации. Поэтому при изменении местности необходимо обновлять карту, а точнее её участки [4]. Некоторые несоответствия карты и действительности снижают точность локализации только в условиях, которые ухудшают лидарное облако [2]. Например мокрый асфальт, который отражает лучи лидара, из-за чего исчезают точки дороги. В [3] показано, что для применения беспилотных автомобилей уровень отказов должен быть не более 10^{-9} ошибок на милю. Там же замечено, что сбор достаточного количества данных для подтверждения этой статистики с реальных проездов может занять десятки лет. В этой работе предлагается симулировать условия, в которых мог бы оказаться беспилотный автомобиль. Это позволит оценить качество локализации по одному проезду в хороших условиях для проездов по этому же маршруту во всех условиях, когда беспилотный автомобиль может проехать. А по качеству локализации можно оценить свежесть карты.

Так как определить точность локализации в реальных условиях сложно из-за трудностей с получением истины, будем оценивать качество локализации по другому. Промоделируем процесс съёмки облака точек в данном месте через случайный генеративный процесс с эвристическим распределением p_{cloud}. Большинство методов локализации использует начальное приближение для своей работы, поэтому также введём случайный процесс, выдающий начальное приближение с нормальным распределением p_{init}. Общее распределение входных данных алгоритма локализации описывается распределением p_{input}, которое является декартовым произведением двух других распределений.

Тогда можно представить локализацию по заданной карте map, как функцию:

$$pose = localization_{map}(cloud, init\_estimate) = localization_{map}(input)$$

Считаем, что при наличии хорошей карты локализация устойчива к возможным деградациям лидарных облаков и начальным приближениям. В нашем методе комбинация облаков и приближений представлена процессом p_{input}. Устойчивость выражается в малом значении второго момента позы. Аналитически вычислить эту величину не представляется возможным, но можно её оценить при помощи метода Монте-Карло, сэмплируя из случайного процесса p_{input}:

$$\mathbb{E}_{x\sim p_{input}}[pose(x)] = \int pose(x)p(x)dx \approx \frac{1}{m}\sum_{i}^{m} pose(x_i)$$

$$Cov_{x\sim p_{input}}[pose(x)] = \int \left(pose\,(x) - \mathbb{E}_{x\sim p_{input}}\left[pose\,(x)\right]\right)^2 p(x)dx$$

$$\approx \frac{1}{m}\sum_{i}^{m}\left(pose\,(x_i) - \frac{1}{m}\sum_{i}^{m} pose(x_i)\right)^2,$$

где m - число сэмплов. Если полученное значение ниже порога, подобранного на датасете с примерами успешной и неуспешной локализации, то качество локализации считается хорошим. Иначе считаем, что в окрестности этой точки карта не соответствует действительности, и надо обновить этот участок карты.

При анализе датасета проездов были выявлены следующие наиболее частые причины деградации лидарных облаков: другие автомобили, ограничивающие обзор, мокрый асфальт, осадки и туман. Для моделирования этих условий в генеративном процессе p_{cloud} были применены следующие модификации: удаление секторов точек с разной протяженностью, удаление точек на поверхности земли, добавление случайных точек и удаление точек на определенном расстоянии.

Для тестирования модификаций был создан датасет из двух частей: первая включает в себя проезды, в которых заметны отличия карты и действительности. Вторая часть включает в себя проезды в участках на очень свежей карте. Модификации облаков оценивались по двум метрикам. `True Positive Rate`, которая показывает насколько хорошо использованные модификации помогают определить участки с устаревшей картой. И `True Negative Rate`, которая показывает, что использование модификации не приведет к ложным показаниям.

По результатам проведенных экспериментов лучше всего показали себя удаление секторов точек, с четырех сторон от автомобиля и удаление точек на

поверхности земли. Выбранная комбинация имела наибольший TPR, при значении TNR равном 1. Другие фильтры имели, либо слишком низкий TPR, либо слишком низкий TNR.

[1] *Seiko H., Hu X.* Autonomous Driving in the iCity—HD Maps as a Key Challenge of the Automotive Industry // Engineering, 2016. Vol. 2(2). Pp. 159–162.

[2] *Endo Y., Javanmardi E., Kamijo S.* Analysis of Occlusion Effects for Map-Based Self-Localization in Urban Areas // Sensors, 2021. Pp. 51–96.

[3] *Reid T., Houts, S. et al* Localization Requirements for Autonomous Vehicles // SAE International Journal of Connected and Automated Vehicles, 2019. Vol. 2(3).

[4] *Jomrich F.* Dynamic Maps for Highly Automated Driving - Generation, Distribution and Provision // Technische Universität Darmstadt, 2020.

Detection of outdated regions in map for lidar localization using synthetic lidar clouds modifications

Ivanov Dmitriy[1,2]⋆ ivanov.dmitriy.99@gmail.com
Olkhovnikov Sergey[2] olkhovnikov@yandex-team.ru
[1]Saint-Petersburg, ITMO University
[2]Moscow, Yandex Selfdriving Group

Localization is the task of determining the robot's position and orientation with respect to some map. One of the most popular sensors for localization is LIDAR. To localize using LIDAR data robot needs to match the point cloud obtained during a ride with the map to determine the position from which the point cloud was obtained.

There are algorithms that simultaneously perform mapping and localization. But in the industry of self-driving cars localization with pre-prepared HD-maps, which gives additional safety guarantees, is preferred [1].

But usage of a prepared map has its problems. The extent to which the map corresponds to reality directly affects the accuracy of localization. Therefore, if reality changes, it is necessary to update the map, or rather its parts[4]. Some inconsistencies of the map and reality reduce the localization accuracy only in conditions that worsen the lidar cloud [2]. For example, wet asphalt reflects the LIDAR beams, so the road points disappear.In [3] it is shown that in order to use self-driving cars, the failure rate should not be more than 10^{-9} errors per mile. Also noted is that collecting enough data to validate these statistics from real-world runs can take dozens of years. In this work, we propose to simulate the conditions in which a self-driving car could be. This will make it possible to assess the quality of localization for one route in all conditions with data from only one ride in good conditions. And by the quality of localization, you can assess the freshness of the map.

Since it is difficult to determine the accuracy of localization in real conditions because it is difficult to obtain the ground truth, we will evaluate the quality of localization in another way. We simulate the process of imaging the point cloud at a given location through a random heuristic generative process with a distribution p_{cloud}. Most localization methods use an initial approximation for their work, so we introduce a random process giving an initial approximation with a normal distribution p_{init}. The general distribution of the input data of the localization algorithm is described by the distribution p_{input} - the cartesian product of two other distribution. Then we can represent localization by a given map map, as a function:

$$pose = localization_{map}(cloud, init\_estimate) = localization_{map}(input)$$

We believe that, given a good map, the localization is robust to possible lidar clouds and initial approximations. In our method, the combination of clouds and

approximations is represented by the process p_{input}. Stability is expressed by the small value of the second moment of the pose. It is not possible to calculate this value analytically, but it can be estimated by Monte Carlo sampling from a random process p_{input}:

$$\mathbb{E}_{x\sim p_{input}}[pose(x)] = \int pose(x)p(x)dx \approx \frac{1}{m}\sum_{i}^{m} pose(x_i)$$

$$Cov_{x\sim p_{input}}[pose(x)] = \int \left(pose\,(x) - \mathbb{E}_{x\sim p_{input}}\left[pose\,(x)\right]\right)^2 p(x)dx$$

$$\approx \frac{1}{m}\sum_{i}^{m}\left(pose\,(x_i) - \frac{1}{m}\sum_{i}^{m} pose(x_i)\right)^2,$$

where m is the number of samples. If the obtained value is lower than the threshold selected on the dataset with examples of successful and unsuccessful localization, the quality of localization is considered good. Otherwise, we consider that in the vicinity of this point the map does not correspond to reality and it is necessary to update this part of the map. The analysis of the dataset of rides revealed the following and most frequent causes of lidar cloud degradation: other cars limiting the view, wet asphalt, precipitation, and fog. To simulate these conditions, the following modifications were applied to the generative process p_{cloud}: removal of point sectors with different lengths, removal of ground points, the addition of random points, and removal of points at a certain distance.

To test the modifications, we created a dataset from two parts: the first consists of rides in which there are noticeable differences between the map and reality. The second part consists of rides in areas on a very fresh map. The modifications of the clouds were evaluated by two metrics. `True Positive Rate`, which shows how well the used modifications help to identify areas with an outdated map. `True Negative Rate`, which shows that using the modification will not lead to false readings.

According to the results of the experiments, the best results were the removal of point sectors, on the four sides of the car, and the removal of ground points. The selected combination had the highest `TPR`, with the `TNR` value of 1. The other filters had either too low `TPR` or too low `TNR`.

[1] *Seiko H., Hu X.* Autonomous Driving in the iCity—HD Maps as a Key Challenge of the Automotive Industry // Engineering, 2016. Vol. 2(2). Pp. 159–162.

[2] *Endo Y., Javanmardi E., Kamijo S.* Analysis of Occlusion Effects for Map-Based Self-Localization in Urban Areas // Sensors, 2021. Pp. 51–96.

[3] *Reid T., Houts, S. et al* Localization Requirements for Autonomous Vehicles // SAE International Journal of Connected and Automated Vehicles, 2019. Vol. 2(3).

[4] *Jomrich F.* Dynamic Maps for Highly Automated Driving - Generation, Distribution and Provision // Technische Universität Darmstadt, 2020.

Параметрическая оценка наблюдаемых объектов по перспективным изображениям на базе методов перспективной геометрии, типизированных элементов и свёрточных нейронных сетей

Рихтер Андрей Александрович[1]⋆ urfin17@yandex.ru
Мурынин Александр Борисович[1,2] amurynin@bk.ru
Гвоздев Олег Геннадьевич[1,3] gvozdev@miigaik.ru
Козуб Владимир Александрович[1] postbox-kozub@yandex.ru
Пуховский Дмитрий Юрьевич[1] dpukhovskiy@yandex.ru

[1]Москва, Научно-исследовательский институт аэрокосмического мониторинга "АЭРОКОСМОС"
[2]Москва, Федеральный исследовательский центр "Информатика и управление" РАН
[3]Москва, Московский государственный университет геодезии и картографии

Предлагается метод восстановления трёхмерной модели объекта по одному перспективному изображению, основанный на универсальных геометрических особенностях объектов – так называемых типизированных элементах. Это объекты или их части, которые имеют типизированные (стандартизированные) размеры. Различаем три вида типизированных элементов: 1) конструкторы – «клетки», из которых состоит объект (например, кирпичи, сайдинговые панели, монолитные блоки); 2) формообразующие элементы – части объекта, задающие его структуру (например, оконные, дверные или арочные проёмы); 3) типизированные объекты – сложные конструкции, все параметры которых стандартизированы. При семантической сегментации типизированных элементов, объектов и их пространственных признаков на перспективном изображении с применением методов перспективной геометрии рассчитываются трёхмерная модель.

Метод состоит из следующих этапов: I) семантическая сегментация; II) определение опорных точек и образа системы координат; III) расчёт параметров образа прямоугольной системы координат; IV) пространственная привязка точек объекта.

I. Семантическая сегментация состоит из двух стадий – интегрального и локального анализа. Она базируется на топологии U-Net с различными расширениями: MultiResUNet, U-Net++, DeepUp v3 и др. Например, может применяться модификация U-Net, разработанная авторами и описанная в работе [1].

Целью интегрального анализа является решение задачи пообъектной сегментации изображения (задача instance segmentation). Искусственная нейронная сеть (ИНС) обучается для сегментации объектов определённых классов. Результатом данной стадии является выделение областей на изображении, где каждая из них соответствует одному объекту.

Для восстановления моделей объектов (зданий) требуется локальный анализ, задачей которого является выделение значимых элементов и характеристик отдельного объекта, по которым можно будет восстановить его форму.

Входными данными для работы ИНС является часть изображения, включающая объект интереса и его значимый контекст. Результатом работы являются сегментация пространственных положений элементов объекта относительно земной поверхности.

Задача локального анализа с точки зрения топологии ИНС сводится к двум типовым: 1) задача извлечения из изображения количественных признаков; 2) задача локализации геометрических примитивов.

II. По результатам семантической сегментации формируются объектные и пространственные признаки.

Опорные точки системы координат получаются на базе типизированных элементов (например, шпалы, рельсы, опоры контактной сети): o, g_i, t_i, t'_i, γ_i, $i = = 1..3$. Точки o , g_i, t_i , t'_i. оси ox_i, прямые γ_i – соответственно образы точек O, G_i, T_i, T'_i, осей OX_i, прямых Γ_i. o – начало координат; g_i – задают направления на оси координат ox_i. g_i могут быть в произвольных местах на оси или совпадать с одной из точек t_i или t'_i, если $t_i, t'_i \in ox_i$. Точки t_i и t'_i задаются так, что $T_iT'_i$ типизирован (например, стандартные расстояние между шпалами или рельсами, высота опоры), а $T_iT'_i \parallel OX_i$ либо $T_iT'_i \in OX_i$. γ_i – линии (построенные на опорных точках), для которых $\Gamma_i \parallel OX_i$.

Опорные точки a_i объекта - в «углах» его изображения.

III. Для преобразования координат на изображении в пространственные координаты необходимо рассчитать параметры образа $ox_1x_2x_3$ системы координат $OX_1X_2X_3$. К ним относятся: X_{mi} и x_{mi} – эталоны пространственных и осевых координат; $\overrightarrow{n_i}$ – направляющие осей (нормированные векторы, сонаправленные векторам $\overrightarrow{oe_i}$); f_i – фокусы на осях ox_i (точки пересечения прямых $t_it'_i$ и ox_i). Фокусом на оси ox_i условно назовём точку схождения образов прямых, параллельных оси OX_i и лежащих в смежных с ней координатных плоскостях.

Эталоны координат задают соответствие отчётам координат X_i на осях OX_i их отчёты x_i на осях ox_i по закону убывающей геометрической прогрессии.

IV. Пространственная привязка точек объекта осуществляется по сегментированным пространственным признакам. Геометрическая фигура Ф в пространстве идентифицируется по k опорным точкам на ней $A_i, i = 1 \dots k$, для которых известны их образы $a_i, i = 1..k$ - координаты на изображении. То есть находится преобразование $X = F(\psi)$, позволяющее найти пространственные координаты $X = (X_1, X_2, X_3)$ точки A, лежащей на этой фигуре, по координатам на изображении $\psi = (p, q)$ образа a этой точки.

Например, плоскость $\alpha = A_1A_2A_3$ идентифицируется по $k = 3$ точкам A_1, A_2, A_3. Тогда берётся произвольная точка A (например, в «углах»), полагая, что она лежит в той же плоскости. И пространственные координаты X точки A задаются по координатам на изображении (p, q) её образа a. Аналогично для других геометрических фигур: отрезки, лучи и прямые – по $k = 2$ точкам; окружности – по $k = 3$; эллипсы – по $k = 5$; сферические поверхности

– по $k = 4$; цилиндрические поверхности – по $k = 5$; конические поверхности – по $k = 6$; эллипсоиды – по $k = 6$ и др.

На разных этапах метода выполняются перспективные операции и операции геометрических достроений. К первым относятся, например, трансформации образов систем координат (сдвиг, поворот и масштабирование). Ко вторым относятся, например, достроения фигур по их частям.

Работа выполнена при финансовой поддержке Российской Федерации в лице Минобрнауки России в рамках соглашения № 075-15-2020-776.

[1] *Гвоздев О. Г., Козуб В. А., Кошелева Н. В., Мурынин А. Б., Рихтер А. А.* Нейросетевой метод построения трехмерных моделей ригидных объектов по спутниковым изображениям // Мехатроника, автоматизация, управление, 2021. Т. 2(1). С. 46–53.

Parametric estimation of observed objects from perspective images based on methods of perspective geometry, typed elements and convolutional neural networks

Rihter Andrej[1]⋆ urfin17@yandex.ru
Murynin Aleksandr[1,2] amurynin@bk.ru
Gvozdev Oleg[1,3] gvozdev@miigaik.ru
Kozub Vladimir[1] postbox-kozub@yandex.ru
Puhovskij Dmitrij[1] dpukhovskiy@yandex.ru

[1]Moscow, State scientific Institution for Scientific Research of Aerospace Monitoring "AEROCOSMOS"
[2]Moscow, Federal Research Center "Informatics and Management" RAS
[3]Moscow, Moscow State University of Geodesy and Cartography

A method is proposed for restoring a three-dimensional model of an object from a single perspective screen, based on the universal geometric features of objects - the so-called typed elements. These are objects or parts of them that have typed (standardized) sizes. We distinguish three types of typed elements: 1) constructors - "cells" that make up an object (for example, bricks, siding panels, monolithic blocks); 2) shaping elements - parts of an object that define its structure (for example, window, door or arched openings); 3) typed objects are complex constructions, all parameters of which are standardized. During the semantic segmentation of typed elements, objects and their spatial features on a perspective image using the methods of perspective geometry, a three-dimensional model is calculated.

The method consists of the following stages: I) semantic segmentation; II) determination of control points and coordinate system image; III) calculation of the parameters of rectangular coordinate system image; Iv) the spatial reference of the points of the object.

I. Semantic segmentation consists of two stages - integral and local analysis. It is based on U-Net topology with various extensions: MultiResUNet, U-Net++, DeepUp v3 etc. For example, a modification of U-Net, developed by the authors and described in the work, can be used. [1].

The purpose of integral analysis is to solve the problem of object segmentation of an screen (instance segmentation problem). An artificial neural network (ANN) is trained to segment objects of certain classes. The result of this stage is the selection of areas in the screen, where each of them corresponds to one object.

To restore models of objects (buildings), local analysis is required, the task of which is to highlight significant elements and characteristics of an individual object, by which it will be possible to restore its shape. The input data for the ANN operation is a part of the screen that includes the object of interest and its meaningful context. The result of the work is the segmentation of the spatial positions of the object's elements relative to the earth's surface.

The task of local analysis from the point of view of ANN topology is reduced to two typical ones: 1) the task of extracting quantitative features from the screen; 2) the problem of localizing geometric primitives.

II. Object and spatial features are formed based on the results of semantic segmentation. The reference points of the coordinate system are obtained on the basis of typed elements (for example, sleepers, rails, contact network supports): o, g_i, t_i, t'_i, γ_i, $i = 1..3$. The points o , g_i, t_i , t'_i. the axes ox_i, the lines γ_i – respectively the images of points O, G_i, T_i, T'_i, the axes OX_i, the lines Γ_i. o – origin; g_i – set directions on the coordinate axis ox_i. g_i can be in arbitrary places on the axis or coincide with one of the points t_i or t'_i, if $t_i, t'_i \in ox_i$. Points t_i and t'_i are given so that $T_iT'_i$ typed (for example, standard distance between sleepers or rails, support height), and $T_iT'_i \parallel OX_i$ or $T_iT'_i \in OX_i$. γ_i – lines (plotted on control points) for which $\Gamma_i \parallel OX_i$.

Object anchor points a_i - in the "corners" of its screen.

III. To transform the coordinates on screen into spatial coordinates, it is necessary to calculate the parameters of the image $ox_1x_2x_3$ of coordinate system $OX_1X_2X_3$. These include: X_{mi} and x_{mi} – standards of spatial and axial coordinates; $\overrightarrow{n_i}$ – axis guides (normalized vectors co-directional vectors $\overrightarrow{oe_i}$); f_i – focuses on the axes ox_i (intersection points of straight lines $t_it'_i$ and ox_i). Focus on the axis ox_i conventionally call the point of convergence of images of straight lines parallel to the axis OX_i and lying in adjacent coordinate planes.

Coordinate references define correspondence to coordinate reports X_i on axles OX_i their reports x_i on axles ox_i according to the law of geometric progression. Namely, the distance between adjacent reports x_i decreases exponentially as the point tends x_i to the point x_{f_i}.

IV. Spatial reference of points of an object is carried out according to segmented spatial features. Geometric figure Φ in space is identified by k anchor points $A_i, i = = 1 \ldots k$ on it, for whom their images are known $a_i, i = 1..k$ - coordinates on the screen. That is, there is a transformation $X = F(\psi)$, allowing to find spatial coordinates $X = (X_1, X_2, X_3)$ of the point A, lying on this figure, according to the coordinates on the screen $\psi = (p, q)$ of the image a of this point.

For example, the plane $\alpha = A_1A_2A_3$ identified by $k = 3$ points A_1, A_2, A_3. Then an arbitrary point is taken A (for example, in the "corners"), assuming that it lies in the same plane. And spatial coordinates X of point A are set by coordinates on screen (p, q) of its image a. Likewise for other geometric shapes: lines, rays and lines – by $k = 2$ points; circles – by $k = 3$; ellipses – by $k = 5$; spherical surfaces – by $k = 4$; cylindrical surfaces – by $k = 5$; tapered surfaces – by $k = 6$; ellipsoids – by $k = 6$ et al.

At different stages of the method, perspective operations and operations of geometric extensions are performed. The first one include, for example, transforming coordinate system images (shift, rotation, and scaling). The second one include, for example, the completion of figures by their parts.

The work was carried out with the financial support of the Russian Federation represented by the Ministry of Education and Science of the Russian Federation under agreement No. 075-15-2020-776.

[1] *Gvozdev O., Kozub V., Kosheleva N., Murynin A., Rihter A.* Neural network method for constructing three-dimensional models of rigid objects from satellite screens // Mechatronics, automation, control, 2021. Vol. 2(1). Pp. 46–53.

Комбинирование сегментированных изображений на основе минимизации информационной избыточности

Мурашов Дмитрий Михайлович[1]⋆ d_murashov@mail.ru
[1]Москва, ФИЦ ИУ РАН

Один из подходов к повышению качества сегментации изображений, основан на комбинировании сегментированных изображений с целью оптимизации заданного функционала качества. Например, известен итерационный метод комбинирования множества грубо сегментированных изображений, полученных в разных цветовых пространствах при различных значениях параметров алгоритма кластеризации [1]. Для обеспечения качества автор использовал критерий минимума средней вариации информации между комбинированным изображением и каждым из грубых вариантов сегментации. С помощью алгоритма ICM (Iteration Conditional Modes) выполнялась коррекция границ сегментов. В качестве начального приближения для оптимизационной процедуры использовалось «медианное разбиение» оригинального изображения относительно критерия средней вариации информации, вычисляемого на множестве рассматриваемых разбиений.

Ранее было предложено решать задачу обеспечения качества сегментации как задачу выбора из множества сегментированных изображений такого изображения, которое бы минимизировало информационную избыточность [2]. Была показана эффективность критерия минимума информационной избыточности.

В представляемой работе предлагается двухуровневый метод комбинирования сегментаций на стадии постобработки на основе критерия минимума информационной избыточности. На первом уровне осуществляется комбинирование сегментов. На втором уровне производится итерационная попиксельная коррекция границ сегментов комбинированного изображения. Комбинирование сегментов позволит получить более точное разбиение оригинального изображения и сохранить информационно важные области, которые могут быть утрачены при работе традиционных алгоритмов сегментации.

Пусть с помощью одного или нескольких алгоритмов сегментации получен набор разбиений V_q, $q = 1, 2, \ldots, Q$ оригинального изображения U: $\mathcal{V} = \{V_1, V_2, \ldots, V_Q\}$.

Пусть изображение V_{qmin} обеспечивает минимум меры информационной избыточности $R(U, V_q) = 1 - \frac{I(U;V_q)}{H(V_q)}$, где $I(U; V_q)$ - средняя взаимная информация между входом и выходом алгоритма сегментации, $H(V_q)$ - энтропия выхода. Необходимо, комбинируя сегменты изображений из $\mathcal{V}$, получить изображение V_{Comb} такое, что $R(U, V_{Comb}) < R(U, V_{qmin})$, $R(U, V_{Comb}) \to min$.

Процедура комбинирования реализуется следующим образом. В качестве начального приближения изображения V^0_{Comb} из множества сегментаций $\mathcal{V}$ выбирается изображение V_{qmin}. Далее выполняется последовательное сравнение сегментов V^0_{Comb} и одного из изображений V_q, $q \neq q_{min}$. Если какой-либо

сегмент из V_q отсутствует на V^0_{Comb}, то он комбинируется с V^0_{Comb}, и, таким образом, формируется новое изображение V^1_{Comb}. Если выполняется условие $R(U, V^1_{Comb}) < R(U, V^0_{Comb})$, то полученное разбиение V^1_{Comb} сохраняется, и проверяется следующий сегмент, который в комбинации с изображением V^1_{Comb} порождает новое разбиение V^2_{Comb}. Если условие убывания избыточности информации не выполняется, то с использованием следующего сегмента формируется новое изображение V^1_{Comb}. После проверки всех сегментов изображения V_q, выбирается следующее разбиение из множества $\mathcal{V}$, и процедура повторяется.

На втором уровне уточняются границы сегментов полученного комбинированного изображения V_{Comb}. В качестве базового алгоритма использована итерационная процедура, предложенная в работе [1]. В предлагаемом алгоритме, в отличие от оригинала, вместо критерия минимума средней вариации информации применяется мера избыточности информации $R(U, V_{Comb})$, содержащейся в комбинированном изображении. Процедура состоит в следующем. Последовательно пикселям, которые находятся на границе сегмента, присваивается метка соседнего сегмента. Вычисляется приращение меры избыточности. Если приращение отрицательно, то сохраняется новая метка сегмента. Если приращение положительно, метка восстанавливается. Далее меняется метка у следующего пикселя на границе сегмента, и описанные выше операции повторяются. Получены формулы для вычисления приращения меры избыточности.

Проведен вычислительный эксперимент на изображениях из базы изображений BSDS500 университета Беркли. Результаты эксперимента показали, что предложенный двухуровневый алгоритм комбинирования сегментированных изображений позволяет улучшить результат сегментации изображений с точки зрения минимума информационной избыточности, и при этом выделить информационно значимые области. Комбинированные разбиения в большинстве случаев продемонстрировали большее информационное сходство с эталонными сегментациями, чем разбиения, полученные традиционным алгоритмом.

[1] *Mignotte M.* A Label Field Fusion Model With a Variation of Information Estimator for Image Segmentation // Information Fusion, 2014. Vol. 20. Pp. 7–20.

[2] *Murashov D.* An Information Model of Image Segmentation Algorithm Based on Redundancy Minimization // 2020 International Conference on Information Technology and Nanotechnology (ITNT), 2020. Pp. 1–7.

Combining segmented images based on information redundancy minimization

Murashov Dmitry[1]⋆ d_murashov@mail.ru

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

One of the approaches to improving the image segmentation quality is based on applying a fusion model to combine segmented images and optimize a given quality measure. For example, an iterative method is known for combining a set of roughly segmented images obtained in different color spaces for different values of the parameters of the clustering algorithm (see [1]). To provide segmentation quality, the author used the criterion of the minimum of the mean variation of information between the combined image and each of the coarse segmentation. The segment boundaries were corrected using the ICM (Iteration Conditional Modes) algorithm. As an initial approximation for the optimization procedure, the author used the "median partition" of the original image with respect to the criterion of the average variation of information calculated on the set of the partitions under consideration.

Earlier it was proposed to solve the problem of segmentation quality as the problem of choosing from a set of segmented images such an image that would minimize information redundancy [2]. The effectiveness of the criterion of minimum information redundancy was demonstrated.

In this paper, we propose a two-level method for combining segmented images based on the criterion of minimum information redundancy. This method should be applied at the post-processing stage. At the first level, we combine the segments. At the second level, an iterative pixel-by-pixel correction of the boundaries of the combined image segments is performed. Combining the segments will allow one to get a more accurate partition of the original image and preserve information about important areas that can be lost when using conventional segmentation algorithms.

Let a set of partitions V_q, $q = 1, 2, \ldots, Q$ of the original image U be obtained using one or several segmentation algorithms: $\mathcal{V} = \{V_1, V_2, \ldots, V_Q\}$. Let the image V_{qmin} provides a minimum measure of information redundancy $R(U, V_q) = 1 - \frac{I(U;V_q)}{H(V_q)}$, where $I(U; V_q)$ is the average mutual information between the input and output of the segmentation algorithm, $H(V_q)$ is the entropy of the output. It is necessary by combining image segments from $\mathcal{V}$ to obtain an image V_{Comb} such that $R(U, V_{Comb}) < R(U, V_{qmin})$, and $R(U, V_{Comb})$ tends to a minimum.

We propose the following combination procedure. As an initial approximation of the V^0_{Comb} image, we choose the V_{qmin} image from the set of segmentations $\mathcal{V}$. Next, we sequentially compare the segments of the image V^0_{Comb} and one of the images V_q, $q \neq q_{min}$. If a segment from V_q is missing in the V^0_{Comb}, then it is combined with the V^0_{Comb}, and thus a new V^1_{Comb} image is formed. If the condition $R(U, V^1_{Comb}) < R(U, V^0_{Comb})$ is satisfied, then we save the resulting partition V^1_{Comb}, and check the next segment, which in combination with the image V^1_{Comb} generates

a new partition V^2_{Comb}. If the condition for decreasing information redundancy is not met, then using the next segment we form a new image V^1_{Comb}. After checking all segments in image V_q, we choose the next partition from the set $\mathcal{V}$ and repeat the described above procedure.

At the second level of the proposed method, we refine the boundaries of the segments of the combined V_{Comb} image. As a basic algorithm, we used the iterative procedure proposed in [1]. In the proposed algorithm, in contrast to the original, instead of the criterion of the minimum average variation of information, we use the measure of redundancy of information $R(U, V_{Comb})$ contained in the combined image. The algorithm includes the following operations. Sequentially, we assign a label of the adjacent segment to pixels of the analyzed segment located on the border. We then calculate the increment of the redundancy measure. If the increment is negative, then a new segment label is assigned to the current pixel. If the increment is positive, the label is restored. Next, we change the label of the next pixel on the segment border and repeat the operations described above. Formulas for calculating the increment of the redundancy measure are obtained.

A computational experiment was carried out on images from the BSDS500 image database of the University of Berkeley. The experimental results showed that the proposed two-level algorithm for combining segmented images makes it possible to improve image segmentation in terms of minimum information redundancy and, at the same time, take into account informationally significant areas. Combined partitions, in most cases, demonstrated greater informational similarity with the reference segmentations than partitions obtained by the traditional algorithm.

[1] *Mignotte M.* A Label Field Fusion Model With a Variation of Information Estimator for Image Segmentation // Information Fusion, 2014. Vol. 20. Pp. 7–20.

[2] *Murashov D.* An Information Model of Image Segmentation Algorithm Based on Redundancy Minimization // 2020 International Conference on Information Technology and Nanotechnology (ITNT), 2020. Pp. 1–7.

Особенности моделей прогнозирования на основе компонент временного ряда в эпидемиологии и экономике

Добролюбова Ольга Анатольевна[1,2] dbrl.olga@gmail.com
[1]Москва, ФБУН ЦНИИ Эпидемиологии Роспотребнадзора
[2]Москва, МГУ имени М.В.Ломоносова

Прогнозирование временных рядов является одним из важнейших элементов анализа для принятия стратегических решений. В данном исследовании рассматриваются ряды, для которых невозможно произвести контролируемый эксперимент. Более того выявление всех факторов, влияющих на динамику ряда, является практически невыполнимой задачей.

Работа посвящена исследованию подходов к прогнозированию рядов в зависимости от результатов декомпозиции ряда. То есть для каждого ряда строится одномерная модель: $y_t = T_t + S_t + C_t + \varepsilon$, где y_t - наблюдение в момент t, T_t - тренд, S_t - сезонная компонента, C_t - цикличная компонента, ε - шумовая составляющая, получаемая после удаления других компонентов.

Для исследования предлагается рассмотреть два типа рядов – экономический ряды и эпидемиологические ряды. Экономические ряды включают в себя временные ряды цен акций, эпидемиологические – ряды неконтролируемых вирусных инфекций. Для динамики цен акций характерно наличие высокого порядка авторегрессионной функции и линейность в трендовой компоненте. В то же время, для эпидемиологических кривых также присутствует высокий порядок авторегрессии, однако тренд является нелинейным - для ряда характерны периоды спада и подъема. Более того предположения о случайности, вообще говоря, не выполняется для эпид-процессов, в отличие от экономических. В связи с этим подход к прогнозированию для этого типа данных должен быть отличен.

В качестве моделей для построения прогноза использовались регрессионные модели и градиентный бустинг. Для оценки качества прогноза использовалась классическая функция потерь – среднеквадратичная ошибка. Также для проверки правильности настройки модели использовался тест Дикки-Фулера для тестирования стационарности остатков.

Результаты позволяют судить о том, что регрессионные модели лучше описывают экономические ряды, в то время как с помощью градиентного бустинга удается построить лучший прогноз для эпидемиологических данных.

[1] *Durbin J., Koopman S.* Time Series Analysis by State Space Methods // Oxford Univ Pr; 1st edition, 2001. 253 p.

Features of forecasting models based on the time series components in epidemiology and economics

Dobroliubova Olga[1,2]⋆ dbrl.olga@gmail.com

[1]Moscow, Central Research Institute of Epidemiology (Rospotrebnadzor)

[2]Moscow, Lomonosov Moscow State University

Time-series forecasting is one of the most critical analysis elements for strategic decisions. This analysis consists of several series. Moreover, the identification of all the factors influencing the dynamics of the series is practically impracticable.

The work is devoted to the study of approaches to forecasting depending on the results of the decomposition of the series. That is, a univariate model is built for each row: $y_t = T_t + S_t + C_t + \varepsilon$, where y_t - refers to the observation vector at time t, T_t - trend, S_t - seasonal component, C_t - cycle, ε - irregular.

For the study, it is proposed to consider two series types - economic series and epidemiological series. The economical series includes the stock price time series, while the epidemiological series includes the sequence of uncontrolled viral infections. A high-order autoregressive function and linearity trend component characterizes the dynamics of prices. At the same time, for epidemiological curves, there is a high order of autoregression. Still, the trend is nonlinear - for several periods, there are periods of decline and rise. Moreover, the assumptions about randomness in question are not fulfilled for epidemiological processes, in contrast to economic ones. In this regard, the approach to forecasting for this type of data should be different.

Regression models and gradient boosting were used as models for making predictions. To assess the quality of the forecast, we used the classical loss function - the root-mean-square error. Also, I used the Dickey-Fuller test was used to test the stationarity of the residuals.

The results suggest that regression models are better at describing economic series, while using gradient boosting, it is possible to construct a better forecast for epidemiological data.

[1] *Durbin J., Koopman S.* Time Series Analysis by State Space Methods // Oxford Univ Pr; 1st edition, 2001. 253 p.

Поиск ключевых слов на изображениях рукописей средневековых исландских нарративных памятников

Качура Александр Сергеевич[1] kachuraalexandr@mail.ru
Липкина Анна Львовна[1]⋆ lipkina96@mail.ru
Литовских Елена Владимировна[2] elitovskih@mail.ru
Рейер Иван Александрович[3] reyer@forecsys.ru
[1]Москва, МГУ, ВМК
[2]Москва, ИВИ РАН
[3]Москва, ФИЦ ИУ РАН

Работа с большим массивом рукописей средневековых исландских нарративных памятников как с текстами в электронном виде затруднена, потому что подготовка такого текста (выверенного и вычитанного) требует отдельных усилий и времени. На настоящий момент электронные версии многих текстов отсутствуют, они доступны только в виде отсканированных изображений листов рукописи.

К техническим проблемам здесь следует отнести нераспространенный язык рукописей (хотя и на основе латиницы), особенности написания букв латинского алфавита (в т.ч. использование нескольких вариантов для одной буквы), нечеткую систему сокращений и использование диакритических знаков.

В докладе рассматриваются задачи количественного анализа текста применительно к «Книге о занятии земли».

«Книга о занятии земли» (*Landnámabók*) – средневековое историческое произведение, повествующее о первом периоде заселения Исландии. Она дошла до нас в пяти редакциях. Представляется важным определить авторский вклад в создание каждой из редакций «Книги». Для этого, в первую очередь, следует разделить формульные (стандартизированные) фрагменты текста и «пряди» и висы (соответственно, прозаические и поэтические вставки) с помощью поиска ключевых слов формул на изображениях листов рукописи.

Вторая редакция (*Hauksbók*, «Книга Хаука», по имени ее редактора, исландского ученого Хаука Эрлендссона) взята нами, поскольку она сохранилась в хорошо читаемом бумажном списке середины XVII в. (AM 105 fol), сделанном рукой известного переписчика Йоуна Эрлендссона, т.е. в рассматриваемой нами рукописи присутствует единственный почерк.

Мы решили начать с формулы *bjó* («жил», вариант *bjógdhu* / *bjógdhi* «жили»), вводящей массив топонимической информации. Формула задается набором образцов ключевых слов с различными вариантами написания.

Дополнительным условием поиска является выделение максимального количества действительных вхождений формулы в тексте. Это объясняется тем, что затраты на визуальную проверку истинно положительных результатов среди выделенных кандидатов не очень высоки. Таким образом, целевой метрикой качества в нашей задаче является полнота (recall), но при условии сохранения

приемлемого для возможности ручного поиска значения точности распознавания (precision).

Предлагаемое нами решение включает следующие этапы: 1) бинаризация изображения; 2) сегментация частей текста (абзацы, строки, слова); 3) подсчет меры сходства образцов формулы и сегментированных слов на основе графемного подхода [1]; 4) выбор слов-кандидатов на основе меры сходства.

Бинаризация цветных изображений листов рукописи проводится с использованием нейросетевой модели, подготовленной С. Лидесом (`https://github.com/sliedes/binarize`). Модель является модифицированным вариантом сети U-Net, обученным на изображениях документов Национального архива Финляндии.

Для сегментации изображения на строки и слова используется свободное ПО Tesseract OCR, в основе которого лежит статистический анализ параметров связных компонент и расположения выделенных элементов на странице [2]. Разбиение страницы на абзацы также проводится на основе анализа расположения черных областей на бинаризованном изображении и применения к ним морфологической операции дилатации. Выделенные абзацы, строки и слова связываются в единую структуру.

Для сравнения заданных образцов формулы с сегментированными словами строится скелетное представление [1] бинаризованных изображений образцов. Сходство слов оценивается по «расстоянию редактирования» (Graph Edit Distance, GED) [3] между графами образцов формулы и графами изображений слов. Итоговое расстояние между словом и формулой определяется как минимальное значение GED среди всех образцов формулы.

Для минимизации количества попарных сравнений между графами мы предварительно отбираем среди всех выделенных слов текста некоторое подмножество слов-кандидатов. Эти слова выбираются на основе меры сходства Intersection over Union (IoU) [4] между словом и образцом формулы. Подсчет IoU реализован таким образом, что возможно выделение части слова, наиболее похожей на заданный образец. Итоговый коэффициент сходства между словом и формулой определяется как максимальное значение IoU среди всех образцов формулы.

Пороговые значения метрик IoU и GED были подобраны эмпирическим путем.

В результате экспериментов на тестовом множестве размеченных изображений листов мы получили следующие значения метрик качества: precision = 0.37, recall = 0.91. Такое качество представляется приемлемым в текущей постановке задачи, поскольку нахождение всех вхождений формул существенно важнее, чем сужение множества предполагаемых кандидатов.

Работа поддержана грантами РФФИ No. 20-01-00664 и No. 20-07-00990.

[1] *Липкина А. Л., Местецкий Л. М.* Метод распознавания шрифтов на основе медиального представления // Труды Межд. конф. по компьютерной графике и зрению «Графикон», 2020. С. 118–129.

[2] *Smith R.* An overview of the Tesseract OCR engine // Ninth int. conf. on document analysis and recognition (ICDAR 2007), 2007. Pp. 629–633.

[3] *Sanfeliu A., Fu K.* A distance measure between attributed relational graphs for pattern recognition // IEEE transactions on systems, man, and cybernetics, 1983. No. 3. Pp. 353–362.

[4] *Rezatofighi H. et al.* Generalized intersection over union: A metric and a loss for bounding box regression // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. Pp. 658–666.

Keyword search in images of manuscripts of medieval Icelandic narrative sources

Kachura Alexandr[1] kachuraalexandr@mail.ru
Lipkina Anna[1]⋆ lipkina96@mail.ru
Litovskikh Elena[2] elitovskih@mail.ru
Reyer Ivan[3] reyer@forecsys.ru
[1]Moscow, MSU
[2]Moscow, IGH RAS
[3]Moscow, FRC CSC RAS

Working with a large corpus of manuscripts of medieval Icelandic narrative sources as electronic texts is complicated, because the preparation of such a text (verified and proofread) requires particular time and effort. Nowadays electronic versions of many texts are available only as scanned images of manuscript pages.

Technical problems in this case include the uncommon language of manuscripts (albeit it is based on the Latin alphabet), the features of writing Latin letters (using several variants for a letter), a variable abbreviation system, and the use of diacritics.

The paper considers the problems of quantitative text analysis in relation to the "Book of Settlements".

Landnámabók ("Book of Settlements") is a medieval historical work that tells about the first period of the settlement of Iceland. It has reached us in five editions. It seems important to determine the authors' contribution to the creation of each edition of the "Book". To do this, it is necessary in the first instance to separate the formalized text fragments from *thættir* and *vísur* (prosaic and poetic insertions respectively) by searching for formula keywords on the images of manuscript pages.

We took the second edition (*Hauksbók*, "Book of Hauk", named after its editor, Icelandic scholar Haukr Erlendsson), since it is preserved in a well readable paper copy of the mid-17th century (AM 105 fol) written by renowned scribe Jón Erlendsson, i.e. the manuscript is made in a single handwriting.

We decided to start with the *bjó* formula ("[he/she] lived", also including the variant *bjógdhu / bjógdhi*, "[they] lived") that introduces a block of toponymic information. The formula is defined by a set of sample keywords with various spellings.

An additional search condition is to identify the maximum number of valid occurrences of the formula in the text. This is because the cost of visual control of the selected candidates is not very high. Thus, the target quality metric in our case is recall, provided the precision of the recognition is acceptable for visual search.

The proposed solution includes the following steps: 1) image binarization; 2) segmentation of text parts (paragraphs, lines, words); 3) calculation of a similarity measure between formula samples and segmented words using the grapheme approach [1]; 4) selection of candidate words based on the similarity measure.

Color images of manuscript pages are binarized using a neural network model prepared by S. Liedes (`https://github.com/sliedes/binarize`). The model is

a modified version of the U-Net trained on documents from the Finnish National Archive.

The Tesseract OCR engine is used for segmentation of a binary image into lines and words. The software is based on a statistical analysis of connected components' parameters and the location of selected elements on a page [2]. Breaking a page into paragraphs is also based on the analysis of black areas' location on the image and the application of morphological dilation to them. Segmented paragraphs, lines, and words are combined into a single structure.

To compare given formula samples with segmented words, a skeletal representation [1] of binarized sample images is constructed. The word similarity is estimated by Graph Edit Distance (GED) [3] between the skeletal graphs of formula samples and word images. The final distance between a word and a formula is defined as the minimum GED value among all formula samples.

To minimize the number of pairwise comparisons between graphs, we pre-select a subset of candidate words. These words are selected using Intersection over Union (IoU) [4] between a word and a formula sample. The IoU calculation is implemented in such a way that it is possible to select a part of a word being the most similar to a given pattern. The final similarity score between a word and a formula is defined as the maximum IoU value among all formula samples.

The threshold values of the IoU and GED metrics were selected empirically.

Experiments on a test set of labeled page images provided the following quality estimates: precision = 0.37, recall = 0.91. This quality seems to be acceptable for the current formulation of the problem, since finding all occurrences of a formula is much more important than narrowing the set of supposed candidates.

This research is funded by RFBR, grants 20-01-00664 and 20-07-00990.

[1] *Lipkina A., Mestetsky L.* Medial representation based font recognition method // Proceedings of the Int. conf. on computer graphics and vision ¡¡ Graphicon ¿¿, 2020. Pp. 118–129.

[2] *Smith R.* An overview of the Tesseract OCR engine // Ninth int. conf. on document analysis and recognition (ICDAR 2007), 2007. Pp. 629–633.

[3] *Sanfeliu A., Fu K.* A distance measure between attributed relational graphs for pattern recognition // IEEE transactions on systems, man, and cybernetics, 1983. No. 3. Pp. 353–362.

[4] *Rezatofighi H. et al.* Generalized intersection over union: A metric and a loss for bounding box regression // Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. Pp. 658–666.

Классификация сообщений новостного потока, предположительно «участвующих» в реализации PUMP-стратегий на фондовом рынке

Инякин Андрей Сергеевич[1]⋆ inyakin@forecsys.ru
Кормаков Георгий Владимирович[1,2] kormakov_georgiy@forecsys.ru
Каширин Даниил Олегович[1] kashirin@forecsys.ru
Мусин Шамиль Наилович[1] musin_shamil@forecsys.ru
Сотнезов Роман Михайлович[1,3] sotnezov@forecsys.ru
Разин Николай Алексеевич[4] razinna@cbr.ru
Саутенков Иван Сергеевич[4] sautenkovis@cbr.ru

[1]Москва, ООО «Форексис»
[2]Москва, МГУ имени М.В.Ломоносова
[3]Москва, МФТИ
[4]Москва, Центральный Банк РФ

В связи со стремительным развитием брокерских приложений в последние годы на рынок попали неквалифицированные инвесторы. Их стратегия часто руководствуется новостным потоком и советами в различных блогах (в том числе, телеграм-каналах).

Со стороны инсайдеров и остальных игроков активнее стали проявляться *PUMP-стратегии* – стратегии, направленные на «накачку» торгуемого инструмента сделками на фоне активного информационного потока. Часто в качестве инструментов выбираются волатильные позиции, позволяющие с помощью повышения объёма торгов влиять на цену (и наоборот).

На текущий момент достаточно сложно оценить признаки, характеризующие осуществление PUMP-стратегии. Это могут быть как характеристики исторической волатильности инструмента (например, наличие рывков в показаниях цен и объёмов рынка по инструменту), так и предшествующие объёмные покупки или продажи некоторыми инсайдерами.

С точки зрения поставленной задачи необходимо создать возможность изъятия информации из новостного потока, оценить некоторые признаки, свидетельствующие об отклонении торгуемого инструмента от «нормального» хода, и осуществить классификацию опубликованной новости на предмет осуществления PUMP-стратегии (т.е. провести бинарную классификацию).

Характеристики инструмента (по минутным японским свечам) оценивались двумя стратегиями: для цен проводилась типизация свечек, выделение тренда по предыстории и оценка тела свечи (описание характеристик см. в [1]); для признаков на объёмах торгов использовались статистические признаки на предыстории в несколько отсчётов, а также оценивались мета-признаки различных алгоритмов выделения выбросов по найденным признакам.

На итоговых признаках обучены две модели случайных лесов (в экспериментах они показали лучшее качество среди взятых моделей). Итоговый алгоритм

классификации свечек – это взвешенное голосование этих моделей. Основной целью модели на характеристиках инструмента – это полное покрытие случаев рывков, связанных с PUMP-новостями, с минимизацией ложно-положительных срабатываний (т.е. максимизация полноты (recall) на экспертно предоставленной разметке с достижением приемлемого процента FP-срабатываний).

Результаты обученного на данных по двум инструментам в период с 2019 по 2020 год (на имеющихся интервалах с разметкой PUMP-рывков) классификатора приведены в таблице 1. Результаты представлены на отложенном инструменте.

Таблица 1. Результаты алгоритма классификации свечей

	precision	recall	f1	support
Не PUMP	1.00	1.00	1.00	1105301
PUMP	0.00	1.00	0.01	14

Количество ложно положительных срабатываний в приведённых результатах равно 3556. Данный результат является лучшим среди проводимых экспериментально.

Для итоговой классификации новости необходима её классификация по текстовой стилистике. Исследовались модели TF-IDF, Word2Vec и Doc2Vec представления текстов. Результаты Word2Vec показали, что модель классификации на данном представлении неуверенно выделяет PUMP-новости на отложенной выборке. Обучение моделей проводилось на текстах, опубликованных в финансовых блогах и телеграм каналах.

Итоговый алгоритм был обучен на Doc2Vec представлениях как взвешенное голосование двух моделей, обученных на двух выборках разнородных текстов. Качество на отложенной выборке продемонстрировано в таблице 2.

Таблица 2. Результаты алгоритма классификации новостей

	precision	recall	f1	support
Не PUMP	1.00	1.00	1.00	6816
PUMP	0.88	0.87	0.87	254

Влияние признаков инструментов на новость оценивалось количеством рывков в окрестности новости. В таблице 3 приведены результаты ансамбля моделей на Doc2Vec признаках и на суммарном числе рывков в окрестности по всем инструментам. Также исследована модель на разреженном пространстве

Таблица 3. Результаты итогового ансамбля на сумме рывков

	precision	recall	f1	support
Не PUMP	1.00	1.00	1.00	6816
PUMP	0.97	0.96	0.97	254

суммарных рывков по каждому инструменту (из фиксированного списка инструментов). Результаты взвешенного голосования данной модели с моделью doc2vec приведены в таблице 4.

Таблица 4. Результаты итогового ансамбля на всех рывках

	precision	recall	f1	support
Не PUMP	1.00	1.00	1.00	6816
PUMP	0.91	0.90	0.90	254

Результаты подтверждают влияние числа рывков по торговому инструменту в окрестности даты публикации на тип новости.

Работа выполнялась с использованием инфраструктуры Центра коллективного пользования «Высокопроизводительные вычисления и большие данные» ФИЦ ИУ РАН (ЦКП «Информатика»). Работа поддержана грантом РФФИ No. 19-07-00885.

[1] *Gdakowicz A.* The application of Japanese candlestick charting on the residential real estate market // Real Estate Management and Valuation, 2014. Pp. 27–34.

Classification of news stream messages supposedly "involved" in the implementation of PUMP strategies on the stock market

Inyakin Andrey[1]⋆ inyakin@forecsys.ru
Kormakov Georgii[1,2] kormakov_georgiy@forecsys.ru
Kashirin Daniil[1] kashirin@forecsys.ru
Musin Shamil[1] musin_shamil@forecsys.ru
Sotnezov Roman[1,3] sotnezov@forecsys.ru
Razin Nikolay[4] razinna@cbr.ru
Sautenkov Ivan[4] sautenkovis@cbr.ru

[1]Moscow, Forecsys
[2]Moscow, Lomonosov Moscow State University
[3]Moscow, Moscow Institute of Physics and Technology
[4]Moscow, Central Bank of Russia

Due to the rapid development of brokerage applications in recent years, unqualified investors have entered the market. Their strategy is often guided by the news flow and advice in various blogs (including telegram channels).

On the part of insiders and other players, in connection with this phenomenon, *PUMP strategies* have become more active – strategies aimed at "pumping" the traded instrument with transactions by increasing the information flow. Most often, volatile positions are chosen as instruments, allowing to influence the price characteristics by increasing the trading volume (and vice versa).

At the moment, it is quite difficult to assess the features characterizing the implementation of the PUMP strategy. These can be both characteristics of the historical volatility of the instrument (for example, the presence of surges in the readings of prices and market volumes for the instrument), and previous volume purchases or sales by some insiders.

From the point of view of the task, it was necessary to create the possibility of removing information from the news stream, evaluate some features indicating a deviation of the traded instrument from the "normal" course, and classify the published news for the implementation of the PUMP strategy (i.e., conduct a binary classification).

To evaluate the characteristics of the instrument (by minute Japanese candlesticks), two strategies for obtaining features were used: for the price values of Japanese candlesticks, candle typing was carried out, trend selection by background and evaluation of the candle body (a description of the characteristics can be found in [1]); to evaluate features on trading volumes, statistical features on the background in several counts were used, and meta-features of various algorithms for allocating emissions by found features were evaluated.

Two models of random forests were trained on the obtained features (as part of the experiments, they showed the best quality among the tested models) and the final algorithm for classifying candlesticks as a weighted vote of two models was

obtained. The main purpose of the model based on the characteristics of the tool was to fully cover the cases of jerks associated with PUMP-news with minimizing false-positive triggers (i.e. maximizing completeness (recall) on expertly provided markup with achieving an acceptable percentage of FP-triggers).

The results of the classifier trained on data for two instruments in the period from 2019 to 2020 (at the available intervals with the marking of PUMP-jerks) are shown in Table 1. The results are presented on the deferred instrument. The number

Table 1. Results of the candle classification algorithm.

	precision	recall	f1	support
Not PUMP	1.00	1.00	1.00	1105301
PUMP	0.00	1.00	0.01	14

of false positives in the above results is 3556. This result is the best among those conducted experimentally.

The next stage for the final classification of the news is the classification of the published news. TF-IDF, Word2Vec and Doc2Vec models of text representation were investigated. The results of Word2Vec showed that the classification model on this view hesitantly highlights PUMP-news on a deferred sample. The models were trained on texts published in financial blogs and telegram channels.

The final algorithm was trained on Doc2Vec representations as a weighted vote of two models trained on two samples of heterogeneous texts. The quality on the deferred sampling is demonstrated in the table 2.

Table 2. Results of the news classification algorithm.

	precision	recall	f1	support
Not PUMP	1.00	1.00	1.00	6816
PUMP	0.88	0.87	0.87	254

As an assumption about the influence of the features of tools on the news, a feature was taken that estimates the number of surges in the vicinity of the news. Table 3 shows the results of an ensemble of models on Doc2Vec features and on the total number of surges in the neighbourhood for all instruments.

Table 3. The results of the final ensemble on the sum of surges.

	precision	recall	f1	support
Not PUMP	1.00	1.00	1.00	6816
PUMP	0.97	0.96	0.97	254

A model on a sparse space of total surges for each instrument (from a fixed list of instruments) is also investigated. The results of weighted voting of this model with the doc2vec model are shown in the table 4.

Table 4. The results of the final ensemble in all surges.

	precision	recall	f1	support
Not PUMP	1.00	1.00	1.00	6816
PUMP	0.91	0.90	0.90	254

The research was carried out using the infrastructure of the shared research facilities ≪High Performance Computing and Big Data≫ of FRC CSC RAS (CKP ≪Informatics≫).

This research is funded by RFBR, grant 19-07-00885.

[1] *Gdakowicz A.* The application of Japanese candlestick charting on the residential real estate market // Real Estate Management and Valuation, 2014. Pp. 27–34.

Согласование смысловых эталонов и взаимная релевантность документов тематического корпуса

Михайлов Дмитрий Владимирович ⋆ mdv74@list.ru
Емельянов Геннадий Мартинович Gennady.Emelyanov@novsu.ru
Великий Новгород, Россия, НовГУ

Настоящая работа посвящена проблеме единства и целостности образа смыслового эталона, выделяемого по фразам для тематического текста. Близость текста эталону при этом оценивается без поиска перифраз, а основой оценки близости текста эталону является разбиение слов каждой его фразы на классы по значению меры TF-IDF относительно текстов корпуса D, предварительно формируемого экспертом. Сам эталон отождествляется с набором единиц текста и их связей, необходимым и достаточным для представления единицы знаний. В качестве анализируемых текстов выступают аннотации научных статей вместе с их заголовками. Суть проблемы: для каждой фразы максимум близости эталону достигается относительно своего документа корпуса и, как следствие, требуется оценить взаимную релевантность таких документов по разным фразам анализируемого текста $\mathbb{T}$s.

Предлагаемое решение [1] основано на введении расстояний между векторами значений меры TF-IDF слов отдельной фразы $Ts_i \in \mathbb{T}$s, получаемых относительно разных документов $d_j \in D$:

$$\boldsymbol{Ts_{ij}} = \left(v_1, \ldots, v_{len(Ts_i)}\right), \tag{1}$$

где $len\,(Ts_i)$ — длина Ts_i в словах. Оценка близости отдельной фразы эталону [1, 2] строится из следующих эмпирических соображений. Во-первых, разделение на общую лексику и термины здесь должно выражаться как можно в большей степени, а слова в кластерах, формируемых по TF-IDF — распределяться более или менее равномерно. Кроме того, число получившихся кластеров должно стремиться к трём при максимуме значений TF-IDF для слов кластера наибольших значений указанной меры.

Пусть $\boldsymbol{Ts_{i,\max(D,i)}}$ — вектор вида (1) для $d_{\max(i)} \in D$, относительно которого достигнут максимум близости эталону по Ts_i. Обозначим последовательность векторов вида (1) по фразе Ts_i для документов $d_j \in D$: $d_j \neq d_{\max(i)}$, отсортированную по убыванию расстояния до $\boldsymbol{Ts_{i,\max(D,i)}}$, как $\mathbb{T}_i$. Разобьём $\mathbb{T}_i$ на кластеры $H_1, \ldots, H_{r(\mathbb{T}_i)}$, где $H_{r(\mathbb{T}_i)}$ по определению будет отвечать документам с наименьшим расстоянием до документа $d_{\max(i)}$.

Определение 1. *Классификацию слов фразы $Ts_i \in \mathbb{T}$s по значению TF-IDF, выполненную относительно некоторого $d_j \in D$, будем считать сопоставимой с аналогичной классификацией относительно $d_{\max(i)}$ при выполнении одного из двух условий:*

— $d_j \in H_{r(\mathbb{T}_i)}$

— $\exists Ts_j \in \mathbb{T}\mathrm{s}\colon Ts_j \neq Ts_i$, $d_j = d_{\max(j)}$, *при этом* $\exists\, d_k \in D\colon$ $d_k \neq d_j$, $d_k \neq d_{\max(i)}$, *причём* d_k *одновременно относится и к* $H_{r(\mathbb{T}_i)}$, *и к* $H_{r(\mathbb{T}_j)}$.

Сами d_j *и* $d_{\max(i)}$ *назовём взаимно релевантными по TF-IDF.*

Введём в рассмотрение граф, где вершины соответствуют документам, относительно которых достигнут максимум близости эталону минимум по одной фразе $Ts_i \in \mathbb{T}\mathrm{s}$, а каждое ребро соединяет вершины для пары взаимно релевантных по TF-IDF документов. Будем называть далее такой граф графом релевантности. Введение указанного графа позволяет сделать важный вывод по применению двух предложенных авторами ранее вариантов оценки близости текста смысловому эталону, предусматривающих минимум среднеквадратического отклонения (СКО) значения близости эталону по всем $Ts_i \in \mathbb{T}\mathrm{s}$. Первый подразумевает максимизацию близости эталону для заголовка, второй — по всем фразам. Допустимо полагать, что указанные оценки будут точнее для того текста, граф релевантности которого — связный. Ключевые сочетания слов из задающих смысловые образы отдельных фраз также с большей вероятностью определяют единый образ текста в случае связности его графа релевантности. Данное предположение естественно согласуется с известной гипотезой о скрытых связях, согласно которой пары слов, встречающиеся в похожих моделях, стремятся иметь близкую смысловую зависимость.

Другой вывод касается построения текстовой иерархии анализом встречаемости слов с наибольшими значениями TF-IDF в разных текстах коллекции (по степени дополнения эталона, [1]).

Утверждение 1. *При объединении графов релевантности всех текстов коллекции вышестоящий текст* $\mathbb{T}\mathrm{s}_i$ *и непосредственно связанный с ним нижестоящий текст* $\mathbb{T}\mathrm{s}_j$ *в формируемой иерархии должны иметь свои графы релевантности подграфами некоторой компоненты связности объединённого графа по коллекции.*

При прочих равных условиях при выборе вышестоящего текста для заданного $\mathbb{T}\mathrm{s}_j$ в формируемой иерархии предпочтение отдаётся тексту $\mathbb{T}\mathrm{s}_i$, который отвечает условию *Утверждения 1.*

Пусть S — последовательность текстов анализируемой коллекции; $\bigcup_S \mathbb{T}\mathrm{s}$ — объединённое множество фраз по всем текстам $\mathbb{T}\mathrm{s}$ в составе S. Тогда значимость документа $d \in D$ для формирования графа релевантности текста $\mathbb{T}\mathrm{s}$ может из геометрических соображений быть оценена [2] как

$$\mathrm{N}(d) = \frac{\big|D\big| - \min\limits_{d}\big(\big|H_{r(\mathbb{T}_i)}\colon Ts_i \in \bigcup_S \mathbb{T}\mathrm{s}\big|\big)}{\sigma\Big(\Big|H_j \in \Big\{H_1, \ldots, H_{r(\mathbb{T}_i)}\Big\}\colon Ts_i \in \bigcup_S \mathbb{T}\mathrm{s}\Big|\Big) + 1}. \tag{2}$$

Первое слагаемое в знаменателе есть СКО числа элементов кластера из полученных относительно d по разным $\mathbb{T}\mathrm{s} \in S$. Вычитаемое в числителе есть минимум числа элементов кластера наименьших расстояний относительно того же d и также по разным текстам $\mathbb{T}\mathrm{s} \in S$. Содержательно оценка (2) позволяет выделить те $d \in D$, относительно которых рассматриваемая классификация по расстоянию до них наиболее выражена. Кроме того, если документы $d \in D$, относительно которых максимум близости эталону был достигнут минимум по одной фразе, разбить на кластеры по значению оценки (2), то при выборе вышестоящего для текста $\mathbb{T}\mathrm{s}_j$ в вышеупомянутой иерархии при прочих равных условиях наименьший приоритет будет у того $\mathbb{T}\mathrm{s}_i$, у которого максимум близости эталону минимум по одной фразе достигается относительно некоторого документа кластера наименьших значений оценки (2).

Работа поддержана грантом РФФИ No. 19-01-00006.

[1] *Mikhaylov D., Emelyanov G.* Analysis of the mutual relevance of topical corpus documents in the problem of assessing the proximity of text to the semantic standard // Pattern Recognition and Image Analysis, 2021. Vol. 31(3). Pp. 588–594.

[2] *Mikhaylov D., Emelyanov G.* 2021. Ranking of documents of topical corpus according to their mutual relevance in the problem of estimating of affinity of a text to the sense standard // Journal of Physics: Conf. Series, (in press).

Coherence of semantic patterns and mutual relevancy of topical corpus documents

Mikhaylov Dmitry ⋆ mdv74@list.ru
Emelyanov Gennady Gennady.Emelyanov@novsu.ru
Russia, Veliky Novgorod, Yaroslav-the-Wise Novgorod State University

The offered work is devoted to the problem of the unity and integrity of the image for a semantic pattern (i. e. sense standard) revealed phrase by phrase for a topical text. The closeness of the text to its standard is estimated without a revelation of periphrases. The base for estimating the closeness of the text to the standard is the splitting of words of each its phrase into classes by the TF"=IDF metric value relative to the texts of a corpus D pre"=formed by an expert. The standard itself is associated with a set of text units and their relations necessary and enough to represent a knowledge unit. The analyzed texts are the abstracts of scientific articles together with their titles. The essence of the problem: for each phrase, its maximum closeness to the standard is reached concerning the individual corpus document, and, consequently, it is necessary to estimate the mutual relevance of such documents concerning different phrases of the analyzed text $\mathbb{T}$s.

The suggested solution [1] is based on entering into consideration distances between vectors of TF"=IDF values for words of a separate phrase $Ts_i \in \mathbb{T}$s relative to different corpus documents $d_j \in D$:

$$\boldsymbol{Ts_{ij}} = \left(v_1, \ldots, v_{len(Ts_i)}\right), \tag{1}$$

where $len\,(Ts_i)$ is the length of the phrase Ts_i measured in words. The estimation of the closeness of a certain phrase to its sense standard [1, 2] is based on the following empirical reasons. First, the division into general vocabulary and terms here should be expressed as much as possible, and words should be distributed more or less evenly in clusters formed by TF"=IDF. In addition, the number of resulted clusters should be close to three as possible at maximum TF"=IDF for words related to the cluster of greatest values of mentioned metric.

Let $\boldsymbol{Ts_{i,\max(D,i)}}$ be the vector of the kind (1) for $d_{\max(i)} \in D$, concerning which the maximum closeness to the standard for Ts_i is reached. Let's designate as $\mathbb{T}_i$ the sequence of vectors like (1) obtained for Ts_i concerning documents $d_j \in D$: $d_j \neq = d_{\max(i)}$. The sequence is sorted by descending the Euclidean distance to vector $\boldsymbol{Ts_{i,\max(D,i)}}$.
Let's split $\mathbb{T}_i$ into clusters $H_1, \ldots, H_{r(\mathbb{T}_i)}$, where $H_{r(\mathbb{T}_i)}$ will correspond to documents closest by distance to $d_{\max(i)}$.

Definition 1. *We'll assume, that the classification of words of the phrase $Ts_i \in \mathbb{T}$s by to the value of TF"=IDF metric concerning some document $d_j \in D$ can be considered as comparable to an analogous classification concerning $d_{\max(i)}$ when one of two conditions is met:*

— $d_j \in H_{r(\mathbb{T}_i)}$
— $\exists Ts_j \in \mathbb{T}\mathrm{s}\colon Ts_j \neq Ts_i$, $d_j = d_{\max(j)}$, *herewith* $\exists\, d_k \in D\colon$
$d_k \neq d_j$, $d_k \neq d_{\max(i)}$, *and* d_k *simultaneously relates to* $H_{r(\mathbb{T}_i)}$ *and* $H_{r(\mathbb{T}_j)}$.

Let's name these d_j *and* $d_{\max(i)}$ *as mutually relevant by TF"=IDF.*

Let's consider the graph, which vertices correspond to the documents concerning to which the maximum affinity to the standard is reached at least for one $Ts_i \in \mathbb{T}\mathrm{s}$, and each edge connects vertices for a pair of mutually TF"=IDF relevant documents. Hereinafter, such a graph we'll name a relevancy graph. The introduction of this graph allows us to make an important conclusion on the usage of two estimation variants proposed earlier for the closeness of a text to the sense standard. Both variants are equally assumed the minimum of root"=mean"=square deviation (RMSD) for value of affinity to the standard for all $Ts_i \in \mathbb{T}\mathrm{s}$. The first variant assumes the maximal closeness to the standard for the article title. The second one assumes maximizing this value for all $Ts_i \in \mathbb{T}\mathrm{s}$. It's permissible to assume that mentioned estimations will be more precise for that text, whose relevancy graph is connected. Key word combinations defining semantic images of separate phrases will also be more likely to determine a single semantic image of $\mathbb{T}\mathrm{s}$ in the case of the relevancy graph connectivity for it. This assumption naturally agrees with the well"=known hypothesis about latent relations, according to which pairs of words occur in similar models tend to have a close semantic dependence.

Another important conclusion concerns the formation of the text hierarchy using the analysis of the occurrence of words with the highest TF"=IDF values in different texts of the given collection (i. e., by the degree of the standard's complementarity).

Statement 1. *If we unite relevancy graphs of all collection texts, the parent text* $\mathbb{T}\mathrm{s}_i$ *and the directly related lower"=level text* $\mathbb{T}\mathrm{s}_j$ *in the formed hierarchy must have their relevancy graphs as subgraphs of some connectivity component of the united graph for the collection.*

All other things being equal, when choosing the higher"=level text for the given $\mathbb{T}\mathrm{s}_j$ in the formed hierarchy, preference will be given to those text $\mathbb{T}\mathrm{s}_i$, that meets the condition of *Statement 1.*

Let S be the sequence of texts of the analyzed collection; $\bigcup_S \mathbb{T}\mathrm{s}$ be the united set of phrases for all texts $\mathbb{T}\mathrm{s}$ within S. Then the significance of document $d \in D$ for the formation of the relevancy graph of text $\mathbb{T}\mathrm{s}$ can be estimated from geometrical considerations [2] as

$$\mathrm{N}(d) = \frac{\big|D\big| - \min\limits_{d}\big(\big|H_{r(\mathbb{T}_i)}\colon Ts_i \in \bigcup_S \mathbb{T}\mathrm{s}\big|\big)}{\sigma\Big(\Big|H_j \in \Big\{H_1, \ldots, H_{r(\mathbb{T}_i)}\Big\}\colon Ts_i \in \bigcup_S \mathbb{T}\mathrm{s}\Big|\Big) + 1}. \tag{2}$$

The first summand in the denominator is the RMSD of the number of elements for clusters obtained relative to d for different $\mathbb{T}\mathrm{s} \in S$. The subtrahend in the nu-

merator is the minimum number of elements of a cluster of least Euclidean distances concerning the same d for different $\mathbb{T}\mathrm{s} \in S$. Essentially, estimation (2) allows revealing those $d \in D$, concerning which the considered classification according to Euclidean distance to them is the most expressed. Besides, if documents $d \in D$, concerning which the maximum of affinity to a standard was reached at least for one phrase, be split into clusters according to the estimation (2), then in a choice of parent text for $\mathbb{T}\mathrm{s}_j$ in the abovementioned hierarchy, all other things being equal, the least priority will have a relationship with that text $\mathbb{T}\mathrm{s}_i$ for which a maximum of affinity to the standard is achieved concerning some document of the cluster of least values of estimation (2) at least for one phrase.

This research is funded by RFBR, grant 19-01-00006.

[1] *Mikhaylov D., Emelyanov G.* Analysis of the mutual relevance of topical corpus documents in the problem of assessing the proximity of text to the semantic standard // Pattern Recognition and Image Analysis, 2021. Vol. 31(3). Pp. 588–594.

[2] *Mikhaylov D., Emelyanov G.* 2021. Ranking of documents of topical corpus according to their mutual relevance in the problem of estimating of affinity of a text to the sense standard // Journal of Physics: Conf. Series, (in press).

Определение факта заимствования в текстовых документах без указания источника

Сафин Камиль Фанисович[1]⋆ kamil.safin@phystech.edu
Чехович Юрий Викторович[2] chehovich@ap-team.ru
[1]Москва, Московский физико-технический институт
[2]Москва, АО «Антиплагиат»

Можно выделить два глобальных подхода к задаче поиска заимствований в тексте: обнаружение «внешних» и «внутренних» заимствований [1]. Первый подход представляет собой поиск по внешней коллекции документов, которые могли были быть использованы для заимствования. Второй подход же, наоборот, не использует никаких внешних данных, а анализирует текст изолированно.

Корпус документов для поиска внешних заимствований, как правило, довольно большой, а значит и поиск по нему является тяжелой вычислительной задачей. Поэтому корпус документов заранее подготавливают (например, индексируют). Тем не менее, задача поиска заимствований по внешнему корпусу остается ресурсоемкой.

При поиске внутренних заимствований не используется никакой внешний корпус документов. При поиске анализируются различные стилистические, синтаксические, орфографические особенности текста.

Предлагается совместить эти два подхода для ускорения поиска текстовых заимствований. При большом потоке документов, которые необходимо проверить на наличие заимствований, система поиска по внешнему корпусу обрабатывает каждый документ и в каждом находит блоки заимствований, если такие имеются. Однако можно использовать поиск внутренних заимствований для определения факта наличия заимствований как таковых. Таким образом, можно сократить число документов для ресурсоемкой процедуры поиска заимствований по внешнему корпусу. Причем при изолированном анализе отдельно взятого документа не нужно пытаться найти конкретные блоки заимствований, эта процедура рассматривается как своеобразный показатель оригинальности документа. В случае, если общая оригинальность на низком уровне, то этот документ стоит отправить на более детальную и точную проверку.

Для определения факта заимствования анализируемый текст разбивается на сегменты. Каждый из полученных сегментов сравнивается с исходным текстом по частотным характеристикам, подобно работе [2]. При наличии достаточного числа сегментов, сильно отклоняющихся от основного текста, анализируемый текст должен быть отправлен на поиск заимствований по внешнему корпусу. В противном случае, текст в такой проверке не нуждается.

Предлагаемый метод позволяет фильтровать тексты с высоким показателем оригинальности, которые не нуждаются в дополнительной проверке. Поэтому основной метрикой при настройке параметров алгоритма является полнота среди текстов с заимствованиями.

В работе используется корпус текстов, подготовленных и размеченных в рамках конкурса PAN-2020. Корпус содержит документы на английском языке. Каждый документ может содержать от 0 до 10 вставок текста другого авторства.

[1] *Сафин К., Кузнецов М., Кузнецова М.* Определение заимствований в тексте без указания источника // Информатика и её применения, 2017.

[2] *Oberreuter G., L'Huillier G., Ríos S., Velásquez J.* Outlier-Based Approaches for Intrinsic and External Plagiarism Detection // Knowlege-Based and Intelligent Information and Engineering Systems, 2011.

Intrinsic methods for plagiarised texts detection

Safin Kamil[1]⋆ kamil.safin@phystech.edu
Chekhovich Yuriy[2] chehovich@ap-team.ru
[1]Moscow, Moscow Institute of Physics and Technology
[2]Moscow, Antiplagiat Company

There are two global approaches to the problem of searching plagiarism in the text: external and intrinsic search [1]. The first approach implies search through an external collection of documents that could have been used for text reuse. The second approach, on the contrary, does not use any external data, but analyzes the text by itself.

The corpus of documents for external search is usually quite large, which means that searching through it is a difficult computational task. Therefore, this corpus is prepared in advance (for example, indexed). Nevertheless, the task of searching plagiarism by the outer corpus remains expensive.

When searching for internal plagiarism, no external corpus of documents is used. During the search, various stylistic, syntactic, spelling features of the text are analyzed.

It is proposed to combine these two approaches to speed up the search for text plagiarism. With a large flow of documents that need to be checked, the outer corpus search system processes each document and finds plagiarised blocks in each document, if there are any. However, intrinsic search could be used to determine the fact of plagiarism. Thus, it is possible to reduce the number of documents for the expensive procedure for searching for plagiarism by the outer corpus. Moreover, in an isolated analysis of a single document, there is no need to try to find specific blocks of plagiarism, this procedure is considered as a unique indicator of the originality of the document. If the overall originality is at a low level, then this document should be sent for a more detailed and accurate check.

To determine the fact of plagiarism the analyzed text is divided into segments. Each of the obtained segments is compared with the original text in terms of frequency characteristics, similar to the work of [2]. If there is a sufficient number of segments that strongly deviate from the main text, the analyzed text should be sent to the external search for plagiarism. Otherwise, the text does not need such a check.

The proposed method allows to filter texts with a high rate of originality that do not need additional verification. Therefore, recall among the texts with plagiarised segments is the main metric during the algorithm parameters adjusting.

The work uses a corpus of texts prepared and marked up within the track of the PAN-2020 competition. The corpus contains documents in English. Each document contains from 0 to 10 inserts of other authorship text.

[1] *Safin K., Kuznetsov M., Kuznetsova M.* Methods for intrinsic plagiarism detection // Informatics and applications, 2017.

[2] *Oberreuter G., L'Huillier G., Ríos S., Velásquez J.* Outlier-Based Approaches for Intrinsic and External Plagiarism Detection // Knowlege-Based and Intelligent Information and Engineering Systems, 2011.

Применение нейронных сетей в вопросно-ответных системах на русском языке

Галеев Денис Талгатович[1]⋆ ra3wvw@mail.ru
Панищев Владимир Славиевич[1] gskunk@yandex.ru
[1]Курск, Юго-Западный государственный университет

В работе рассматривается задача поиска ответа на вопрос в тексте на русском языке. Под поиском ответа на вопрос в тексте будем подразумевать: наличие текста и вопроса к тексту, система должна выбрать в качестве ответа на вопрос - непрерывный фрагмент из данного текста либо сообщить, что ответа в тексте нет.

Основным набором данных для вопросно-ответных систем на русском языке является SberQuAD [1]. Данный датасет содержит 45328 тренировочных наборов из текста, вопроса, и ответа, 5036 валидационных наборов и 23936 проверочных наборов.

Для данной задачи широко применяются нейронные сети с архитектурой Трансформер, либо сети, которые основаны на данной архитектуре и используют только энкодер или декодер из неё. Для данных сетей данная задача является задачей классификации, в которой они пытаются найти ответы на вопросы: "Является ли данное слово из текста началом ответа на заданный вопрос? "Является ли данное слово из текста концом ответа на заданный вопрос?".

Успешность в решении поставленной задачи часто является следствием предварительного обучения эмбеддингов слов для данных сетей на больших объёмах текстов, решая различные задачи языкового моделирования. К тому же создание контекстуальных эмбеддингов слов при помощи BPE (byte-pair encoding) положительно сказывается на понимании сетями полученных текстов. Лучшими результатами среди задач обработки текста обладают следующие сети: BERT [2], RoBERTa [3], AlBERT [4], T5 [5].

В работе исследованы представленные сети. Рассмотрены специальные версии данных сетей, которые были дотренированы на больших объёмах русских текстов, а также которые были натренированы на нескольких языках. Изучаются возможные варианты модификаций данных сетей, такие как: обучение эмбеддингов, настройка гиперпараметров обучения, модификации архитектур под поставленную задачу.

[1] *Pavel E., Andrey Ch., Leonid B., Pavel B.* SberQuAD – Russian Reading Comprehension Dataset: Description and Analysis // arXiv:1912.09723, 2020.

[2] *Devlin J., Chang M., Lee K., Toutanova K.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019. Vol. 1. Pp. 4171–4186.

[3] *Liu Y., Ott M., Goyal N., Du J. Joshi M. Chen D. Levy O. Lewis M. Zettlemoyer L. Stoyanov V.* RoBERTa: A Robustly Optimized BERT Pretraining Approach // arXiv:1907.11692, 2019.

[4] *Devlin J., Chang M.-W., Lee K., Toutanova K.* Albert: A lite bert for self-supervised learning of language representations // arXiv:1909.11942, 2019.

[5] *Raffel C., Shazeer N., Roberts A., Lee K. Matena M. Zhou Y. Li W. Liu P.J.,* Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer // arXiv:1910.10683, 2020.

Application of neural networks in question answering systems in Russian

Galeev Denis[1]⋆ ra3wvw@mail.ru
Panishchev Vladimir[1] gskunk@yandex.ru
[1]Kursk, The Southwest State University

The paper considers the problem of finding an answer to a question in a text in Russian. By searching for an answer to a question in the text, we mean: the presence of a text and a question to the text, the system must choose as an answer to the question - a continuous fragment from the given text or report that there is no answer in the text.

The main dataset for question-answer systems in Russian is SberQuAD [1]. This dataset contains 45328 training sets from text, question, and answer, 5036 validation sets and 23936 test sets.

For this task, neural networks with the Transformer architecture are widely used, or networks that are based on this architecture and use only an encoder or a decoder from it. For these networks, this task is a classification task in which they try to find answers to the questions: "Is this word from the text the beginning of the answer to the question asked?", "Is this word from the text the end of the answer to the question?"

Success in solving this problem is often a consequence of the preliminary training of word embeddings for these networks on large volumes of texts, solving various problems of language modeling. In addition, the creation of contextual word embeddings using BPE (byte-pair encoding) has a positive effect on the understanding of the received texts by the networks. The following networks have the best results among word processing tasks: BERT [2], RoBERTa [3], AlBERT [4], T5 [5].

The presented networks are investigated in the work. Special versions of these networks are considered, which were trained on large volumes of Russian texts, as well as which were trained in several languages. Possible variants of modifications of these networks are studied, such as: training embeddings, tuning hyperparameters of training, modifications of architectures for the task at hand.

[1] *Pavel E., Andrey Ch., Leonid B., Pavel B.* SberQuAD – Russian Reading Comprehension Dataset: Description and Analysis // arXiv:1912.09723, 2020.

[2] *Devlin J., Chang M., Lee K., Toutanova K.* BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019. Vol. 1. Pp. 4171–4186.

[3] *Liu Y., Ott M., Goyal N., Du J. Joshi M. Chen D. Levy O. Lewis M. Zettlemoyer L. Stoyanov V.* RoBERTa: A Robustly Optimized BERT Pretraining Approach // arXiv:1907.11692, 2019.

[4] *Devlin J., Chang M.-W., Lee K., Toutanova K.* Albert: A lite bert for self-supervised learning of language representations // arXiv:1909.11942, 2019.

[5] *Raffel C., Shazeer N., Roberts A., Lee K. Matena M. Zhou Y. Li W. Liu P.J.,* Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer // arXiv:1910.10683, 2020.

Улучшение качества машинного перевода с использованием обратной модели

Скачков Николай Андреевич[1]⋆ nikolaj-skachkov@yandex.ru
Воронцов Константин Вячеславович[1] vokov@forecsys.ru
[1]Москва, ВЦ ФИЦ ИУ РАН

Задача машинного перевода является одной из важнейших задач анализа текстов естественного языка. Ручной перевод текстов делается специалистами, владеющими несколькими языками, и затрачивает существенное количество времени. Данное ограничение делает перевод больших объемов текстов сложной задачей, если не использовать системы машинного перевода.

Современные модели перевода использует нейросетевые модели, обученные на параллельных предложениях. Нейросети способны переводят не просто каждое слово на язык оригинала, но и способны учитывать контекстный смысл на уровне предложения. Данное свойство помогает нейросетям правильно улавливать смыслы исходных слов или фраз и согласовывать их переводы в соответствии с нормами языка. Именно с использованием нейросетевых моделей качество машинного перевода приблизилось к человеческому.

При этом задача машинного перевода является задачей обучения с учителем, и для обучения современных нейросетевых моделей требуются миллионы пар параллельных текстов. При таком количестве данных достаточно сложно гарантировать их высокое качество и параллельность. Более того, при сборе данных, исследователи зачастую сталкиваются с тем, что домены, в которых присутствуют параллельные тексты не покрывают всего разнообразия языка. Из-за всех этих проблем, при обучении моделей перевода можно столкнуться с появлением систематических ошибок перевода, а также с низким качеством перевода определенных тематик, плохо вписывающихся в распределение входных данных.

Одним из методов решения данной проблемы является использование обратной модели при обучении машинного перевода. Обратная модель позволяет использовать синтетические переведенные данные при обучении, получаемые переводом одноязычных непараллельных документов обратной моделью. Одноязычных документов существенно больше чем параллельных, и с их помощью можно улучшить качество на сложных входных доменах. Однако такое использование даже более сильной обратной модели не решает проблему ошибок в параллельных данных. Модель перевода всё ещё может систематически недопереводить исходный текст из-за неправильного выравнивания в данных.

Другим методом использования модели обратного перевода является реранжирование гипотез во время инференса перевода. При таком подходе, из получившихся гипотез перевода выбирается та, что имеет высокую вероятность как

с точки зрения прямой, так и с точки зрения обратной моделей:

$$\text{score}(\text{hypo}) = \log P(\text{hypo}|\text{src}) + \alpha \log P'(\text{src}|\text{hypo}),$$

где P — прямая модель перевода, а P' — обратная. Однако данный подход требует оценки вероятности каждой гипотезы обратной моделью, что вычислительно неэффективно и замедляет перевод в десятки раз.

В работе [1] было предложен метод совместного обучения прямой и обратной моделей перевода с использованием дополнительных языковых моделей исходного и целевого языков. Основной идеей было использование обратной модели в качестве функции награды при обучении прямой, а языковые модели использовались для стабилизации оптимизационного алгоритма. При всей сложности данного подхода, он не усложняет инференс модели и позволяет оценивать параллельные данные с помощью обратной модели, оставляя при этом возможность использования синтетических данных.

В данной работе описан способ более стабильного совместного обучения прямой и обратной моделей. Для этого была использована совместная функция потерь, полученная из правдоподобия для циклических переводов:

$$\mathcal{L}(\text{src}) = P(\text{dst}|\text{src}) \log P'(\text{src}|\text{dst}), \quad \text{dst} \sim P(\text{dst}|\text{src})$$

где P — прямая модель, P' — обратная. При этом обратная модель использовалась не с начала обучения, а после сходимости. Это позволило существенно повысить стабильность процесса обучения.

Эксперименты на англо-финском направлении показали прирост как при использовании синтетических данных. Это объясняется малым количеством параллельных данных для этой пары языков. При этом, совместное доучивание с обратной моделью дало дополнительный прирост качества порядка 0.5 BLEU [2]. Результаты эксперимента можно увидеть в таблице ниже.

Модель	en-fi-wmt-15, BLEU
base (with backtrans)	23.4
base+finetune	25.0
base+joint	25.5

В данной работе удалось рассмотреть различные способы улучшения качества при использования обратной модели перевода. Кроме использования синтетических данных, удалось интегрировать совместное обучение прямой модели с обратной, благодаря предложенной функции потерь. В итоге, совместное использование описанных подходов дало максимальный прирост качества. Это подтверждает гипотезу, что обратная модель дает полезную информацию в процессе обучения на параллельных текстах и может использоваться не только для генерации синтетики.

[1] *Di H, Yingce X., Tao;Q., Liwei W., Nenghai Yu, Tie-Yan L., Wei-Ying M.* Dual learning for machine translation // Proceedings of the 30th International Conference on Neural Information Processing Systems, 2016. Pp. 820–828.

[2] *Roukos P., Ward S., Zhu T.* Bleu: a Method for Automatic Evaluation of Machine Translation // Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002. Pp. 311–318.

Machine translation quality improvement using reverse translation model

Skachkov Nikolay[1]⋆ nikolaj-skachkov@yandex.ru
Vorontsov Konstantin[1] vokov@forecsys.ru
[1]Moscow, CC FRC CSC RAS

Machine translation is one of the most important tasks in NLP. Manual translation of texts is usually done by specialists who speak several languages. That makes manual translation expensive and time-consuming. This limitation makes the translation of large volumes of texts a difficult task if one doesn't use machine translation systems.

Modern translation models use neural network trained on parallel bilingual data. Neural networks are able to translate the sentences using contextual information. This feature helps neural networks to correctly capture the meanings of the source words or phrases and build the fluent translation. It is with the use of neural network models that the quality of machine translation has approached human.

Although, the task of machine translation is a supervised learning task, and millions of parallel data is required to train modern neural network models. With such a large amount of data, it is quite difficult to guarantee its high quality and exact alignment of the sentence pairs. Moreover, collected data topic distribution might not match the topic distribution of real data. All of these problems, cause the appearance of systematic translation errors, as well as poor translation quality of certain domains that do not match the distribution of input data.

To solve some of this problems reverse model is used for machine translation training. The reverse model allows the use of synthetic translated data in training, obtained by back-translating monolingual documents with the reverse model. There are significantly more monolingual documents than parallel ones and they can help to improve the machine translation quality of some complex input domains. However, this does not solve the problem of errors in parallel data. The translation model can still systematically undertranslate the source text due to incorrect alignment in the data.

Another method of using the reverse translation model is the reranking of hypotheses during the translation inference. In this approach, the resulting translation must have high probability according to both direct and reverse translation models:

$$\text{score}(\text{hypo}) = \log P(\text{hypo}|\text{src}) + \alpha \log P'(\text{src}|\text{hypo}),$$

where P is a direct translation model, and P' is the reverse model. However, this approach requires estimating the probability of each hypothesis by the reverse model. This is computationally inefficient and slows down translation by tens of times.

In [1], a method was proposed for joint training of direct and reverse translation models with using additional language models of the source and target languages.

The main idea was to use the reverse model as a reward function in direct learning, and language models were used to stabilize the optimization algorithm. Despite the complexity of this approach, it does not complicate the inference of the model and allows to score parallel data using the reverse model, while leaving the possibility of using synthetic data.

This paper describes a method for more stable joint learning of direct and reverse models. To do this, a joint loss function is obtained from the likelihood of cyclic translations:

$$\mathcal{L}(\text{src}) = P(\text{dst}|\text{src}) \log P'(\text{src}|\text{dst}), \quad \text{dst} \sim P(\text{dst}|\text{src})$$

where P is the direct model, P' is the reverse model. At the same time, the joint training was not used from the beginning of training. That significantly increases the stability of the learning process.

Experiments in the English-Finnish direction showed quality improvement with the use of synthetic data. This is due to the small amount of parallel data given for this pair of languages. At the same time, joint training with the reverse model gave an additional quality improvement of about 0.5 BLEU [2]. The results can be seen in the table below.

Model	en-fi-wmt-15, BLEU
base (with backtrans)	23.4
base+finetune	25.0
base+joint	25.5

In this paper, there were presented various ways of machine translation quality improvement via reverse translation model integration. Besides back-translation generation, it is possible to integrate joint training of the direct and reverse translation models. As a result, consecutive application of the described approaches gave the best quality improvement. This confirms the statement that the reverse models provide useful information on parallel texts quality in the training process and can be used not only for back-translation generation.

[1] *Di H, Yingce X., Tao;Q., Liwei W., Nenghai Yu, Tie-Yan L., Wei-Ying M.* Dual learning for machine translation // Proceedings of the 30th International Conference on Neural Information Processing Systems, 2016. Pp. 820–828.

[2] *Roukos P., Ward S., Zhu T.* Bleu: a Method for Automatic Evaluation of Machine Translation // Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002. Pp. 311–318.

Банк тем: сбор интерпретируемых тем с помощью множественного обучения тематических моделей и их дальнейшее использование для оценки качества тематических моделей

Алексеев Василий Антонович[1]⋆ vasiliy.alekseyev@phystech.edu
Воронцов Константин Вячеславович[1] vokov@forecsys.ru

[1]Москва, Московский физико-технический институт (национальный исследовательский университет) (МФТИ)

Вероятностное тематическое моделирование используется для выявления тематической структуры текстовой коллекции. Получая на вход набор текстов, тематическая модель должна найти скрытые в этой коллекции текстов *темы* как вероятностные распределения на множестве слов. Однако тематические модели обладают несколькими недостатками. Тематические модели *неустойчивы*, то есть итоговые темы могут зависеть, например, от инициализации модели. Также тематические модели *неполны*, то есть каждая вновь обученная на одной коллекции документов тематическая модель может выявлять новые темы. В связи с отмеченными недостатками тематических моделей стоит отметить, что идеи и гипотезы, принятые в тематическом моделировании, позволяют свести задачу поиска тем в документах к *некорректно поставленной* задаче матричного разложения, которая решается итерационным алгоритмом. При решении такой задачи может применяться регуляризация, с помощью которой можно наложить дополнительные ограничения на множество допустимых решений. Но всё равно при обучении нескольких моделей на одной и той же коллекции документов получается так, что некоторые из найденных ими тем могут быть похожи, а некоторые вообще неинтерпретируемыми, состоящими из слабо связанных друг с другом слов. Это означает, что исследование данных с помощью тематического моделирования, поиск модели, наилучшим образом описывающей данные, связан с проведением большого числа экспериментов, просмотром и оценкой качества большого числа тем.

В работе мы представляем новый способ применения тематического моделирования для исследования текстовых коллекций, в котором учитываются неустойчивость и неполнота тематических моделей. Предлагаемый метод состоит из двух шагов и может рассматриваться как способ организации эксперимента с использованием тематического моделирования. Сначала с помощью множественного обучения тематических моделей происходит отбор интерпретируемых тем, которые в совокупности образуют “Банк тем”. Далее банк тем используется для оценки качества вновь обученных моделей. Таким образом, информация о текстовой коллекции постепенно накапливается, помогая при анализе тем новой модели. Далее опишем подробнее процессы создания и применения банка тем.

Первый шаг — постепенное накопление хороших тем для данной текстовой коллекции и объединение их в банк тем. Темы отбираются в процессе множе-

ственного обучения тематических моделей. Добавляемые в банк темы должны быть хорошего качества, также темы банка должны быть различны. Качество тем оценивается по когерентности темы. Чтобы обеспечивать различность тем в процессе создания банка тем, мы предлагаем метод сравнения тем двух моделей, основанный на двухуровневой иерархической тематической модели. Темы вновь обученной модели представляют дочерний уровень в иерархической тематической модели, темы банка тем — родительский. Таким образом мы можем выяснить отношения между темами двух моделей (например, отношение "родительская тема — дочерняя тема"), что позволяет исключить шумовые темы и темы-дубликаты.

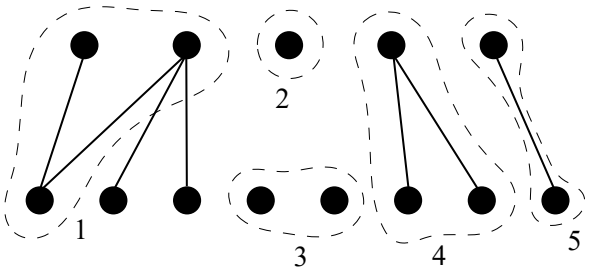


Рис. 1. Верхний уровень точек представляет родительские темы (темы из банка тем), нижний — дочерние темы (темы вновь обученной модели). Возможные отношения между темами: соединение нескольких тем (1), когда у темы нет дочерних (2), когда у темы нет родителя (3), когда тема "расщепляется" на несколько тем (4), и когда у темы только один потомок (5).

Второй шаг — применение банка тем для оценки качества тематических моделей. Качество модели оценивается путём сравнения её тем и тем, собранных заранее в банк тем, которые по построению банка тем являются хорошими и различны. Тематическая модель считается тем лучше, чем больше тем из банка тем ей удалось найти.

Для того, чтобы показать, что банк тем можно использовать для оценки качества тематических моделей, мы провели следующий эксперимент. Мы взяли несколько датасетов естественного языка: ПостНаука, Reutes, Brown, Twenty Newsgroups, AG News, Хабрахабр, Watan2004. Собрали для этих датасетов Банки тем. И далее провалидировали несколько тематических моделей на всех датасетах: PLSA; LDA; модели с декоррелирующим, разреживающим, сглаживающим регуляризаторами (APTM); модели с неслучайной инициализацией. Выяснилось, что банк тем позволяет определить модель, находящую максимальное число интерпретируемых тем среди всех рассмотренных моделей. Что доказывает возможность применения Банка тем для оценки качества тематических моделей.

Работа поддержана грантом РФФИ No. 20-07-00936.

[1] *Alekseev V., Egorov E., Vorontsov K., Goncharov A., Nurumov K., Buldybayev T.* TopicBank: Collection of coherent topics using multiple model training with their further use for topic model validation // Data & Knowledge Engineering, 2021. Vol. 135. Pp. 10–19.

TopicBank: Collection of coherent topics using multiple model training with their further use for topic model validation

Alekseev Vasiliy[1]⋆ vasiliy.alekseyev@phystech.edu
Vorontsov Konstantin[1] vokov@forecsys.ru
[1]Moscow, Moscow Institute of Physics and Technology (National Research University) (MIPT)

Probabilistic topic modeling of a text collection is a tool for unsupervised learning of the inherent thematic structure of the collection. Given only the text of documents as input, the topic model aims to reveal latent topics as probability distributions over words. However, topic models are *unstable* in the sense that topics may depend on the random initialization, and *incomplete* in the sense that each new run of the model on the same collection may discover some new topics. It should be noted that ideas and hypotheses adopted in topic modeling ultimately allow reducing the original problem of finding topics in documents to the *incorrectly posed* matrix decomposition problem, which is solved by an iterative algorithm. One of the possible ways to overcome this is, for example, to apply regularization techniques, which impose additional constraints and even lead to a better solution. Additionally, the result of the iterative algorithm depends on the initialization of the model: different initializations may lead to different resulting topics. Some resulting topics may be similar for many topic models with different initializations, some require certain initialization of a topic model, and some topics may be uninterpretable: include words from unrelated areas. Best parameters and hyper-parameters selection is also about model stability. This means that data exploration using topic modeling usually requires too many experiments for looking over many topic models and tuning their parameters in search of a model that describes the data best.

This paper presents a simpler, more understandable, and more straightforward way of exploring data using topic models which focuses not on model training but on model validation. The proposed method to deal with the instability and incompleteness of topic models consists of two steps. Overall, this can be seen as a way of conducting an experiment using topic modeling.

First, unsupervised or semi-supervised gradual collection of good topics in a "topic bank" using multiple model training. Such an approach naturally forms a deposit of good and bad examples of topics from the collection, hence the name: Topic Bank, or TopicBank. Thus, TopicBank is a kind of wrapper over topic modeling when information about the dataset accumulates gradually (see Fig. 1). We propose an algorithm for creating a topic bank for a given dataset. When collecting topics for the topic bank, we pay attention to the following: so that the topics added to the bank are of good quality, and the topics in the topic bank are different. Topic quality is assessed using topic coherence. Then, we introduce a method to compare the topics of two topic models based on the use of a two-level hierarchical topic model where each level represents topics of one particular topic model. To add topics into

the bank, we learn a child level in a hierarchical topic model, then we analyze the coherence of child subtopics and their relationships with parent bank topics in order to exclude irrelevant and duplicate subtopics instead of adding them to the bank. This allows not only to assess the closeness of topics but also to understand whether the topic of one model can be considered a parent for the topic of another model.

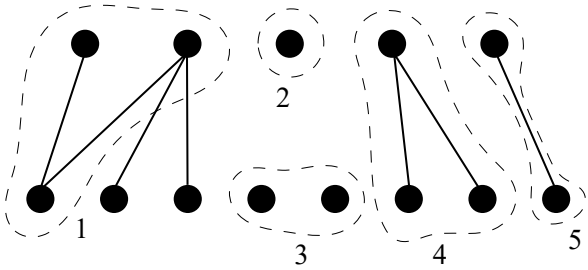


Fig. 1. The upper layer of points represents parent topics (bank topics), the lower layer represents child topics (new model topics). Possible relationships between topics: junction of several topics (1), when a topic has no child topics (2), when child topics have no parent topic (3), when one topic has several child topics and these child topics have only one topic as a parent (4), when a topic moves to the next level without splitting (5).

Second, automatic validation of new topic models using the collected topics. Then we introduce a new way to topic model evaluation by comparing the topics found by the model with the ones that were collected beforehand in a bank. The more the model finds topics similar to ones stored in the topic bank, the better the model is.

To show that the proposed method actually does help in better data exploration, we address the following question: is there such a way to train topic models so that it leads to the best topic model quality estimated using the collected topic bank. We take several natural language datasets: PostNauka, Reuters, Brown, Twenty Newsgroups, AG News, Habrahabr, Watan2004. Create a topic bank for each of the datasets. Then we train a range of topic models on each of the datasets and estimate the quality of the models using the corresponding topic banks. The models we used include: PLSA, LDA, several models with regularizers (ARTM), several models with nonrandom initialization. Hypothetically, as the considered datasets are basically similar, there should be such a model which should describe well the majority of the datasets. Experiments demonstrate the applicability of the topic bank: it does help in determining a model with more interpretable topics. Thus, the collected topic bank can be used to automatically assess the quality of newly trained topic models.

This research is funded by RFBR, grant 20-07-00936.

[1] *Alekseev V., Egorov E., Vorontsov K., Goncharov A., Nurumov K., Buldybayev T.* TopicBank: Collection of coherent topics using multiple model training with their further

use for topic model validation // Data & Knowledge Engineering, 2021. Vol. 135. Pp. 10–19.

Дата или не дата? Опыт обучения нейросети разрешению неоднозначности темпоральных выражений

Сулейманова Елена Анатольевна[1]⋆ yes2helen@gmail.com
Трофимов Игорь Владимирович[1] itrofimov@gmail.com
[1]Переславль-Залесский, ИПС им. А. К. Айламазяна РАН

В докладе представлен опыт обучения нейросетевого классификатора разрешению трех видов языковой неоднозначности — лексической омонимии, многозначности (полисемии) и типа референции — в контексте задачи извлечения темпоральной информации из текста. Извлечение темпоральной информации включает, среди прочего, обнаружение и нормализацию (стандартизованную запись значения) дат. Дата — это текстовое выражение, которое обозначает конкретный календарный день, месяц, год и т. п. В целом, обнаружение дат вполне успешно выполняется правилами (pattern matching). Однако возможны ситуации, когда требуется отличить дату от выражения, которое датой не является, например:

а) ***Сегодня*** (дата) *все обстоит так же, но журналистов меньше, потому что разрешена трансляция из зала суда.*

б) ***Сегодня*** (не дата) *этот вид деятельности не запрещен.*

в) *Специалисты в* ***среду*** (дата) *погрузили на судно 13 белух из «китовой тюрьмы».*

г) *Затем трансплантат наполовину погрузили в* ***среду*** (не дата) *культивирования.*

д) *Такие встречи проходят теперь в* ***среду*** (не дата).

е) *В* ***среду*** (не дата) *он редко свободен.*

Снятие такого рода неоднозначностей можно отнести к задачам различения значений слов (Word Sense Disambiguation, WSD). При решении таких задач хорошо себя зарекомендовали предобученные языковые модели на основе трансформеров. Такие модели довольствуются существенно меньшим объемом обучающих данных, по сравнению с традиционным машинным обучением, но высокая трудоемкость подготовки тренировочных данных остается узким местом и этих технологий.

Реализованные нами решения различаются методом получения тренировочных данных. В качестве ядра все решения используют предобученную модель RuBERT [1]. Контекстуализованные векторные представления подавались на вход классифицирующей двуслойной сети прямого распространения. Задача разрешения неоднозначности во всех трех случаях формулировалась как бинарная классификация на релевантные и нерелевантные употребления. В качестве источника обучающих примеров для всех трех задач использовался корпус предложений русского языка PaRuS (https://parus-proj.github.io/PaRuS).

Для задачи разрешения полисемии слова *сегодня* (‘день’/‘настоящее время’) обучающее множество было построено вручную и состояло из 2 тыс. предложе-

ний, половина из которых содержала релевантные вхождения целевого слова и половина — нерелевантные. Тестовое множество было сформировано как случайная выборка из обучающего (по 100 примеров на класс). На тестовом множестве классификатор достиг точности (accuracy) $93.2 \pm 1.5\,\%$ (здесь и далее приводится среднее значение в 10 экспериментах и стандартное отклонение).

Для снятия лексической омонимии слова *среда* ('день недели'/'окружение, субстанция') обучающее множество строилось автоматически по следующему принципу: положительный (отрицательный) обучающий пример должен содержать целевой термин, который с высокой вероятностью употреблен в релевантном (нерелевантном) значении. Для извлечения из корпуса обучающих примеров (10 тыс.) использовались несложные паттерны. Тестовое множество было составлено вручную из числа предложений, в которых не было отмечено совпадений с паттернами (по 100 положительных и отрицательных примеров). Оценка классификатора на тестовом множестве дала точность $99.1 \pm 0.6\,\%$.

Третья задача состояла в том, чтобы распознать, когда целевой термин соотносится с конкретной календарной единицей (пример «в» выше), а когда он употреблен не конкретно-референтно («д», «е»). Целевые термины — названия дней недели, месяцев и времен года в единственном числе. Эта задача отличается от двух других тем, что разница между релевантным и нерелевантным употреблением термина лежит не в плоскости лексического значения, а имеет другую — референциальную — природу. Обучающее множество для этой задачи мы формировали из дифференцирующих контекстов. Такой контекст не обязательно содержит сам целевой термин, но предположительно обладает признаками, позволяющими отличить релевантное употребление от нерелевантного. Положительный (отрицательный) пример — это такой контекст, в котором целевой термин — если бы он там встретился — с большой вероятностью был бы употреблен релевантно (нерелевантно). Большую часть положительных и отрицательных обучающих примеров составили предложения, содержащие референциально однозначные (маркированные) темпоральные выражения (*ближайший вторник*, *прошлый апрель* для релевантных контекстов, *каждую пятницу*, множественное число для нерелевантных). Чтобы увеличить число контекстов с признаком многократности, в обучающее множество были включены примеры со словами *ежедневно*, *ежемесячно*, *ежеквартально* и *ежегодно*. Сложнее всего было набрать надежные контексты для распознавания нерелевантных случаев с родовой, атрибутивной референцией. Примеры такого рода извлекались на основе паттернов. Всего для задачи разрешения референциальной неоднозначности было получено более 86 тыс. положительных и более 38 тыс. отрицательных обучающих примеров.

Для оценки классификатора были созданы тестовые множества по видам выражений. Тестовые множества создавались сбалансированными и репрезентативными. Тестовые примеры набирались не из корпуса (всего около 500 примеров) и не содержали совпадений с паттернами. Результаты оценки: названия

дней недели — 84.3 ± 1.9 %, названия месяцев — 84.5 ± 1.3 %, названия времён года — 86.7 ± 3.3 %.

В целом, результаты эксперимента обнадеживают. Самые низкие результаты получены для весьма нетривиальной и трудноформализуемой задачи, при том что на дообучение нейросети были затрачены довольно скромные усилия и возможности улучшения качества обучающих данных не исчерпаны.

Работа поддержана грантом РФФИ No. 19-07-00991.

[1] *Kuratov Yu., Arkhipov M.* Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language // arxiv.org/pdf/1905.07213, 2019.

A date or not a date? Word sense disambiguation for temporal expression recognition

Suleymanova Elena[1]⋆ yes2helen@gmail.com
Trofimov Igor[1] itrofimov@gmail.com
[1]Pereslavl-Zalessky, Program Systems Institute, Russian Academy of Sciences

The paper presents the author's experience in training a neural network classifier to handle three kinds of language ambiguity — lexical homonymy, polysemy and type of reference — as part of the task of temporal information extraction. Extracting temporal information includes, among other things, recognition and normalization of dates. A date is an expression referencing a specific calendar point like a day, a month, a year, etc. In most cases, pattern matching suffices to recognize a date in a text. The problem is that sometimes an expression that looks like a date may in fact be something different. Here are a few Russian examples to illustrate the point:

a) ***Сегодня*** (a date) *все обстоит так же, но журналистов меньше, потому что разрешена трансляция из зала суда.* Today everithing is the same, but there are fewer journalists because broadcasting is permitted in the courtroom.

b) ***Сегодня*** (not a date) *этот вид деятельности не запрещен.* Today this type of activity is not prohibited.

c) Специалисты в ***среду*** (a date) *погрузили на судно 13 белух из «китовой тюрьмы».* On Wednesday, wildlife experts loaded 13 belugas from the "whale prison" onto the ship.

d) *Затем трансплантат наполовину погрузили в* ***среду*** (a date) *культивирования.* Then the graft was half immersed in culture medium.

e) *Такие встречи проходят теперь в* ***среду*** (not a date). Such meetings are now held on Wednesday

f) *В* ***среду*** (not a date) *он редко свободен.* He is seldom free on Wednesday.

Resolving ambiguities of this kind falls within the scope of word sense disambiguation (WSD). Transformer-based pretrained language models have proven to perform well on WSD tasks. Such models need significantly smaller amount of training data compared to traditional machine learning, but the high cost of training data generation remains a bottleneck of these technologies.

We applied a different approach to training data preparation, for each of the disambiguation tasks. All of the three solutions use the RuBERT [1] pretrained model as the kernel component. Contextualized embeddings were fed into a two-layer fully connected feed-forward network. The disambiguation task was set as a binary classification into relevant and irrelevant cases. As a source of training data we used PaRuS — a corpus of Russian sentences (https://parus-proj.github.io/PaRuS).

Polysemy resolution: *сегодня* (today: 'the current day' or 'now'). The training set (2,000 sentences) was compiled by hand. The test set was formed as a random subset of the training data (100 sentences per class). For the test set, average accuracy over 10 runs was 93.2 % (with standard deviation 1.5 %).

Lexical homonymy resolution: *среда* ('Wednesday' or 'environment, medium'). As positive (negative) training examples, we accepted those sentences that were very likely to contain the target term in relevant (irrelevant) sense. To collect such examples (a total of 10,000) we searched the corpus for some simple patterns that we deemed indicative of either sense. The test set was compliled by hand out of the sentences that did not match any of the patterns (100 positive and 100 negative examples). An average accuracy of 99.1 ± 0.6 % was obtained.

The third disambiguation task had to deal with distinctions of a different nature. As opposed to the other two tasks, here the difference between relevant and irrelevant uses has nothing to do with the lexical meaning of the target term. The goal was to distinguish the cases when a name of day of week, month or season is used to refer to a particular calendar point (like *Wednesday* in example "c" above) from the cases of non-specific reference ("e", "f"). To train the classifier to make such distinctions, we used a set of differentiating contexts. A differentiating context is an example that need not contain a target term itself, but presumably has some features that would be helpful for disambiguating between a 'date' and 'not-a-date' reading. In order to identify differentiating contexts, we retrieved sentences with referentially unambiguous temporal expressions (e. g. *coming Tuesday*, *last April* for positive examples; *each Tuesday*, plural forms for negative contexts). To augment the number of irrelevant recurrence contexts, we added sentences with the words *daily*, *quarterly*, *monthly*, *yearly*. To collect examples of generic (non set-referring) occurrences, we relied on a number of usage patterns. The training data for the task of type-of-reference disambiguation consisted of more than 86,000 positive and 38,000 negative examples.

The test sets for all terms were made balanced and as representative as possible. Test examples were collected from other sources than the corpus. We avoided including any matches with the training patterns. Evaluation results: days of week — 84.3 ± 1.9 %, names of months — 84.5 ± 1.3 %, seasons — 86.7 ± 3.3 %.

The results of the experiments are rather encouraging. The lowest accuracy was obtained for a non-trivial, hard to formalize task, with modest effort spent on training data preparation, so there is still room for training data quality improvement.

This research is funded by RFBR, grant 19-07-00991.

[1] *Kuratov Yu., Arkhipov M.* Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language // arxiv.org/pdf/1905.07213, 2019.

Когнитивно-подобный документальный информационный поиск: концепция и технологии

Максимов Николай Вениаминович[1] ⋆ nv-maks@yandex.ru

[1]Москва, Национальный исследовательский ядерный университет МИФИ

В основе когнитивного поиска лежит представление познания как процесса построения модели действительности (т.е. знаний), причем для одного и того же объекта, процесса или явления в зависимости от аксиоматики и т.п. может быть построено несколько моделей. Существенно, что в этом процессе изменяются не только знания, но и инструменты познания: понятийная система, схемы познания, формы представления знаний и т.п.

В [1] на концептуальном уровне когнитивный информационный поиск, исходя из семиотической природы знания, определяется как процесс построения образа решения задачи пользователя - его онтологического представления в виде системы трех взаимосвязанных, относительно независимо развивающихся систем: функциональной, понятийной и терминологической.

На даталогическом уровне когнитивный информационный поиск представляется как построение на множестве разрозненных хаотизированных фактов (триплетов, выделяемых средствами традиционных информационно-поисковых систем) пути от исходного факта к факту-результату на мультимета-гиперграфе, представляющем знания как семиотическую систему (объекты – понятийный базис - текст).

Автоматизированная система когнитивного информационного поиска ориентирована на информационную поддержку сложных процессов синтеза знания и основывается на полнотекстовом глубинном семантическом индексировании научных и технических текстов, для которых характерно наличие новизны. Система предоставляет пользователю информацию о возможных (потенциальная новизна) связях между сущностями, а также соответствующие фрагменты текста.

Это принципиально отличается от концепции классического поиска, предполагающей формирование конечной "цельной" выдачи в ответ на семантически завершенный запрос, а процесс поиска - это итеративная последовательность согласования (обучения) информационных представлений предметной области в виде множеств документов и лексики, генерируемых человеком и системой.

Разработанная система [2], имеющая структуры/процессы, подобные "механике", познания (такие как восприятие, понимание, идентификация, абстрагирование, обобщение), помимо возможностей классического информационного поиска, включает функции визуализации и преобразования (объединения, пересечения, проекции, семантического масштабирования) графов онтологии документа(ов). Поиск на графах сводится к операциям поиска цепочки фактов и поиска окрестности факта. Для обеспечения адекватности форм визуализации используются соответствующие разным когнитивным состояниям (фоку-

су внимания) пользователя метафоры представления графа: гравитационная и функциональная модель, упорядочение путей по значимости и вершин согласно изложению материала.

Здесь фактически реализуется технологическая и семантическая интеграция основной и информационной деятельностей в режиме интерактивного "квантования", когда запросы будут относиться к минимальным, но значимым для принятия решения смысловым "блокам", а выдачи (фрагменты документов, как блоки, полезные при решении задачи) будут предопределять последующие запросы.

Такой процесс поиска, соответствуя направлению процесса познания, связывает отдельные фрагменты находимых текстов, формируя некоторую «цитатную» основу решения задачи пользователя. Это отвечает положению, что эволюция развивающейся системы определяется не сколько ее прошлым, сколько ее будущим. Структурой, организующей и отражающей эту эволюцию, является когнитивный рубрикатор (КР) - иерархическая классификационная структура пользовательского представления предметной области, отражающая текущую степень полноты знания. Узлы иерархии образуют элементы состоявшегося знания (документы, термины, фрагменты онтологий, запросы и т.д.), соответствующие теме. При этом, анализируя различия/сходства элементов КР, можно делать выводы о наличии пробелов, дисбалансов и противоречий в имеющихся знаниях. Кроме того, поскольку такая структура соединяет интенсиональное и экстенсиональное начала познания и вариантно представляет направление (отражает видение отдельных субъектов/проектов в распределенной сетевой среде), это позволит "встроить" развитие понятийно-терминологической составляющей в процесс познания.

Для семантического индексирования и графовых операций используется технология автоматического построения онтологий по полным текстам. Формируемый ориентированный граф онтологии документа, содержит вершины, соответствующие всем сущностям, представленным в тексте (и в соответствии с их расположением), и дуги, соответствующие типизированным отношениям между ними. Типизация основывается на таксономиях сущностей и отношений, что позволяет соотносить сходные по смыслу понятия и связи, которые в текстах в следствие свойств естественного языка могут быть представлены различающимися вербальными конструкциями.

В целом это позволяет повысить эффективность интерактивного процесса построения нового знания за счет совмещения процессов поиска, анализа и синтеза информации. Это достигается сочетанием кибернетического и синергетического подходов, в том числе (1) использованием отрицательной обратной связи для обеспечения сходимости процесса уточнения лексики запроса и итеративного процесса, обеспечивающего полноту отбора, и (2) использованием положительной обратной связи для развития запроса (расширения лексики запроса) и нахождения новых композиций фактов и понятий.

Работа выполнена при поддержке Министерства науки и высшего образования РФ (проект государственного задания No. 0723-2020-0036)

[1] *Maksimov N.*, *Golitsina O.*, *Monankov K.*, *Gavrilkina A.* Methods of visual graph-analytical presentation and retrieval of scientific and technical texts // Scientific Visualization, 2021, Vol. 13(1). Pp. 138–161.

[2] *Максимов Н. В.*, *Голицына О. Л.*, *Монанков К. В.*, *Гаврилкина А. С.* Документальная информационно-аналитическая система xIRBIS (редакция 6.0): программа для ЭВМ // Свидетельство о гос. регистрации №2020661683 от 29.09.2020.

Cognitive-like documentary information search: concept and technologies

Maksimov Nikolay[1] ⋆ nv-maks@yandex.ru

[1]Moscow, National Research Nuclear University MEPhI

Cognitive search is based on cognition representation as a reality model (i.e. knowledge) constructing process, and for the same object, process or phenomenon, several models can be built depending on an axiomatic etc. It is essential that in this process not only knowledge changes, but also a cognition tools: conceptual system, cognition schemes, knowledge representation forms etc.

In [1] cognitive information retrieval at a conceptual level, based on knowledge semiotic nature, is defined as a constructing process of a user's problem solution image - its ontological representation in form of a system of three interconnected, relatively independently developing systems: functional, conceptual and terminological.

At datalogical level cognitive information retrieval is represented as path construction, on a scattered chaotic facts set (triplets, extracted by means of traditional information retrieval systems), from an initial fact to a fact-result on a multi-meta-hypergraph, representing knowledge as a semiotic system (*objects - conceptual basis - text*).

An automated cognitive information retrieval system is focused on information support for complex knowledge synthesis processes and based on full-text deep semantic indexing of scientific and technical texts that are characterized by novelty. System provides to user information about *possible* (potential novelty) relationships between entities, as well as corresponding text fragments.

This is fundamentally different from the classical information retrieval concept, which presupposes a final "integral" search result formation in response to a semantically complete query, and a search process - is matching (training) iterative sequence of subject area information representations, in documents sets form, and vocabulary, generated by a person and a system.

The developed system [2], which has structures / processes similar to a cognition "mechanics" (such as perception, understanding, identification, abstraction, generalization), in addition to the classical information retrieval capabilities, includes functions of document(s) ontology graphs visualization and transformation (unification, intersection, projection, semantic scaling). Search on graphs is reduced to facts chain searching operations and searching for fact neighborhood. To ensure the visualization forms adequacy, graph representation metaphors are used, corresponding to different user cognitive states (attention focus): gravitational and functional model, paths ordering by importance and vertices according to content statement.

Here, in fact, technological and semantic integration of main and informational activities is implemented in interactive "quantization" mode, when requests will relate to a minimal, but significant for decision making semantic "blocks", and

results (document fragments, as blocks, useful in problem solving) predetermine subsequent requests.

Such search process, corresponding to cognition process direction, connects separate fragments of found texts, forming a certain "quotation" basis for user's problem solution. This corresponds to proposition that evolution of a developing system is mainly determined not by its past, but by the future. A structure that organizes and reflects this evolution is a cognitive rubricator (CR) - a subject area user presentation hierarchical classification structure, reflecting knowledge completeness current level. Hierarchy nodes form an established knowledge elements (documents, terms, ontologies fragments, queries, etc.) corresponding to the topic. At the same time, analyzing the CR elements differences / similarities, one can draw conclusions about the presence of gaps, imbalances and contradictions in available knowledge. In addition, since such a structure connects an intensional and extensional cognition beginnings and variantly represents a direction (reflects an individual subjects / projects vision in a distributed network environment), this will make it possible to "embed" a conceptual and terminological component development into cognition process.

For semantic indexing and graph operations, an ontologies automatic construction technology on full texts is used. Generated directed graph of document ontology contains nodes corresponding to all entities presented in text (and in accordance with their position), and arcs corresponding to typed relationships between them. Typification is based on entities and relations taxonomies, which allows one to correlate concepts and relationships that are similar in meaning, which in natural language texts can be represented by different verbal constructions.

In general, this makes it possible to increase the building new knowledge interactive process efficiency by combining processes of searching, analyzing and synthesizing information. This is achieved through a combination of cybernetic and synergistic approaches, including (1) using negative feedback to ensure query vocabulary refinement process convergence and an iterative process that ensures selection completeness, and (2) using positive feedback to develop a query (expanding query vocabulary), and finding of facts and concepts new compositions.

This research is funded by the Ministry of Science and Higher Education of the Russian Federation (state assignment project No. 0723-2020-0036).

[1] *Maksimov N.*, *Golitsina O.*, *Monankov K.*, *Gavrilkina A.* Methods of visual graph-analytical presentation and retrieval of scientific and technical texts // Scientific Visualization, 2021, Vol. 13(1). Pp. 138–161.

[2] *Maksimov N.*, *Golitsina O.*, *Monankov K.*, *Gavrilkina A.* Documentary information and analytical system xIRBIS (revision 6.0): computer program // Certificate of state. registration No. 2020661683 dated 09/29/2020.

Многозадачное обучение в задаче рубрикации научных документов

Шевченко Олег Владимирович[1]⋆ shevchenko@ap-team.ru
Гращенков Кирилл Владимирович[1] grashchenkov@ap-team.ru
Чащин Артем Валерьевич[1] chashchin@ap-team.ru
Грабовой Андрей Валериевич[1,2] grabovoy@ap-team.ru
[1]Москва, АО «Антиплагиат»
[2]Москва, Московский физико-технический институт (национальный исследовательский университет)

Работа посвящена решению задачи рубрикации научных публикаций, тезисов, диссертаций, отчетов и других научноемких текстов. Рубрикация традиционно является задачей многоклассовой классификации текстов, где в качестве класса рассматривается рубрика в используемом классификаторе. В настоящей работе описано применение подхода многозадачного обучения [1] к задаче рубрикации научных документов. Этап принятия решения по различным рубрикаторам разделяется на несколько задач классификации. Каждому рубрикатору соответствует своя задача классификации в рамках многозадачного подхода. В качестве рубрикаторов научных документов могут использоваться, например, рубрики разного уровня Государственного рубрикатора научно-технической информации (ГРНТИ)[1] или классификатора Организации экономического сотрудничества и развития (Organisation for Economic Co-operation and Development, OECD)[2]. В качестве признакового описания документов используются предобученные тематические вектора [2], при этом тематический вектор документа характеризует вероятность принадлежности этого документа каждой теме.

Многозадачное обучение разделяется на *жесткое* или *мягкое* разделение параметров скрытых слоев. При мягком разделении параметров каждая задача имеет свою собственную модель со своими параметрами, схожесть которых настраивается путем регуляризации. В жестком разделении предполагается, что параметры первых слоев нейросетевой модели обучаются совместно для разных задач в единой процедуре. Это позволяет построить общее скрытое представление объектов для связанных задачах. Остальные параметры обучаются, аппроксимируя выборку для каждой задачи отдельно. Жесткое совместное использование параметров значительно снижает риск переобучения [3]. В настоящей работе используется только жесткое разделение, таким образом структура модели выглядит, как нейронная сеть прямого распространения с использованием совместных скрытых слоев между всеми задачами и нескольких выходных слоев для каждой отдельной задачи в рамках многозадачного подхода. В данной работе показано, что использование многозадачного обучения с жестким

[1]https://grnti.ru
[2]https://www.oecd.org

	ГРНТИ 1 уровня	ГРНТИ 2 уровня	OECD 1 уровня	OECD 2 уровня
Кол-во классов	69	833	6	38
Кол-во документов	281842	122463	122463	122463

Таблица 1. Распределение документов по разным рубрикаторам

разделением параметров скрытых слоев помогает поднять итоговое качество для каждой из задач по отдельности.

Для анализа качества предложенного подхода был проведен вычислительный эксперимент, целью которого являлась проверка значения многозадачного обучения для задачи рубрикации научных документов. В качестве данных использовалась выборка открытых научных статей из научной электронной библиотеки eLIBRARY.ru[3], состоящая из 280 тыс. документов. Для каждого документа частично доступна информация по рубрикам, которая представлена в таблице 1. Важно, что для каждого документа доступна информация не о всех кодах одновременно, а только для некоторого подмножества. Были рассмотрены четыре модели, решающие каждую задачу по отдельности без использования рассматриваемого подхода. Также дополнительно была рассмотрена модель, которая включает в себя все уровни и рубрикаторы, которая обучается с использованием метода многозадачного обучения. Качество оценивалось по метрикам точности и полноты. В рамках данного вычислительного эксперимента было получено, что при использовании многозадачного обучения качество общей модели повышается.

Исследование выполнено в рамках проекта государственной поддержки ведущих компаний No.1 / 549/2020 от 19.06.2020 г.

[1] *Caruana R.* Multitask learning: A knowledge-based source of inductive bias. // Proceedings of the Tenth International Conference on Machine Learning, 1993.

[2] *Vorontsov K., Frei O., Apishev M., Romov P., Dudarenko M.* BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections // Analysis of Images, Social Networks and Texts, 2015.

[3] *Baxter J.* A Bayesian/Information Theoretic Model of Learning to Learn via Multiple Task Sampling. // Machine Learning, 1997.

[3] https://www.elibrary.ru

Multi-task Learning in the Problem of Rubrication of Scientific Documents

Shevchenko Oleg[1]⋆ shevchenko@ap-team.ru
Grashchenkov Kirill[1] grashchenkov@ap-team.ru
Chashchin Artem[1] chashchin@ap-team.ru
Grabovoy Andrey[1,2] grabovoy@ap-team.ru
[1]Moscow, Russia, Antiplagiat Company
[2]Moscow, Moscow Institute of Physics and Technology

The paper is devoted to solving the problem of rubricating scientific publications, theses, dissertations, reports and other scientifically intensive texts. Rubrication is traditionally posed as a problem of multiclass classification of texts, where a rubric in classifier used is considered as a class. In this paper, we describes the application of the multi-task learning approach [1] to the problem of rubricating scientific documents. The decision-making stage for various rubricators is divided into several classification tasks. Each rubricator has its classification task within the multi-task approach. As rubricators of scientific documents can be used, for example, rubrics of different levels of the State Rubricator of Scientific and Technical Information (SRSTI)[1] or the classifier of the Organization for Economic Cooperation and Development (OECD)[2]. Pretrained thematic vectors are used as a feature description of documents [2], while the document's thematic vector describes the probability that the document belongs to a particular topic.

Multi-task learning is divided into the *hard* or the *soft* separation of hidden layer parameters. In soft parameter sharing each task has its own model with its own parameters, the similarity of which is adjusted by regularization. In hard sharing, it is assumed that the parameters of the first layers of the neural network model are trained together for different tasks in a single procedure. This provides an opportunity to build a general latent representation of objects for related tasks. The rest of the parameters are trained by approximating the sample for each task separately. Hard parameter sharing significantly mitigates the risk of overfitting [3]. In this approach, only hard parameter sharing is considered, so the structure of the model looks like a feedforward neural network using joint hidden layers between all tasks and several output layers for each task in a multi-task approach. This paper shows that the use of multi-task training with the hard parameter sharing of hidden layers helps to raise the final quality for each of the tasks separately.

To analyze the quality of the proposed approach, a computational experiment was carried out, the purpose of which was to test the significance of multi-task learning for the task of rubricating scientific documents. A sample of free scientific articles from the scientific electronic library eLIBRARY.ru[3], consisting of 280

[1]https://grnti.ru
[2]https://www.oecd.org
[3]https://www.elibrary.ru

	SRSTI 1 level	SRSTI 2 level	OECD 1 level	OECD 2 level
number of classes	69	833	6	38
number of documents	281842	122463	122463	122463

Table 1. Distribution of documents by different rubricators

thousand documents, was used as data. For each document, information on rubrics is partially available, which is presented in the table 1. It is worth mentioning that for each document, information is available not about all codes simultaneously, but only for a subset. Four models were considered that solve each problem separately without using the considered approach. We also additionally considered a model that includes all levels and rubricators, which is trained using the multi-task learning method. The quality was assessed by the metrics of precision and recall. Within the framework of this computational experiment, it was found that when using multi-task learning, the quality of the general model increases.

This research has done within the project of governmental support of leading companies No.1 / 549/2020 dated 19.06.2020

[1] *Caruana R.* Multitask learning: A knowledge-based source of inductive bias. // Proceedings of the Tenth International Conference on Machine Learning, 1993.

[2] *Vorontsov K., Frei O., Apishev M., Romov P., Dudarenko M.* BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections // Analysis of Images, Social Networks and Texts, 2015.

[3] *Baxter J.* A Bayesian/Information Theoretic Model of Learning to Learn via Multiple Task Sampling. // Machine Learning, 1997.

Полуавтоматическая суммаризация тематических подборок научных публикаций: задачи и подходы

Крыжановская Светлана Юрьевна[1]⋆ skryzhanovskaya@yandex.ru
Власов Андрей Валерьевич[2] andrey.v.vlasov@phystech.edu
Еремеев Максим Алексеевич[3] eremeev@nyu.edu
Воронцов Константин Вячеславович[1,2] voron@forecsys.ru
[1]Москва, МГУ им. М.В. Ломоносова
[2]Москва, МФТИ
[3]Нью-Йорк, Нью-Йоркский Университет

Целью автоматической суммаризации обычно является синтез краткого изложения подборки текстовых документов для быстрого понимания ключевых идей. Мы рассматриваем полуавтоматическую суммаризацию подборок научных публикаций, которая имеет другую цель — помочь пользователю реализовать свой авторский замысел. Обзоры, написанные по одной и той же подборке разными авторами и/или для разных целей (диссертация, отчёт, учебник) могут различаться существенно. В таких случаях пользователю нужен не готовый текст, а рекомендательный сервис, выполняющий рутинные операции информационного поиска, например, подбирающий релевантные фразы. Такие системы называются *автоматизированной авторской суммаризацией текстов* (machine aided human summarization, MAHS).

В данной работе предлагается процесс MAHS, основанный на решении серии задач машинного обучения [1].

Задача 1: *построение сценария обзора.* Дана подборка текстов, требуется ранжировать их в том порядке, в котором они будут упоминаться в обзоре. Обучающая выборка может быть составлена автоматически по большой коллекции научных статей. Каждая статья порождает обучающий объект, в котором роль подборки выполняет список литературы, а обучающим ранжированием является последовательность ссылок в обзорной части статьи. Информативными признаками являются год публикации, её цитируемость, цитируемость её авторов, семантическая близость публикации к обзору, к его локальному контексту, и т. д.

После того, как сценарий синтезирован и скорректирован пользователем, начинается собственно процесс MAHS. Для каждого документа из подборки (в порядке, определённом сценарием) система предлагает пользователю ранжированный список фраз-подсказок, из которых он может выбирать фразы для продолжения своего обзора. Существует множество разумных способов отбора и ранжирования фраз. Они реализуются в системе как функции ранжирования, называемые *суфлёрами*. В пользовательском интерфейсе суфлёр — это кнопка, нажатие которой перестраивает ранжированный список фраз. Например, *суф-*

лёр аннотации — это простейшая необучаемая функция, которая выдаёт фразы из аннотации статьи в их исходном порядке.

Задача 2: *обучение экстрактивного суфлёра.* Классические методы экстрактивной суммаризации основаны на выделении наиболее важных предложений в документе. Важность может определяться по-разному, в том числе относительно аспектов «актуальности», «новизны», «методов», «выводов», «преимуществ», «недостатков», и т. д. Для обучения суфлёра по каждому из аспектов строится обучающая выборка документов, в каждом из которых разметчики выделяют фразы, релевантные аспекту.

Задача 3: *обучение цитирующего суфлёра.* Данная функция ранжирует фразы о статье, сопровождающие ссылку на неё. Задача выделения упоминаний требует размеченной обучающей выборки вида «ссылка, упоминающий фрагмент». Альтернативный подход заключается в том, чтобы выбрать из текста цитируемой статьи фразы, семантически близкие к локальным контекстам ссылок в упоминающих статьях. При этом точное определение границ упоминающих фрагментов уже не требуется, а цитирующий суфлёр становится разновидностью экстрактивного.

Задача 4: *оценивание качества системы суфлёров.* Валидационная выборка составляется автоматически по большой коллекции статей, аналогично первой задаче. В роли подборки выступает список литературы, в роли «идеального» обзора — объединение обзорных разделов статьи. Для каждого предложения обзора подбирается суфлёр, дающий самую близкую фразу среди первых k позиций списка поисковой выдачи. Оценкой качества является средняя позиция наиболее релевантных фраз в списках выдачи суфлёров. Заодно оценивается полезность каждого суфлёра как число его фраз, которые вошли в обзор.

В эксперименте для построения сценария обзора выделялись обзорные части статей из коллекции S2ORC [3], затем обучалась модель ранжирования на основе бустинга. Коэффициент корреляции Кенделла полученной модели $\tau = 0.48$ (при ранжировании по году публикации $\tau = 0.1$). Для обучения и оценивания цитирующего суфлёра использовалась коллекция CL-SciSumm 2018 [2]. При автоматическом составлении сценария наши результаты сопоставимы с наилучшими результатами на CL-SciSumm 2018. При полуавтоматическом составлении сценария (выбор из первых трёх позиций) удаётся добиться повышения качества по метрике ROUGE до 7 процентов (Таб. 1).

В дальнейшем планируется оценивать качество ранжирования каждого суфлёра по логам поисково-рекомендательной системы SciSearch.ai как среднюю позицию тех его фраз, которые пользователи отбирали для обзора.

В заключении заметим, что предлагаемый способ полуавтоматической суммаризации может также рассматриваться как способ *нелинейного чтения*, когда перед пользователем стоит задача не только разобраться в мало знакомой для него области по обширной тематической коллекции, но и одновременно произвести информационный продукт в виде обзора.

	Abstract	Community summary	Human summary
top-1 MLP	0.30	0.33	0.22
top-1 boosting	0.25	0.32	0.20
top-3 MLP	0.30	**0.40**	0.23
top-3 boosting	0.29	0.36	**0.24**
best 2018	**0.33**	0.25	0.21

Таблица 1. $ROUGE_2$ цитирующего суфлёра на CL-SciSumm 2018.

Работа поддержана грантом РФФИ No. 20-07-00936.

[1] *Власов А. В.* Методы полуавтоматической суммаризации подборок научных статей // Магистерская диссертация, МФТИ, 2020.

[2] *Jaidka K., Yasunga M., Chandrasekaran M., Radev D., Kan M.* The CL-SciSumm Shared Task 2018: Results and Key Insights, 2018.

[3] *Lo K., Wang L., Neumann M., Kinney R., Weld D.* S2ORC: The Semantic Scholar Open Research Corpus, 2020.

Machine Aided Human Summarization of scientific articles: tasks and approaches

Kryzhanovskaya Svetlana[1]⋆ skryzhanovskaya@yandex.ru
Vlasov Andrey[2] andrey.v.vlasov@phystech.edu
Eremeev Maksim[3] eremeev@nyu.edu
Vorontsov Konstantin[1,2] voron@forecsys.ru
[1]Moscow, Lomonosov Moscow State University
[2]Moscow, Moscow Institute of Physics and Technology
[2]New York, New York University

The goal of automatic summarization is usually to synthesize a concise analogue of a set of documents for quickly understanding of key ideas. We are considering a semi-automatic summarization of a collection of scientific publications with another purpose: to help the user realize his authorial intent. Reviews written on the same selection by different authors and/or for different purposes (thesis, report, textbook) can differ significantly. In such cases, the user does not need a ready-made text, but rather a recommender service which performs routine information retrieval operations, such as selecting relevant phrases. Such systems are called *machine aided human summarization (MAHS)*.

This paper proposes a MAHS system based on a series of machine learning tasks [1].

Task 1: *generating review scenario.* A collection of texts is given, and we need to rank them in the order in which they will be mentioned in the review. A training sample can be automatically generated from a large collection of scientific articles. Each article generates a training object in which the list of references acts as a collection, and the learning ranking is the sequence of references to these publications in the review part of the article. Informative features are the year of the publication, its citation, the citation of its authors, the semantic proximity of the publication to the review, to its local context, etc.

After the scenario is generated and adjusted by the user, the MAHS process begins. For each document in the collection (in the order defined by the scenario), the system offers the user a ranked list of clue phrases from which to select phrases to continue the review. There are many reasonable ways of selecting and ranking phrases, which are implemented in the system as ranking functions, called *prompters*. In the user interface, a prompter is a button which, when clicked, rearranges the ranked list of phrases. For example, a prompter is a simple, unlearnable function which rearranges the phrases from the abstract in their original order.

Task 2: *training an extractive prompter.* Classical extractive summarization methods are based on extracting the most important sentences in a source document. Importance can be defined in various ways with respect to aspects of "relevance", "newness", "approach", "data", "experiments", "results", "conclusions",

"advantages", "faults", etc. To train a prompter for each aspect, a training sample of documents is constructed, in which assessors select phrases relevant to the aspect.

Task 3: *training a citation prompter.* This function ranks phrases (or larger pieces of text) that refer to a given article. Peeking at how other authors mention a given article when it is cited has long been a ubiquitous practice when writing reviews. The task of finding all references to a given article is purely technical. However, the task of selecting a text fragment with mentioning a reference is no longer trivial and requires a marked training sample of the form ⟨reference, mentioning fragment⟩. An alternative approach, which does not require manual markup, is to select phrases from the text of the cited article that are semantically close to the local contexts of the references in the referenced articles. This does not require the exact boundaries of the referencing fragments, and the citing prompter becomes a kind of extractive prompter. Although the phrases are chosen from the original cited article, they reflect the opinion of the scientific community.

Task 4: *evaluating the quality of the system of prompters.* A validation sample is collected automatically from a large collection of articles, similar to the first problem. The selection serves as the list of references, and the "ideal" review as a union of the review sections of the article. For each review sentence, a prompter that gives the closest phrase among the first k positions of the search engine list is chosen. The quality score is the average position of the most relevant phrases from the prompters. At the same time, the usefulness of each prompter is evaluated as the number of its phrases that were included into the review.

In the experiment, to build a review scenario, we first solved the problem of selecting review parts on a subsample of articles from the S2ORC collection. Then, a ranking model was trained to rank the articles from the selected review parts based on boosting. The Kendell correlation coefficient of the resulting model $\tau = 0.48\%$ (ranking by year of publication gives $\tau = 0.1\%$). In addition, the work of the citing prompter was considered in detail. The CL-SciSumm 2018 collection was used for training and quality assessment. The results show that with automatic review scenario, our results are comparable to the best results on CL-SciSumm 2018. Semi-automatic scenario (selecting from the first 3 phrases) achieves a quality improvement of up to 7 percent on the ROUGE metric (Tab. 1).

	Abstract	Community summary	Human summary
top-1 MLP	0.30	0.33	0.22
top-1 boosting	0.25	0.32	0.20
top-3 MLP	0.30	**0.40**	0.23
top-3 boosting	0.29	0.36	**0.24**
best 2018	**0.33**	0.25	0.21

Table 1. $ROUGE_2(f1)$ on CL-SciSumm 2018 for the citing prompter.

In the future, the quality of each prompter's ranking will be evaluated in the SciSearch.ai search-recommendation system as the average position of those phrases which users selected for review.

In conclusion, it should be noticed that the proposed method can also be seen as a way of *nonlinear reading*, when the user has the task not only to understand a little familiar area to him, but also simultaneously to produce an information product in the form of a review.

This research is funded by RFBR grant 20-07-00936.

[1] *Vlasov A.* Methods of machine aided human summarization of scientific articles // Master's thesis, MIPT, Moscow, 2020.

[2] *Jaidka K., Yasunga M., Chandrasekaran M., Radev D., Kan M.* The CL-SciSumm Shared Task 2018: Results and Key Insights, 2018.

[3] *Lo K., Wang L., Neumann M., Kinney R., Weld D.* S2ORC: The Semantic Scholar Open Research Corpus, 2020.

Оптимизация весов модальностей в тематических моделях транзакционных данных

Хрыльченко Кирилл Ярославович[1]⋆ elightelol@gmail.com
Воронцов Константин Вячеславович[1,2] vokov@forecsys.ru
[1]Москва, ФИЦ «Информатика и управление» РАН
[2]Москва, Московский физико-технический институт

Вероятностные тематические модели (ВТМ) обычно применяются для выявления кластерной структуры текстовых коллекций. Тематическая модель определяет, к каким темам относится каждый документ и какие слова составляют каждую тему. Преимуществом ВТМ является возможность интерпретации каждой темы средствами естественного языка, а также возможность обобщения на случай мультимодальных данных, когда документы содержат не только слова, но и термы других модальностей, таких как авторы, ссылки, теги, названия, пользователи и т. д.

В данной работе ВТМ применяется для анализа транзакционных данных корпоративных клиентов банка. Документ порождается транзакциями клиента на заданном интервале времени. Модальностями являются: контрагенты (в двух ролях — продавцы и покупатели), товары или услуги (также в двух ролях), а также бизнес-сегмент клиента. В отличие от текстов, в данной задаче возникают числовые модальности — суммы транзакций покупки и продажи. Темы характеризуют виды экономической деятельности компаний. Тематическое векторное представление корпоративного клиента используется для прогнозирования его дефолта. Точнее, решается задача бинарной классификации для предсказания 90-дневной просрочки на горизонте года.

Пусть D — коллекция документов, W_m — словарь термов модальности $m \in M$, терм $w \in W_m$ в документе d встречается n_{dw} раз, n_d^m — число термов модальности m в документе d. Мультимодальная тематическая модель максимизирует взвешенную сумму логарифмов правдоподобия модальностей:

$$\sum_{m\in M} \lambda_m \sum_{d\in D} \sum_{w\in W_m} n_{dw} \log \sum_{t\in T} \varphi_{wt}\theta_{td} \to \max_{\Phi,\Theta},$$

где T — множество тем, λ_m — веса модальностей, Φ и Θ — матрицы параметров модели $\varphi_{wt} = p(w \mid t)$ и $\theta_{td} = p(t \mid d)$.

Для решения данной задачи используется ЕМ-алгоритм.

Нами доказано, что оптимальное тематическое представление документа d представимо в виде выпуклой комбинации

$$\theta_{td} = \sum_{m\in M} \tau_d^m \theta_{td}^m, \quad \tau_d^m = \frac{\lambda_m n_d^m}{\sum_{m'\in M} \lambda_{m'} n_d^{m'}},$$

где $\theta_{td}^m = p(t \mid d, m)$ — оптимальное унимодальное векторное представление документа d по модальности m.

Модель	**ROC-AUC**
$\lambda_m = 1$	0.6153 ± 0.0108
λ_m^*	0.6686 ± 0.0064
вещественные модальности	0.7126 ± 0.0109
лучшая унимодальная модель	0.7131 ± 0.0027
оптимизация λ_m	$\mathbf{0.7356} \pm 0.0055$
Конкатенация векторов θ_m	0.7427 ± 0.0057

Таблица 1. Сравнение стратегий балансировки модальностей.

Проблема возникает в тех случаях, когда n_d^m существенно различаются для разных m. Для балансировки модальностей предлагается либо использовать средневзвешенные веса

$$\lambda_m^* = \sum_{d \in D} \frac{n_d}{\sum_{d \in D} n_d} (n_d^m)^{-1},$$

либо оптимизировать на каждом шаге EM-алгоритма целевой критерий $J(\Theta_m, \lambda_m \colon m \in M)$ по скалярным переменным λ_m методом градиентного спуска, где Θ_m — оптимальные унимодальные тематические представления документов.

Для добавления числовых модальностей для каждой темы t вводится гауссовское распределение $\mathcal{N}(\mu_t, \sigma_t^2)$, параметры которого обучаются по числовым данным документов, также путём максимизацией правдоподобия с помощью EM-алгоритма.

В качестве модели классификации клиентов используется градиентный бустинг lightgbm с фиксированными гиперпараметрами. Результат классификации оценивается с помощью метрики ROC-AUC по отложенной выборке. Каждый эксперимент повторяется 10 раз со случайной инициализацией параметров. Число документов в коллекции 84 тысячи.

Таблица 1 показывает, что введение числовых модальностей совместно с оптимизацией λ_m значительно улучшает качество модели предсказания дефолта. При отсутствии дополнительного критерия оптимизации J использование нормированных весов модальностей λ_m^* существенно лучше единичных весов.

Мультимодальная тематическая модель фактически пытается «сжать» информацию о всех модальностях в общее векторное представление размерности $|T|$. Поэтому конкатенацию унимодальных тематических моделей, т. е. одновременное обучение отдельных тематических моделей для каждой модальности с последующей конкатенацией полученных векторных представлений, можно рассматривать как способ получения верхней оценки качества мультимодальной модели. Полученные результаты говорят о том, что оптимизация весов модальностей по критерию J сохраняет почти всю полезную информацию, содержащуюся в модальностях.

Предложенные техники подбора и оптимизации весов модальностей, а также введения числовых модальностей, могут быть использованы для тематического моделирования как банковских транзакционных данных, так и мультимодальных данных любой другой природы.

Работа поддержана грантом РФФИ No. 20-07-00936.

[1] *Хрыльченко К. Я.* Обобщенные модальности в вероятностных тематических моделях для транзакционных данных // Магистерская диссертация, ВМК МГУ, 2020.

Optimizing modality weights for topic models of transaction data

Khrylchenko Kirill[1]⋆ elightelol@gmail.com
Vorontsov Konstantin[1,2] vokov@forecsys.ru
[1]Moscow, Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences
[2]Moscow, Moscow Institute of Physics and Technology

Probabilistic topic models (PTM) are usually applied to detection of clusters in textual collections. Topic model detects for each document to which topics it relates and also which words compose each topic. Ability to interpret each topic with natural language tools is an advantage of PTM, as well as an ability to generalize for multimodal data when documents consist not only of words but of other modalities' terms, such as authors, references, tags, titles, users, etc.

In this work PTM is used to analyze transactional data of corporate bank clients. Document is generated by client's transactions within a fixed time interval. Modalities are: counterparties (in two roles — as sellers and as buyers), goods or services (in two roles as well), and also business-segment of a client. Unlike textual data, given task has real-valued modalities —-paid transaction sums and received sums. Topics describe profiles of companies' economical activites. Topic embedding of corporate bank client is used for clients' default risk prediction. To be precise, we solve a binary classification task for prediction of 90 days long overdue on a one year horizon.

Let D be a corpus of documents, W_m — term vocabulary for modality $m \in M$, term $w \in W_m$ in a document d occures n_{dw} times, n_d^m — amount of terms of modality m in a document d. Multimodal topic model maximizes weighted sum of logarithms of modalities' likelihoods:

$$\sum_{m\in M} \lambda_m \sum_{d\in D} \sum_{w\in W_m} n_{dw} \log \sum_{t\in T} \varphi_{wt}\theta_{td} \to \max_{\Phi,\Theta},$$

where T is a set of topics, λ_m — modality weights, Φ and Θ — model parameters, $\varphi_{wt} = p(w \mid t)$, and $\theta_{td} = p(t \mid d)$.

To solve this task we use EM-algorithm.

We prove that optimal topic embedding for document d can be decomposed into convex combination

$$\theta_{td} = \sum_{m\in M} \tau_d^m \theta_{td}^m, \quad \tau_d^m = \frac{\lambda_m n_d^m}{\sum_{m'\in M} \lambda_{m'} n_d^{m'}},$$

where $\theta_{td}^m = p(t \mid d, m)$ is an optimal unimodal embedding for document d constructed only from modality m.

Model	ROC-AUC
$\lambda_m = 1$	0.6153 ± 0.0108
λ_m^*	0.6686 ± 0.0064
real-valued modalities	0.7126 ± 0.0109
best unimodal model	0.7131 ± 0.0027
optimized λ_m	$\mathbf{0.7356} \pm 0.0055$
Concatenation of θ_m	0.7427 ± 0.0057

Table 1. Comparison of different modality weighting strategies.

Problems arise when n_d^m differ signficantly for different m. To balance modalities we propose either using

$$\lambda_m^* = \sum_{d \in D} \frac{n_d}{\sum_{d \in D} n_d} (n_d^m)^{-1},$$

either optimizing a downstream criteria on each step of EM-algorithm: $J(\Theta_m, \lambda_m : m \in M)$ w.r.t. scalar variables λ_m with gradient descent method, where Θ_m is an optimal unimodal document embedding for modaltiy m.

To include real-valued modalities in each topic t we introduce a normal distribution $\mathcal{N}(\mu_t, \sigma_t^2)$ with learnable parameters which are adjusted to real-valued document data by likelihood maximization with EM-algorithm.

As a classification model we use gradient boosting lightgbm with fixed hyperparameters. Classification result is evaluated with ROC-AUC metric using holdout data. Every experiment is repeated 10 times with a different random seed. There are 84 thousands of documents in the corpus.

Table 1 demonstrates that introducing real-valued modalities together with optimization of λ_m significantly improves quality of default risk prediction model. In the abscence of any arbitrary optimization criteria J it is significantly better to use normalized modality weights λ_m^* in comparison to unit weights.

Multimodal topic model is in fact trying to "compress" information about all modalities into a single $|T|$-sized embedding. Thus concatenation of unimodal topic models, which implies simultaneous training of separate topic models for each modality followed by concatenation of resulting embeddings, can be regarded as a way to construct an upper bound for quality of multimodal model. Results indicate that optimization of criteria J w.r.t. modality weights retains almost all useful information stored in modalities.

Proposed techniques of adjustment and optimization of modality weights, together with real-valued modalities, can be applied to topic modeling of transactional bank data, as well as any other multimodal data.

This research is funded by RFBR grant 20-07-00936.

[1] *Khrylchenko K.* Genelalized modalities for topic modeling of transaction data // Master's thesis, MSU, Moscow, 2020.

Инкрементные тематические модели с аддитивной регуляризацией для выделения трендовых научных тем

Герасименко Николай Александрович[1,2]⋆ nikgerasimenko@gmail.com
Чернявский Александр Сергеевич[2,3] alschernyavskiy@gmail.com
Никифорова Мария Андреевна[2,3] labenzom@gmail.com
Никитин Максим Дмитриевич[2] mdnikitin@sberbank.ru
Воронцов Константин Вячеславович[1] vokov@forecsys.ru
[1]Москва, ФИЦ ИУ РАН
[2]Москва, ПАО Сбербанк
[3]Москва, НИУ «Высшая школа экономики»

Стремительный рост числа научных публикаций, интенсивное появление новых направлений и подходов ставит перед научным сообществом задачу своевременного автоматического выявления трендов. Под трендом мы понимаем семантически однородную тему, которая характеризуется устойчивым во времени лексическим ядром и резким, зачастую экспоненциальным, ростом числа публикаций [1]. Кроме того, тренд часто характеризуется ключевым термином, например, названием задачи, теории или метода. Примерами трендов в машинном обучении являются «LSTM», «deep learning» «word2vec», «BERT», «fake news detection».

Для выделения трендовых тем в потоке научных публикаций мы используем инкрементные методы вероятностного тематического моделирования. Для выбора количества новых тем мы предлагаем критерий, основанный на количестве терминов, частота употребления которых сильно выросла с момента последнего обновления модели. Для оценки качества мы вручную сформировали и сделали общедоступным датасет из 87 тредов, в котором каждый тренд характеризуется набором из не менее, чем 10 ключевых статей и 5 ключевых терминов.

Эксперименты по выделению трендов производились на коллекции из 68762 статей, опубликованных с 2000 по 2021 год на конференциях по машинному обучению с h-индексом, превышающим 100. Про каждую статью известна дата ее публикации. Обучение моделей производилось полностью без учителя, валидационная разметка использовалась только для финальной оценки качества. Подход считался тем лучше, чем меньше дней проходило в среднем между моментами возникновения трендов и моментами их извлечения.

Мы провели эксперименты для двух видов тематических моделей, PLSA и ARTM [2], и для двух вариантов последовательностей временных меток: 30 и 180 дней между обновлениями. Таким образом, мы получили результаты четырех моделей, обозначаемых как PLSA-30, ARTM-30, PLSA-180 и ARTM-180. При поступлении новой порции документов D' словарь пополняется новыми терминами W' и могут образоваться новые темы T'. Предполагается, что новая лексика, появившаяся в новых документах, относится преимущественно к но-

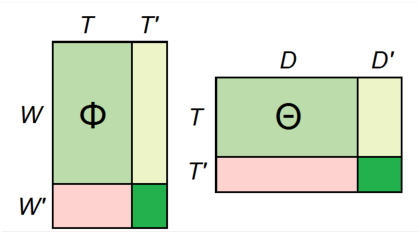


Рис. 1. Инкрементная тематическая модель. Нулевые блоки выделены красным цветом, а сильно разреженные – светло-зеленым.

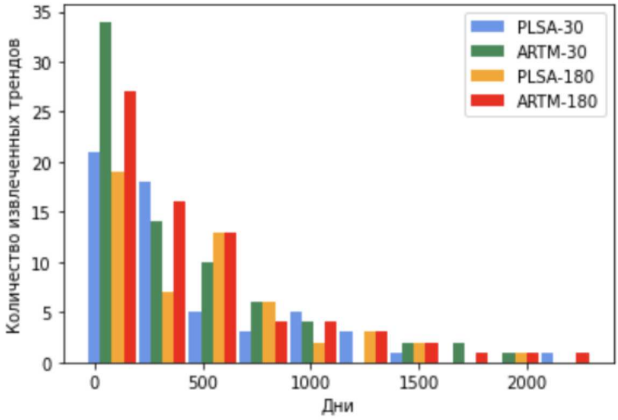


Рис. 2. Количество дней, затраченных моделями на выделение трендов при различном временном шаге.

вым темам, Рис. 1. На каждом шаге модель дообучалась на всех документах, опубликованных к соответствующему моменту времени, а не только на новых. Такой подход улучшает устойчивость тем во времени. Дополнительные ограничения на тематическую модель накладываются в рамках подхода аддитивной регуляризации ARTM с использованием библиотеки BigARTM [3]. В частности, для повышения различности тем используется регуляризатор декоррелирования.

Для определения момента извлечения тренда использовалась метрика Recall@k для подборок ключевых слов и ключевых документов тренда. Тренд считается выделенным, если в топ-50 терминах хотя бы одной из тем модели встретилось хотя бы 30% его ключевых терминов и одновременно в топ-50 ее документах встретилось хотя бы 10% его ключевых документов.

Распределение времен выделения отражено на Рис. 2. Наилучший результат показала модель ARTM-30, а модель ARTM-180 превосходит PLSA-180 и даже

PLSA-30 в задаче раннего обнаружения. Также ARTM-модели извлекли более 70 трендов, в то время, как PLSA-модели только около 50.

Таким образом, эксперименты показали, что ARTM почти во всех случаях превосходит PLSA. Кроме того, лучшая модель способна определить большую часть научных трендов в течение первого года их существования.

Работа поддержана грантом РФФИ No. 20-07-00936.

[1] *Kontostathis A., Galitsky M., Pottenger M., Roy S., Phelps J.* A Survey of emerging trend detection in textual data mining // Springer New York, 2004. Pp. 185–224.

[2] *Kochedykov D., Apishev M., Golitsyn L., Vorontsov K.* Fast and modular regularized topic modelling // Proc. of the 21st Conference of FRUCT, 2017. Pp. 182–193.

[3] *Vorontsov K., Frei O., Apishev M., Romov P., Dudarenko M.* BigARTM: Open source library for regularized multimodal topic modeling of large collections // AIST, Springer International Publishing, 2015. Pp. 370–381.

Incremental ARTM for Scientific Trend Topics Detection

Gerasimenko Nikolai[1,2]⋆ nikgerasimenko@gmail.com
Chernyavskiy Alexander[2,3] alschernyavskiy@gmail.com
Nikiforova Maria[2,3] labenzom@gmail.com
Nikitin Maxim[2] mdnikitin@sberbank.ru
Vorontsov Konstantin[1] vokov@forecsys.ru
[1]Moscow, FRC CSC RAS
[2]Moscow, Sberbank
[3]Moscow, Higher School of Economics

The rapid growth in the number of scientific publications, the intensive emergence of new directions and approaches poses a challenge to the scientific community to identify trends in a timely and automatic manner. By trend we mean a semantically homogeneous theme that is characterized by a steadily time lexical kernel and a sharp, often exponential, increase in the number of publications [1]. In addition, a trend is often characterized by a main term, such as the name of a problem, a theory, or a method. Examples of trends in machine learning are: "LSTM", "deep learning" "word2vec", "BERT", "fake news detection".

In order to extract trend themes in a stream of scientific publications, we use incremental methods of probabilistic topic modelling. To select the number of new topics, we propose a criterion based on the number of terms that have increased significantly since the last update of the model. To assess quality, we manually formed and made publicly available a dataset of 87 trends, in which each trend is characterized by a set of at least 10 key articles and 5 key terms.

Trend extraction experiments were carried out on a collection of 68,762 articles published between 2000 and 2021 at an h-index over 100 machine-learning conferences. The publication date of each article is known. The models were taught entirely unsupervised, and validation dataset were used only for final quality assessments. The approach was considered to be the better the less days, on average, passed between the occurrence of trends and the moment of their extraction.

We conducted experiments for two types of theme models, PLSA and ARTM [2], and for two variants of timestamps sequences: 30 and 180 days between updates. We have thus obtained the results of four models, labeled PLSA-30, ARTM-30, PLSA-180 and ARTM-180. Upon receipt of a new batch of documents D' the dictionary is replenished with new terms W' and new themes T'can be created. It is assumed that the new vocabulary, which has appeared in new documents, relates primarily to new topics, Figure 1. At each step, the model completed all documents published at the time, not just new ones. This approach improves the sustainability of topics over time. Additional constraints on the subject model are imposed within the framework of the additive regularization approach ARTM using BigARTM open source library [3]. In particular, a decoration regularizer is used to increase the variety of topics.

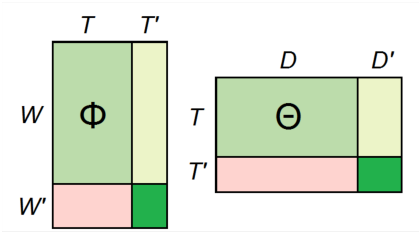


Fig. 1. Incremental topic modeling. Zeros are marked by red, a more sparse matrix is marked with a lighter green.

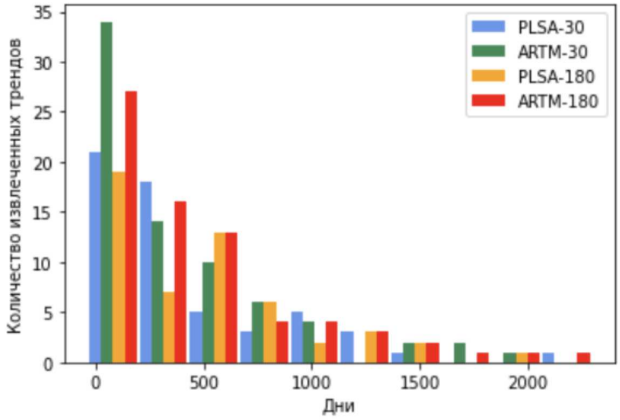


Fig. 2. Histogram represents the number of extracted trend topics depending on delay from its start.

To determine when to extract the trend, the metric Recall@k was used for key terms and key articles. The trend is considered extracted if at least 30% of the key terms are found in top-50 terms of at least one of the topics and at the same time 10% of the key documents are found in the top-50 documents of this topic.

Based on the matched trends, we calculated the delay between inception and extraction dates of each trend. Distributions of delays are demonstrated in Figure 2. The ARTM-30 are the best model, and the ARTM-180 outperformed the PLSA-180 and even the PLSA-30 in the early detection task. The ARTM-based models also extracted more than 70 trends, while the PLSA-based models were only about 50.

The quality is limited by several factors: the sizes of topics and their presence in the validation dataset (for instance, “em algorithm” and “pattern recognition” present quite weakly); the occurrence of keywords in articles (the keyword “gpt”

usually appears in a paper only several times); the dataset quality and the quality of internal components of the approach such as the matching procedure.

Thus, experiments have shown that ARTM is almost always superior to PLSA. In addition, the best model is able to detect most of the scientific trends during their first year of existence. There are possible directions for further research: tuning and improving the components of the current approach, as well as the investigation of the trend identification approaches.

This research is funded by RFBR grant 20-07-00936.

[1] *Kontostathis A., Galitsky M., Pottenger M., Roy S., Phelps J.* A Survey of emerging trend detection in textual data mining // Springer New York, 2004. Pp. 185–224.

[2] *Kochedykov D., Apishev M., Golitsyn L., Vorontsov K.* Fast and modular regularized topic modelling // Proc. of the 21st Conference of FRUCT, 2017. Pp. 182–193.

[3] *Vorontsov K., Frei O., Apishev M., Romov P., Dudarenko M.* BigARTM: Open source library for regularized multimodal topic modeling of large collections // AIST, Springer International Publishing, 2015. Pp. 370–381.

Нейронные тематические модели для рекомендации статей

Рамазанова Аяжан[1]⋆ a.ramazanova@phystech.edu
Янина Анастасия Олеговна[1] yanina@phystech.edu
Воронцов Константин Вячеславович[1] vokov@forecsys.ru
[1]Московский физико-технический институт

Данное исследование направлено на изучение эффективности нейросетевых тематических моделей для рекомендации текстовых документов пользователям электронной библиотеки. Одним из ограничений обычных «плоских» тематических моделей является то, что они представляют тематическую структуру в виде одноуровневого списка тем, в то время как для многих приложений важно выявлять иерархические взаимосвязи между темами и подтемами. В поисковых и рекомендательных сервисах иерархическая структура позволяет улучшать качество поиска [1], уточнять поисковые потребности пользователя, структурировать и интерпретировать результаты поиска, автоматизировать построение иерархических рубрикаторов.

Мы исследуем два типа иерархических тематических моделей: вероятностный многокритериальный подход — аддитивную регуляризацию тематических моделей ARTM [2] и древовидную нейронную тематическую модель TSNTM [3]. Регуляризация позволяет контролировать разреженность отношения между родительскими и дочерними темами, но подбор коэффициентов регуляризации и других гиперпараметров модели требует больших вычислительных ресурсов. Метод TSNTM определяет число тем автоматически, что делает его перспективным для извлечения иерархической тематической структуры текстовой коллекции.

В экспериментах мы сравниваем рекомендательные модели, основанные на различных векторных представлениях текста. Используются стандартные способы предварительной обработки: удаляются знаки препинания, текст переводится в нижний регистр, токенизируется в слова и лемматизируется. Далее текст преобразуется в векторное представление (эмбединг) следующими методами: TF-IDF, Fasttext, BERT, LDA, hARTM, TSNTM. Для поиска семантически близких документов эмбединги запросов и документов сравниваются с помощью скалярного произведения и косинусной меры. Все подходы протестированы на наборе данных триплетов ArXiv, который содержит 19 876 триплетов вида «запрос, релевантная статья, нерелевантная статья» для выбора наилучшего сочетания метода векторизации и меры сходства.

Таблица 1 показывает превосходство нейросетевой иерархической тематической модели TSNTM над другими подходами.

Чтобы объединить преимущества вероятностных и нейросетевых тематических эмбедингов, мы дополнительно исследовали несколько гибридных моделей (Таблица 2). В гибридной модели конкатенируется несколько тематических эмбедингов (в следующих таблицах «+» означает конкатенацию векторов). В свя-

Метод	Dot ↑	Cosine ↑
TFIDF	0.505	0.505
BERT	0.489	0.563
LDA	0.693	0.694
hARTM	0.501	0.498
Fasttext	0.708	0.702
TSNTM	**0.744**	**0.750**

Таблица 1. Качество рекомендации статей: сравнение различных способов векторизации текстов (TF-IDF, Fasttext, BERT, LDA, hARTM и TSNTM) и мер сходства векторов (скалярное произведение и косинусное расстояние)

зи с тем, что в наших экспериментах TSNTM значительно превосходит все другие модели (как тематические, так и предобученные нейронные сети), мы решили работать с комбинациями на основе LDA и hARTM. Добавление эмбедингов на основе hARTM к BERT или Fasttext приводит к некоторому снижению качества рекомендаций, возможно, из-за недостаточно тщательной настройки гиперпараметров в hARTM. Объединение тематических векторов LDA с BERT или Fasttext показывают небольшое увеличение качества для большинства комбинаций.

Метод	Dot ↑	Cosine ↑
LDA	0.693	0.694
hARTM	0.501	0.498
BERT	0.489	0.563
LDA + BERT	0.489	0.564
LDA + σ(BERT)	0.697	0.692
LDA + Fasttext	**0.699**	**0.688**
LDA + σ(Fasttext)	0.697	0.686
hARTM+BERT	0.489	0.563
hARTM+σ(BERT)	0.501	0.496
hARTM+Fasttext	**0.706**	**0.685**
hARTM+σ(Fasttext)	0.501	0.496

Таблица 2. Гибридные нейро-тематические представления. σ означает применения softmax к эмбедингам для нормализации.

Нейросетевые тематические модели с древовидной структурой TSNTM дают более высокое качество рекомендаций статей по сравнению с другими тематическими моделями (LDA и иерархическая аддитивная регуляризация тематических моделей), эмбедингами на основе BERT, Fasttext, TF-IDF, а также комбинированными представлениями (на основе BERT и тематических векторов). Полученные результаты могут быть использованы для разработки новых поисково-рекомендательных сервисов, использующих дополнительную тематическую информацию о текстах.

Работа поддержана грантом РФФИ No. 20-37-90025.

[1] *Ianina A., Vorontsov K.* Hierarchical Interpretable Topical Embeddings for Exploratory Search and Real-Time Document Tracking // International Journal of Embedded and Real-Time Communication Systems (IJERTCS), 2020. Vol.11(4).

[2] *Vorontsov K. and Potapenko A.* Additive regularization of topic models // Machine Learning, 2015. Vol. 101(1). Pp. 303–323.

[3] *Isonuma M. et al.* Tree-Structured Neural Topic Model // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020. Pp. 800–806.

Neural Topic Models for Article Recommendation

Ayazhan Ramazanova[1]⋆ a.ramazanova@phystech.edu
Anastasia Ianina[1] yanina@phystech.edu
Konstantin Vorontsov[1] vokov@forecsys.ru
[1]Moscow Institute of Physics and Technology

The present study aims to examine the effectiveness of neural topic models applied to article recommendation task. One of the limitations of conventional "flat" topic models is that they present the topic structure as a single-level list of topics, while for many applications it is important to identify hierarchical relationships between topics and subtopics. Applied to an information retrieval or recommendation scenario, hierarchical structure is preferable, because it allows to improve the quality of search [1], refine the user's search needs, structure and interpret search results, and automate the construction of hierarchical rubricators.

We investigate two types of hierarchical topic models: non-Bayesian multicriteria approach called Additive Regularization of Topic Models (ARTM, [2]) and a tree-structured neural topic model (TSNTM, [3]). While regularization in the first approach controls the sparsity of the relations, finding the right model hyperparameters requires excessive compute resources, which results in insufficient performance. Moreover, TSNTM can control the number of topics automatically, which makes it a good choice for hierarchical topical feature extractor.

We conduct experiments to compare recommendation models based on different vector representations of text. First, each text is preprocessed: reduced to lower case, stripped of punctuation marks, tokenized into words and lemmatized where needed. Second, it is transformed into a vectorized representation following one of the following approaches: TF-IDF, Fasttext, BERT, LDA, hARTM or TSNTM. Then we compare query and document representatons using dot similarity and cosine similarity. All the discussed above approaches are tested on the dataset of ArXiv triplets which contains $19,876$ document triplets of the form "query - relevant article - irrelevant article" which lets us to find the best combination of textual vector representation and the similarity measure.

In Table 1 we show the superiority of a neural hierarchical topic model (TSNTM) compared to other approaches.

To exploit advantages of both topic modeling and neural-based embeddings several hybrid models were further investigated (Table 2). In a hybrid model several topical embeddings are simply concatenated to form the final representation (in the following tables "+" stands for vectors concatenation). Due to that fact that TSNTM significantly outperforms all the other models (both topical and neural), we decided to combine only LDA and hARTM representations with pretrained neural embeddings. Adding hARTM embeddings to BERT or Fasttext results in a slight decrease in scores, probably, due to inaccurate parameter fine-tuning for hARTM.

Method	Dot ↑	Cosine ↑
TFIDF	0.505	0.505
BERT	0.489	0.563
LDA	0.693	0.694
hARTM	0.501	0.498
Fasttext	0.708	0.702
TSNTM	**0.744**	**0.750**

Table 1. Article recommendation quality: comparison between different textual feature extractors (TF-IDF, Fasttext, BERT, LDA, hARTM and TSNTM) and similarity measures (dot product, cosine similarity, Jaccard similarity coefficient)

LDA vectors concatenated with BERT and Fasttext show a subtle increase in metrics for most of the combinations.

Method	Dot ↑	Cosine ↑
LDA	0.693	0.694
hARTM	0.501	0.498
BERT	0.489	0.563
LDA + BERT	0.489	0.564
LDA + σ(BERT)	0.697	0.692
LDA + Fasttext	**0.699**	**0.688**
LDA + σ(Fasttext)	0.697	0.686
hARTM+BERT	0.489	0.563
hARTM+σ(BERT)	0.501	0.496
hARTM+Fasttext	**0.706**	**0.685**
hARTM+σ(Fasttext)	0.501	0.496

Table 2. Combined neural+topical representations. σ stands for softmax applied to embeddings.

All in all, we demonstrate the effectiveness of tree-structured neural topic models applied to article recommendation task. They bring consistent improvements compared to other topic models (including Latent Dirichlet Allocation and Hierarchival Additive Regularization of Topic Models), BERT-based representations, Fasttext, TF-IDF and also combined representations (BERT-based + topical features). The achieved results can be used as a starting point to study and develop new article recommendation methods that use additional topical information about the texts.

This research is funded by RFBR, grant 20-37-90025.

[1] *Ianina A., Vorontsov K.* Hierarchical Interpretable Topical Embeddings for Exploratory Search and Real-Time Document Tracking // International Journal of Embedded and Real-Time Communication Systems (IJERTCS), 2020. Vol.11(4).

[2] *Vorontsov K. and Potapenko A.* Additive regularization of topic models // Machine Learning, 2015. Vol. 101(1). Pp. 303–323.

[3] *Isonuma M. et al.* Tree-Structured Neural Topic Model // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020. Pp. 800–806.

Реализация EM-алгоритма для аддитивно регуляризованных тематических моделей на GPU

Сердюк Юлиан Анатольевич[1]⋆ jserdyuk@cs.msu.ru
Воронцов Константин Вячеславович[1] vokov@forecsys.ru
[1]Москва, МГУ им. М. В. Ломоносова

Вероятностная тематическая модель коллекции текстовых документов описывает каждый документ d дискретным распределением $p(t|d) = \theta_{td}$ по темам, а каждую тему t — дискретным распределением $p(w|t) = \varphi_{wt}$ по словам w. Для определения параметров модели, матриц $\Phi = (\varphi_{wt})$ и $\Theta = (\theta_{td})$, максимизируется критерий логарифма правдоподобия, к которому добавляется взвешенная сумма критериев регуляризации $R(\Phi, \Theta)$. Для решения оптимизационной задачи применяется EM-алгоритм [1]:

$$p_{tdw} = \mathrm{norm}_t\big(\varphi_{wt}\theta_{td}\big); \tag{E}$$

$$\varphi_{wt} = \mathrm{norm}_w\left(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right); \qquad n_{wt} = \sum_d n_{dw}p_{tdw}; \tag{M$_\Phi$}$$

$$\theta_{td} = \mathrm{norm}_t\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right); \qquad n_{td} = \sum_w n_{dw}p_{tdw}. \tag{M$_\Theta$}$$

где n_{dw} — исходные данные о частотах слов w в документах d, $\mathrm{norm}_t(x_t) = \frac{\max\{x_t,0\}}{\sum_s \max\{x_s,0\}}$ — операция нормирования вектора (x_t), $p_{tdw} = p(t|d,w)$ — вспомогательные переменные.

Онлайновый EM-алгоритм с возможностями пакетной обработки данных и аддитивной регуляризации реализован в библиотеке с открытым кодом BigARTM [2]. Параллельные вычисления в ней производятся на ядрах центрального процессора (CPU). Существует несколько мало известных реализаций тематического моделирования с использованием графических процессоров (GPU), преимущественно для модели латентного размещения Дирихле и без поддержки аддитивной регуляризации [3]. В данной работе предлагается способ распараллеливания регуляризованного EM-алгоритма на GPU.

Согласно формулам (E) и (M) шагов, значения p_{tdw}, φ_{wt} и θ_{td} вычисляются независимо друг от друга. Следовательно, их вычисление может быть разбито на независимые подзадачи и выполнено параллельно. В случае обработки больших коллекций документов с большим числом тем матрицы (p_{tdw}), Φ и Θ, скорее всего, не поместятся в памяти GPU. Как и в BigARTM, мы разделяем коллекцию на пакеты документов, которые обрабатываются по очереди. Это позволяет обновлять по формулам M-шага только те ячейки матриц Φ и Θ, которые соответствуют документам d из загруженного в память пакета. Некоторые трудности возникают при вычислении матрицы (p_{tdw}), поскольку она зависит одновременно от обеих матриц Φ и Θ. Значение p_{tdw} вычисляется

через скалярное произведение строки w матрицы Φ и столбца d матрицы Θ. Поэтому достаточно подгружать лишь те части матриц Φ и Θ, которые соответствуют словам и документам из текущего пакета. Перенормировка столбцов в матрицах Φ и Θ производится после обновления всех значений n_{wt} и n_{td} для текущего пакета. Отметим, что операция матричного умножения является векторной, следовательно, во многих специализированных библиотеках она эффективно реализована на уровне CPU. Однако такие библиотеки часто требуют хранения данных в некотором внутреннем формате, что может привести к дополнительным затратам на перевод из одного формата в другой.

При реализации аддитивно регуляризованных тематических моделей на GPU требуется уделить особое внимание регуляризаторам. В общем случае нельзя гарантировать эффективную реализацию M-шага для произвольного регуляризатора $R(\Phi, \Theta)$ на GPU. Если регуляризатор требует интенсивных обращений к матрице Φ, то вычисление на GPU может оказаться даже медленнее, чем на CPU. Однако основные регуляризаторы, часто применяемые в аддитивной регуляризации, легко реализуются на GPU. В частности, регуляризаторы сглаживания и разреживания требуют лишь добавления заданных констант к значениям n_{wt} и n_{td}, что хорошо реализуются на GPU, поскольку каждое значение вычисляется независимо от других. Регуляризатор декоррелирования требует обращения ко всей строке w матрицы Φ при обновлении значения n_{wt}. Это требует выделения дополнительной памяти и незначительно увеличивает время вычислений.

Реализация EM-алгоритма на GPU может превосходить производительность BigARTM в десятки раз. Поскольку BigARTM требует настройки большого числа гиперпараметров (числа тем, коэффициентов регуляризации), это существенно ускоряет процесс подбора модели и может приводить к улучшению качества модели при увеличении числа экспериментов.

Для реализации аддитивно регуляризованных тематических моделей на GPU использовалась библиотека numba, которая позволяет создавать код на python, который может быть запущен как на CPU, так и на GPU (требуется CUDA-совместимая графическая карта).

В таблице приводится сравнение производительности (в секундах) BigARTM, CPU-версии и GPU-версий разработанной библиотеки для некоторых известных коллекций в задаче классификации текстовых документов. классификации текстовых документов. Во всех случаях число тем равно 100, выполнялось 10 итераций, регуляризаторы не использовались. Эксперименты проводились на машине с CPU Ryzen 2600 (6 ядер, 12 потоков) и GPU RTX 2070 (8GB VRAM).

коллекция	документов	слов	BigARTM	CPU	GPU
20 Newsgroups	18 821	62 707	12.9	24.6	1.7
AG's News	127 600	59 371	29.2	52.5	3.4
DBPedia	630 000	650 142	184.3	285.9	12.5

Таким образом, GPU-версия на порядок быстрее CPU реализаций EM-алгоритма тематического моделирования. Дальнейшее улучшение GPU-версии связано с обеспечением совместимости с BigARTM по форматам данных, интерфейсам и библиотекам регуляризаторов.

Работа поддержана грантом РФФИ No. 20-07-00936.

[1] *Kochedykov D., Apishev M., Golitsyn L., Vorontsov K.* Fast and modular regularized topic modelling // Proc. of the 21st Conference of FRUCT, 2017. Pp. 182–193.

[2] *Vorontsov K., Frei O., Apishev M., Romov P., Dudarenko M.* BigARTM: Open source library for regularized multimodal topic modeling of large collections // AIST, Springer International Publishing, 2015. Pp. 370–381.

[3] *Апишев М.* Эффективные реализации алгоритмов тематического моделирования // Труды ИСП РАН, 2020. Т. 32(1). С.137–152.

GPU implementation of parallel EM-algorithm for additive regularization of topic models

Serdyuk Julian[1] jserdyuk@cs.msu.ru
Vorontsov Konstantin[1] vokov@forecsys.ru
[1]Moscow, Lomonosov Moscow State University

Topic Model of a collection of text documents describes each document d with discrete distribution $p(t|d) = \theta_{td}$ over topics, where each topic t is a discrete distribution $p(w|t) = \varphi_{wt}$ of words w. To determine the parameters of the model, the matrices $\Phi = (\varphi_{wt})$ and $\Theta = (\theta_{td})$, the log likelihood criterion with the weighted sum of the regularization criteria $R(\Phi, \Theta)$ is maximized. To solve the optimization problem, the EM-algorithm [1] is used:

$$p_{tdw} = \operatorname{norm}_t(\varphi_{wt}\theta_{td}); \tag{E}$$

$$\varphi_{wt} = \operatorname{norm}_w\left(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right); \qquad n_{wt} = \sum_d n_{dw}p_{tdw}; \tag{M_Φ}$$

$$\theta_{td} = \operatorname{norm}_t\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right); \qquad n_{td} = \sum_w n_{dw}p_{tdw}. \tag{M_Θ}$$

where n_{dw} is the frequency of the word w in the document d, $\operatorname{norm}_t(x_t) = \frac{\max\{x_t,0\}}{\sum_s \max\{x_s,0\}}$ is the operation of vector (x_t) normalization, $p_{tdw} = p(t|d,w)$ are the auxiliary variables.

An online EM algorithm with batch processing capabilities and additive regularization is implemented in the BigARTM [2] open source library. Parallel computations in it are performed on cores of the central processing unit (CPU). There are several little known implementations of topic modeling using graphics processing units (GPUs), usually for Latent Dirichlet Allocation model and without support for additive regularization [3]. In this paper, we propose a method for parallelizing the regularized EM-algorithm on GPU.

According to formulas (E) and (M) steps, the values of p_{tdw}, φ_{wt} and θ_{td} are calculated independently of each other. Therefore, their calculation can be split into independent subtasks and performed in parallel. In the case of processing large collections of documents with a large number of topics, the matrices (p_{tdw}), Φ and Θ, most likely, will not fit in GPU memory. As in BigARTM, we divide the collection into batches of documents, which are processed in turn. This allows updating by M-step formulas only those cells of the matrices Φ and Θ, which correspond to the d documents from the package loaded into memory. Some difficulties arise when calculating the matrix (p_{tdw}), since it depends simultaneously on both matrices Φ and Θ. The value of p_{tdw} is calculated as the dot product of the row w of the matrix Φ and the column d of the matrix Θ. Therefore, it is enough to load only those parts of the matrices Φ and Θ that correspond to words and documents from

the current package. Renormalization of columns in matrices Φ and Θ is performed after updating all the values of n_{wt} and n_{td} for the current batch. Note that the operation of matrix multiplication is a vector operation, therefore, in many specialized libraries it is effectively implemented at the CPU level. However, such libraries often require storing data in some internal format, which can lead to additional costs for translation from one format to another.

When implementing additive regularization of topic models on the GPU, special attention should be paid to regularizers. In the general case, it is impossible to guarantee an efficient implementation of the M-step for an arbitrary $R(\Phi, \Theta)$ regularizer on the GPU. If the regularizer requires intensive calls to the matrix Φ, then the computation on the GPU may turn out to be even slower than on the CPU. However, the basic regularizers often used in additive regularization are easily implemented on the GPU. In particular, smooth and sparse Φ/Θ regularizers only require the addition of specified constants to the n_{wt} and n_{td} values, which are well implemented on the GPU, since each value is computed independently of the others. The decorrelation regularizer requires access to the entire row w of the matrix Φ when updating the value n_{wt}. This requires additional memory allocation and slightly increases the computation time.

The implementation of the EM algorithm on the GPU can exceed the performance of BigARTM dozens of times. Since additive regularization of topic models requires tuning a large number of hyperparameters (number of topics, regularization coefficients), this significantly speeds up the process of fitting a model and can lead to an improvement in the quality of the model with an increase in the number of experiments.

To implement additive regularization of topic models on the GPU, the numba library was used, which allows to write python code that can be run on both CPU and GPU (requires a CUDA-compatible graphics card).

The table below contains a comparison of the performance (in seconds) of BigARTM, the CPU version and the GPU versions of the developed library for some well-known collections in the task of classifying text documents. In all cases, the number of topics is 100, 10 iterations were performed, no regularizers were used. The experiments were performed on a machine with CPU Ryzen 2600 (6 cores, 12 threads) and GPU RTX 2070 (8GB VRAM) installed.

collection	documents	words	BigARTM	CPU	GPU
20 Newsgroups	18 821	62 707	12.9	24.6	1.7
AG's News	127 600	59 371	29.2	52.5	3.4
DBPedia	630 000	650 142	184.3	285.9	12.5

Thus, the GPU version is an order of magnitude faster than the CPU implementations of the EM topic modeling algorithm. Further improvement of the GPU version is associated with ensuring compatibility with BigARTM in terms of data formats, interfaces and regularizer libraries.

This research is funded by RFBR grant 20-07-00936.

[1] *Kochedykov D., Apishev M., Golitsyn L., Vorontsov K.* Fast and modular regularized topic modelling // Proc. of the 21st Conference of FRUCT, 2017. Pp. 182–193.

[2] *Vorontsov K., Frei O., Apishev M., Romov P., Dudarenko M.* BigARTM: Open source library for regularized multimodal topic modeling of large collections // AIST, Springer International Publishing, 2015. Pp. 370–381.

[3] *Apishev M.* Effective implementations of topic modeling algorithms // Trudy ISP RAN, 2020. Vol. 32(1).

Задачи и методы понимания естественного языка для мониторинга медиа-пространства

Воронцов Константин Вячеславович[1] vokov@forecsys.ru
[1]Москва, ФИЦ ИУ РАН, МФТИ, ВМК МГУ

Фейковые новости, слухи, лженаучные и конспирологические теории, пропагандистские компании, идеологические атаки и информационные войны составляют весомую долю современного медиа контента. Информационные технологии способствуют распространению любых идей со скоростью эпидемии. Фейки сосуществуют в медиа-пространстве с их опровержениями, продолжая воздействовать на умы людей, искажая их картину мира и ценностные установки. Противостоять явлениям постправды могли бы технологии машинного обучения, нацеленные на выявление потенциально опасного контента в медиапространстве.

В области детекции фейков принято выделять семь основных задач, решение которых предполагает формирование размеченных обучающих выборок и применение технологий обработки текстов и машинного обучения [1].

Выявление обмана в тексте (deception detection) исследуется в психологии и криминологии достаточно давно. Задача сводится к классификации текста на два класса «правда/ложь». Обучающие выборки формируются с помощью контролируемых экспериментов, либо по материалам судебных разбирательств. В роли признаков выступают десятки известных из психолингвистики вербальных маркеров — экспрессивность, категоричность, уклончивость, плеоназмы, отрыв от контекста и др. Достигаемые уровни точности или F-меры 70–92% в зависимости от задачи.

Автоматическая проверка фактов (automated fact-checking) основана на опыте ручной перепроверки фактов в журналистских сообществах. За последнее десятилетие появились десятки платформ для проверки фактов: Politifact, FullFact, FactCheck и др. Задача ставится как классификация текста, обычно по пятибалльной шкале от «ложь» до «правда». Регулярно проводятся соревнования: CLEF, FEVER, Rumour-Eval и др. Обученные модели позволяют находить переформулированные версии фейков, для которых в базах данных уже собраны опровержения.

Выявление позиции относительно запроса (stance detection) сводится к известной задаче текстового следования — классифицировать пару текстов ⟨запрос, гипотеза⟩ на три класса: «следует», «противоречит», «нерелевантно». Для соревнований доступны выборки данных Emergent, SemEval, FakeNewsChellenge и др., на которых достигались значения F1-меры до 97% (на новостях), точности до 68% (на Твиттере).

Выявление несоответствия заголовка и контента (clickbait detection) также аналогично задаче текстового следования — пара ⟨заголовок, контент⟩ классифицируется на два класса. В качестве признаков используются вербальные

маркеры гиперболизации, экспрессии, противоречия, а также признаки повышенного web-трафика. На открытых соревнованиях Webis-Clickbait достигались значения F1-меры до 68%, точности до 86%.

Выявление разногласий (controversy detection) сводится к кластеризации мнений по заданном вопросу, обычно по данным социальных сетей. Достигаемые уровни точности порядка 73–83% на выборке обсуждений статей Википедии.

Выявление поляризации общественного мнения (polarization detection) нацелено на поиск разногласий по заданной совокупности запросов или тем. Выборки данных формируются по открытым тематическим сообществам в социальных сетях.

Оценивание достоверности (credibility scores) является известной задачей в социологии, психологии, маркетинге. Требуется дать числовую оценку уровня доверия источнику или отдельной новости. В качестве признаков используются данные о распространении обмана, фейков или спама в прошлом, о медиакампаниях источника, о стиле контента, о геолокации, о среднем образовательном уровне читателей. На отдельных задачах достигались уровни AUC до 89%, точности до 81%.

Интересно, что универсальные нейросетевые модели языка далеко не всегда достигают лучших результатов в этих задачах. При малых объёмах данных задачи понимания естественного языка могут лучше решаться узко специализированными моделями.

Типология потенциально опасного дискурса в средствах массовой коммуникации не исчерпывается фейками. Выявление пропаганды и информационных войн расширяет спектр задач. Детекция пропаганды предполагает выявление её структурных элементов (таких как подмена фактов мнениями, замалчивание, деконтекстуализация и реконтекстуализация), путей распространения и новых трендов, приёмов речевой манипуляции, идеологем и мифологем, целевой аудитории воздействия и возможных её психо-эмоциональных реакций.

Несмотря на содержательное разнообразие задач, с точки зрения технологий машинного обучения они делятся на четыре способа разметки данных, и требуют, соответственно, четырёх типов моделей: (1) классификация текста целиком, (2) классификация пары текстов, (3) выделение и классификация фрагмента текста, (4) кластеризация множества текстов. Выделение некоторых вербальных маркеров в качестве признаков может требовать решения отдельных задач обучения типа (3).

В качестве примера рассмотрим задачу выделения полярных мнений о политическом событии [2]. Модель основана на предположении, что мнения различаются вербальными маркерами трёх типов: (а) факты в представлении троек «объект–субъект–действие», (б) семантические роли слов, (в) тонально окрашенная лексика. По каждому из трёх типов маркеров был построен отдельный словарь, который использовался в качестве модальности в вероятностной тема-

тической модели для кластеризации текстов по мнениям. Модели, построенные по отдельным модальностям, давали значения F1-меры до 70%. Модели, в которых использовались две из трёх модальностей, улучшали этот результат до 80%. Комбинация всех трёх модальностей давала наилучший результат 85–86%. Был сделан вывод о необходимости использования всех трёх типов маркеров.

В предлагаемом подходе формирование размеченных выборок становится основным способом формализации гуманитарных знаний для автоматизации мониторинга медиа-пространства.

Работа поддержана грантом РФФИ No. 20-07-00936.

[1] *Saquete E., Tomas D., Moreda P., Martinez-Barco P., Palomar M.* Fighting post-truth using natural language processing: A review and open challenges // Expert Systems With Applications, Elsevier, 2020.

[2] *Feldman D., Sadekova T., Vorontsov K.* Combining facts, semantic roles and sentiment lexicon in a generative model for opinion mining // Computational Linguistics and Intellectual Technologies, 2020. Pp. 268–283.

Problems and approaches of natural language understanding for media monitoring

Vorontsov Konstantin[1] vokov@forecsys.ru

[1]Moscow, Lomonosov Moscow State University

Fake news, rumors, pseudoscientific and conspiracy theories, propaganda campaigns, ideological attacks and information wars make up a significant proportion of modern media content. Information technology contributes to the spread of any ideas at the speed of an epidemic. Fakes coexist in the media space with their refutations, continuing to influence the minds of people, distorting their worldview and value attitudes. Machine learning technologies aimed at identifying potentially dangerous media content could fight the phenomena of post-truth.

In the field of fake news detection, it is customary to distinguish seven main problems, the solution of which involves the formation of labeled training samples and the use of natural language processing and machine learning technologies [1].

Deception detection has been studied in psychology and criminology for a long time. The task is reduced to the classification of the text into two classes "true / false". Training samples are formed using controlled experiments or from court proceedings. Dozens of verbal markers known from psycholinguistics can be used as features: expressiveness, categoricality, evasiveness, pleonasms, detachment from the context, etc. Achievable levels of accuracy or F-measures of 70-92% depending on the task.

Automated fact-checking is based on the experience of manual fact-checking in journalistic communities. Over the past decade, dozens of fact-checking platforms have emerged: Politifact, FullFact, FactCheck, and more. The problem is posed as a classification of the text, usually on a five-point scale from "false" to "true". Competitions are regularly held: CLEF, FEVER, Rumour-Eval, etc. The trained models allow you to find rewritten versions of fakes, for which refutations have already been collected in the databases.

Stance detection is a special case of a well-known textual entailment task. Given the pair of texts ⟨task, hypothesis⟩ we are to classify them into three classes: "follows", "contradicts", "irrelevant". For competitions, data samples of Emergent, SemEval, FakeNewsChellenge, etc. are available, which achieved F1-measure values up to 97% (on news), accuracy up to 68% (on Twitter).

Clickbait detection reveals the mismatch between the title and the content. It is also a special case of textual entailment because of the pair ⟨title, content⟩ is classified into two classes. Verbal markers of hyperbolization, expression, contradictions, as well as increasing of web traffic are used as features. In the open competitions Webis-Clickbait, F1-measure is up to 68%, accuracy is up to 86%.

Controversy detection boils down to clustering opinions on a given question, usually according to social media data. Achieved levels of accuracy are in the order of 73-83% on a sample of Wikipedia article discussions.

Polarization detection is aimed at finding disagreements on a given set of claims or topics. Data samples are formed from open thematic communities in social networks.

Assessing credibility scores is a well-known task in sociology, psychology, and marketing. It is required to give a numerical estimate of the level of trust in the source or individual news. The spread of deceit, fakes or spam in the past, the source's media campaigns, the style of content, geolocation, and the average educational level of readers are used as features. On individual tasks, AUC levels up to 89% were achieved, accuracy were about 81%.

It is interesting that universal neural language models do not always achieve the best results in these tasks. With small amounts of data, the nontrivial tasks of natural language understanding can be better solved by highly specialized models.

The typology of potentially dangerous discourse in the mass media is not limited to fakes. Revealing propaganda and information wars expands the range of tasks. The detection of propaganda involves the identification of its structural elements (such as substitution of facts with opinions, suppression, decontextualization and recontextualization), ways of spreading and new trends, methods of speech manipulation, ideologemes and mythologemes, target audience of influence and its possible psycho-emotional reactions.

Despite the substantial variety of tasks, from the machine learning point of view, they are divided into four ways of data labeling, and require four types of models respectively: (1) classification of the entire text, (2) classification of a pair of texts, (3) highlighting and classifying a piece of text, (4) clustering of a set of texts. Extraction of some verbal markers as features may require the solution of separate learning problems of type (3).

As an example, consider the problem of detecting polarizing opinions about a political event [2]. The model is based on the assumption that opinions differ in three types of verbal markers: (a) facts in the representation of triples "object–predicate–subject", (b) the semantic roles of words, (c) sentiment words. For each of the three types of markers, a separate vocabulary was built, which was used as a modality in a probabilistic topic model for clustering texts by opinions. Models built for individual modalities yielded F1-measure up to 70%. Models using two of the three modalities improved this result by up to 80%. The combination of all three modalities gave the best result 85–86%. Note that a small size labeled sample was used only for validation and comparing unsupervised models. It was concluded that it was necessary to use all three types of markers.

In the proposed approach, text labeling becomes the mainstream way of humanitarian knowledge formalization for the problem of automation of the media monitoring process.

This research is funded by RFBR grant 20-07-00936.

[1] *Saquete E., Tomas D., Moreda P., Martinez-Barco P., Palomar M.* Fighting post-truth using natural language processing: A review and open challenges // Expert Systems With Applications, Elsevier, 2020.

[2] *Feldman D., Sadekova T., Vorontsov K.* Combining facts, semantic roles and sentiment lexicon in a generative model for opinion mining // Computational Linguistics and Intellectual Technologies, 2020. Pp. 268–283.

Подходы к выявлению тревожных расстройств на основе автоматического анализа текстов комментариев в социальных сетях

Дюличева Юлия Юрьевна dyulicheva@gmail.com
Симферополь, Крымский федеральный университет имени В. И. Вернадского

Активное использование социальных сетей привело к накапливанию огромного количества комментариев, которые оставляют пользователи. В последнее время со стороны исследователей наблюдается интерес к такому направлению киберпсихологии как оценивание психического состояния на основе анализа комментариев в социальных сетях и влияния различного контента на физическое и ментальное здоровье человека. Тревожные и депрессивные расстройства являются одними из наиболее распространенных типов расстройств, которые проявляются на начальных этапах как негативное отношение и выражение беспокойства. Изменение отношения и поведения пользователя можно отслеживать на основе анализа текстов комментариев в социальных сетях, при условии, что пользователь не пытается скрыть своё истинное эмоциональное состояние путём самоконтроля. Ярко выраженное эмоциональное состояние и желание намеренно исказить смысл влияют на употребление устойчивых словосочетаний и лингвистические показатели текста, и, следовательно, представляют интерес для повышения качества алгоритмов распознавания.

Таблица 1. Оценки качества подходов

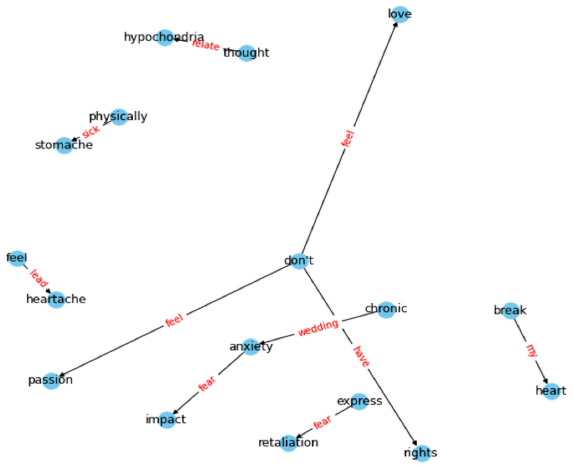
Подходы	Accuracy	Precision	Recall	F1
TF-IDF+RF	0.805	0.857	0.895	0.875
BERT+RF	0.896	0.893	0.893	0.893
Collgram+RF	0.681	0.791	0.522	0.629
TF-IDF+Collgram+RF	**0.864**	**0.861**	**0.902**	**0.881**
BERT+Collgram+RF	**0.933**	**0.931**	**0.948**	**0.939**

В проведённом исследовании в качестве датасета рассматривался фрагмент сбалансированной бинарной выборки из 3206 комментариев: 1756 комментариев с описанием тревожностей и 1450 обычных комментариев. Исходный датасет с комментариями, извлеченными из сети Reddit и результатами анализа стрессовых состояний описан в работе [1].

Для решения задачи классификации применялся случайный лес и изучались различные подходы к векторизации текста: применение меры TF-IDF, модели представления языка BERT, показателей анализа коллграмм, а также их различные комбинации. Анализ коллграмм представляет собой анализ N-грамм на основе двух показателей: показателя взаимной информации (MI) и показателя t-score, которые могут быть дополнены модификацией меры Дайса и т.д. Вы-

числение MI и t-score основано на вычислении частот появления отдельных слов и частоты совместного появления слов, образующих биграмму.

Рис. 1. Построение фраз на основе анализа биграмм из кластера, отнесённого на основании словаря к типу психофизиологических стрессоров, связанных со здоровьем, здравоохранением и социальным окружением



Как видно из таблицы 1, использование модели представления языка BERT совместно с анализом коллграмм позволило увеличить качество распознавания до 93.3%.

Рассмотрим решение задачи кластеризации 1756 комментариев, содержащих описания тревожностей. Следуя работе [2], мы рассматривали следующие типы психосоциологических стрессоров: профессиональные и экономические проблемы, проблемы с ожидаемым жизненным курсом, социальным окружением, образованием, трудоустройством и занятостью, здоровьем и здравоохранением и т.п. Решение задачи кластеризации включало оценивание оптимального количества кластеров путём голосования различных подходов и выделение 7 кластеров на основании BERT+kMeans. Для построенных кластеров с целью выявления описаний и причин тревожных состояний из комментариев был применён метод LDA, позволяющий выделить темы. Ключевые слова для выделенных 4 тем в каждом из кластеров сопоставлялись со словами из словаря, описывающего типы психосоциологических стрессоров.

Анализ коллграмм был использован для выявления списка низкочастотных и сильно связанных слов и высокочастотных слов, после чего из списка удалялись биграммы, не имеющие отношения к словам из словаря, описывающим типы психосоциологических стрессоров. Полученный список биграмм был рас-

ширен биграммами, в которых одно или оба слова имеют негативный сентимент. Для биграмм из построенного списка извлекались смежные биграммы (левая и правая) и выполнялась операция склейки по общим словам. На основе выделенных троек слов был построен граф знаний: левое и правое слова каждой тройки — вершины графа, центральное слово — ребро (рис. 1).

Таким образом, в работе исследовалась эффективность подходов к векторизации текстов, содержащих описания тревожностей, на основе модели представления языка BERT и анализа биграмм. Кроме задачи, бинарной классификации комментариев рассматривалась задача кластеризации комментариев с тревожностями и подходы к выявлению описаний и причин тревожностей для построенных кластеров на основе LDA и анализа коллграмм с учётом слов тематического словаря и негативного сентимента. Для визуализации выделенных описаний тревожностей и их причин использовался граф знаний.

[1] *Turcan E., McKeown K.* Dreddit: A Reddit Datasets for Stress Analysis in Social Media // arXiv: 1911.00133v1, 2019.

[2] *Mowery D., Smith H., Cheney T., Stoddart G., Coppersmith G., Bryan C., Conway M.* Understanding Depressive Symptoms and Psychosocial Stressors on Twitter:A Corpus-Based Study // Journal of Medical Internet Research, 2017.

The Approaches to the Anxiety Disorders Detection based on the Text Mining of Social Media Comments

Dyulicheva Yulia dyulicheva@gmail.com

Simferopol, V. I. Vernadsky Crimean Federal University

The active use of social networks has led to the accumulation of a huge number of comments that users leave. Recently, such a direction of cyberpsychology has been actively developing as the assessment of the mental state based on the analysis of comments in social networks and the influence of various content on the physical and mental health of a person. The anxiety and depressive disorders are some of the most common types of disorders that manifest in the initial stages as negative attitudes and expressions of worry. The changes in the user's attitude and behavior can be monitored based on the analysis of comment texts on social networks, provided that the user does not try to hide his true emotional state through self-control. The altered emotional state and the desire to deliberately distort the meaning affect the use of stable phrases and linguistic indicators of the text, and, therefore, are of interest for improving the quality of recognition algorithms.

Table 1. The evaluation of recognition algorithms

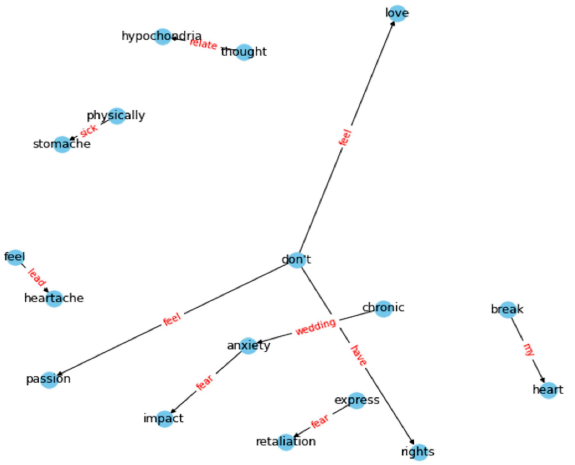
Approaches	Accuracy	Precision	Recall	F1
TF-IDF+RF	0.805	0.857	0.895	0.875
BERT+RF	0.896	0.893	0.893	0.893
Collgram+RF	0.681	0.791	0.522	0.629
TF-IDF+Collgram+RF	**0.864**	**0.861**	**0.902**	**0.881**
BERT+Collgram+RF	**0.933**	**0.931**	**0.948**	**0.939**

In the study, a fragment of balanced binary sample with 3206 comments was considered as a dataset: 1756 comments describing anxiety and 1450 ordinary comments. The original dataset with comments scrapped from the Reddit, and the results of stress analysis is described in [1].

To solve the classification problem, a random forest was used and various approaches to text vectorization were studied: the use of the TF-IDF measure, the BERT language representation model, CollGram analysis indicators, as well as their various combinations. CollGram analysis is an analysis of N-grams based on two indicators: the mutual information (MI) and the t-score, which can be supplemented by a modification of the Dice measure, etc. The calculation of MI and t-score is based on calculating of the frequency of occurrence of individual words and the frequency of co-occurrence of words forming a bigram.

As you can see from the table 1, the use of the BERT language representation model together with the analysis of CollGrams allowed us to increase the recognition quality up to 93.3 %.

Fig. 1. Construction of phrases based on the analysis of bigrams from a cluster classified on the basis of a dictionary as a type of psychophysiological stressors associated with health, healthcare and social environment



Consider the solution to the clustering of 1756 comments containing descriptions of anxieties. Following the [2], we considered the such types of psychosocial stressors as professional and economic problems, problems with the expected life course, social environment, education, employment, health and healthcare, etc. The solution of the clustering included the optimal number of clusters detection by voting between different approaches and identifying of 7 clusters based on BERT + kMeans. For the constructed clusters, in order to identify the descriptions and causes of anxiety states from the comments, the LDA method was applied, which allowed to extract the topics. Key words for the 4 extracted topics in each of the clusters are compared with words from the corpora describing the types of psychosocial stressors.

CollGram analysis was used to identify a list of low-frequency and strongly related words and high-frequency words, after which bigrams were removed from the list that were not related to words from the corpora describing types of psychosocial stressors. The resulting list of bigrams was expanded with bigrams in which one or both words have negative sentiment. For bigrams, adjacent bigrams (left and right) were extracted from the constructed list and the operation of merging was performed according to common words. On the basis of the selected triplets of words, a knowledge graph was created: the left and right words of each triple are the vertices of the graph, the central word is an edge (figure 1).

Thus, the effectiveness of approaches to vectorization of texts containing descriptions of anxieties, based on the BERT language representation model and bigram analysis was investigated. In addition to the challenge of binary classification of

comments, the problem of clustering comments with anxieties and approaches to identifying descriptions and causes of anxieties for the constructed clusters based on LDA and analysis of CollGrams, taking into account the words of the thematic vocabulary and negative sentiment, were considered. The knowledge graph was used to visualize the extracted descriptions of anxieties and their causes.

[1] *Turcan E., McKeown K.* Dreddit: A Reddit Datasets for Stress Analysis in Social Media // arXiv: 1911.00133v1, 2019.

[2] *Mowery D., Smith H., Cheney T., Stoddart G., Coppersmith G., Bryan C., Conway M.* Understanding Depressive Symptoms and Psychosocial Stressors on Twitter:A Corpus-Based Study // Journal of Medical Internet Research, 2017.

Классификация видов активности домашних животных по данным, получаемым с сенсоров носимых устройств

Инякин Андрей Сергеевич[1]⋆ inyakin@forecsys.ru
Мотренко Анастасия Петровна[1] motrenko@forecsys.ru
Руденко Иван[1] rudenko_ivan@forecsys.ru
Кормаков Георгий Владимирович[1,2] kormakov_georgiy@forecsys.ru
Каширин Даниил Олегович[1] kashirin@forecsys.ru
Чипак Егор Олегович[1,3] chipak_egor@forecsys.ru

[1]Москва, ООО «Форексис»
[2]Москва, МГУ имени М.В.Ломоносова
[3]Москва, МФТИ

Классификация активности животных даёт возможность их владельцам осуществлять анализ их поведения и вовремя обращаться к специалистам за лечением. Решение этой задачи может быть полезно для владельцев домашних животных и для анализа здоровья служебных животных.

В последние годы широкое распространение получили исследования активности собак с помощью датчиков ускорения (трёх-осевого акселерометра и гироскопа). Наибольшую эффективность показали методы машинного обучения для классификации активности на данных датчиках.

В работе проанализированы имеющиеся алгоритмы анализа активностей собак с использованием алгоритмов машинного обучения и проведены эксперименты с ансамблированием математических моделей, обученных процедурой бустинга.

Основное внимание уделялось классам активностей животных, выделяемых в уже существующих исследованиях, способам расположения датчиков на собаках, размерам собак в выборке и способам составления математических моделей.

Было принято следующее разбиение на классы активностей: бег, покой и ходьба. Данные классы выбраны из-за поставленной задачи анализа активности животного. Отметим, что в случае постановки задачи в виде мониторинга здоровья используется более широкий набор классов [1].

Важным фактом для исследований являлось расположение датчика на животном. На данных, снятых с блока датчиков на спине у собаки, модели дают качественный прогноз на более широком классе активностей [2]. В рамках собственных экспериментов датчики помещались на ошейнике, обеспечивая достаточную возможность классификации выбранных активностей.

Был осуществлён выбор качественного способа генерации признакового пространства, получаемого из исходных данных носимых устройств. Признаки считались скользящим окном на трёх видах данных: *raw* – сырые данные трёх осей, *abs* – данные энергетической характеристики ряда акселерометра, *pca* – главные компоненты сегмента исходных данных.

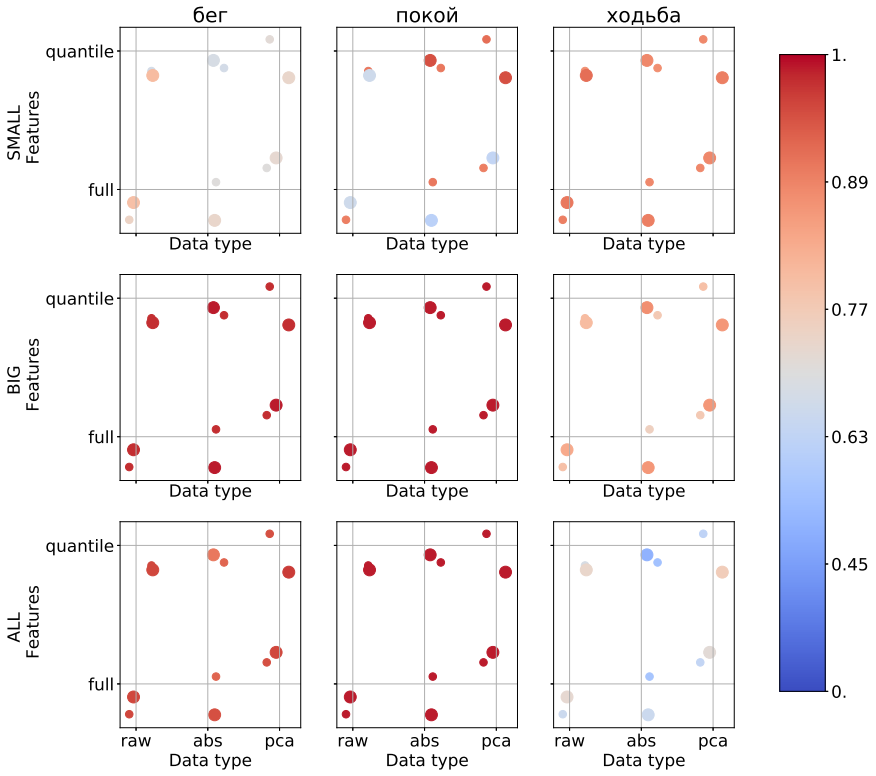


Рис. 1. Сравнение качества для различных стратегий выбора признакового пространства

Исследовались два набора признаков: *full* – статистические, частотные и автокорреляционные признаки, *quantile* – те же и квантильные признаки. Оптимальное признаковое пространство выбрано анализом поведения качества на окнах разной длины (2 и 5 секунд – на рис. 1 изображены малыми и большими кругами) и на собаках разных размеров (маленьких (SMALL), больших (BIG) и всех (ALL)).

На рис. 1 цветом отражены значения метрики F1 для данной комбинации параметров. Сегменты длиной 2 секунды более устойчивы к изменениям остальных структурных параметров. Квантильные признаки и признаки, опирающиеся на главные компоненты, являлись наиболее информативными для моделей

бустинга на практически всех данных и собаках. Итоговые лучшие результаты F1-меры рис. 1 собраны в таблице 1.

Таблица 1. Лучшие значения F1 по классам для всех размеров собак

	бег	покой	ходьба	average
SMALL	0.81	0.95	0.92	0.85
BIG	0.99	0.99	0.88	0.95
ALL	0.97	0.99	0.77	0.90

В результате, определено оптимальное признаковое пространство для собак различных размеров с сохранением высокого качества классификации.

Работа выполнялась с использованием инфраструктуры Центра коллективного пользования «Высокопроизводительные вычисления и большие данные» ФИЦ ИУ РАН (ЦКП «Информатика»).Работа поддержана грантом РФФИ No. 19-07-00885.

[1] *Uijl I., Gómez C., Bartram D., Dror Y., Holland R., Cook A.* External validation of a collar-mounted triaxial accelerometer for second-by-second monitoring of eight behavioural states in dogs // PLoS One, 2017.

[2] *Preston T., Baltzer W., Trost S.* Accelerometer validity and placement for detection of changes in physical activity in dogs under controlled conditions on a treadmill // Res Vet Sci, 2012.

Pets activities classification based on analysis of data obtained from sensors of wearable devices

Inyakin Andrey[1]⋆ inyakin@forecsys.ru
Motrenko Anastasia[1] motrenko@forecsys.ru
Rudenko Ivan[1] rudenko_ivan@forecsys.ru
Kormakov Georgii[1,2] kormakov_georgiy@forecsys.ru
Kashirin Daniil[1] kashirin@forecsys.ru
Chipak Egor[1,3] chipak_egor@forecsys.ru

[1]Moscow, Forecsys
[2]Moscow, Lomonosov Moscow State University
[3]Moscow, Moscow Institute of Physics and Technology

Classification of animal activity allows their owners to analyze their behaviour and contact specialists for treatment in time. Solving this problem could be useful for pet owners and for analyzing the health of service animals.

Wearable devices have been actively used to analyze activity in recent years. Studies of dog activity using acceleration sensors (three-axis accelerometer and gyroscope) have become widespread. Machine learning methods for classifying activity on these sensors have shown the greatest effectiveness.

In this paper, the available algorithms for analyzing dog activities using machine learning algorithms are analyzed and experiments with the ensembling of mathematical models trained by the boosting procedure are carried out.

The main attention was paid to the classes of animal activities identified in existing studies, the ways of placing sensors on dogs, the size of dogs in the sample and the mathematical models' methods producing.

The following division into activity classes was adopted: running, rest and walking. These classes were chosen because of the task of analyzing the activity of the animal. Note that in the case of setting a task in the form of health monitoring, a wider set of classes is used [1].

An important fact for the research was the location of the sensor on the animal. Based on the data taken from the sensor unit on the dog's back, the models give a qualitative forecast for a wider class of activities [2]. As part of our experiments, the sensors were placed on the collar, providing sufficient opportunity to classify the selected activities.

The choice of a qualitative method for generating a feature space obtained from the source data of wearable devices was carried out. The following strategies were used to generate features by a sliding window from the data of a three-axis device: *raw* - on the raw data of the three axes, *abs* - on the data of the energy characteristics of the accelerometer series, *pca* - on the main components of the segment of the initial data.

Two sets of features were studied: *full* - statistical, frequency and autocorrelation features, *quantile* - the same and quantile features. To determine the optimal feature

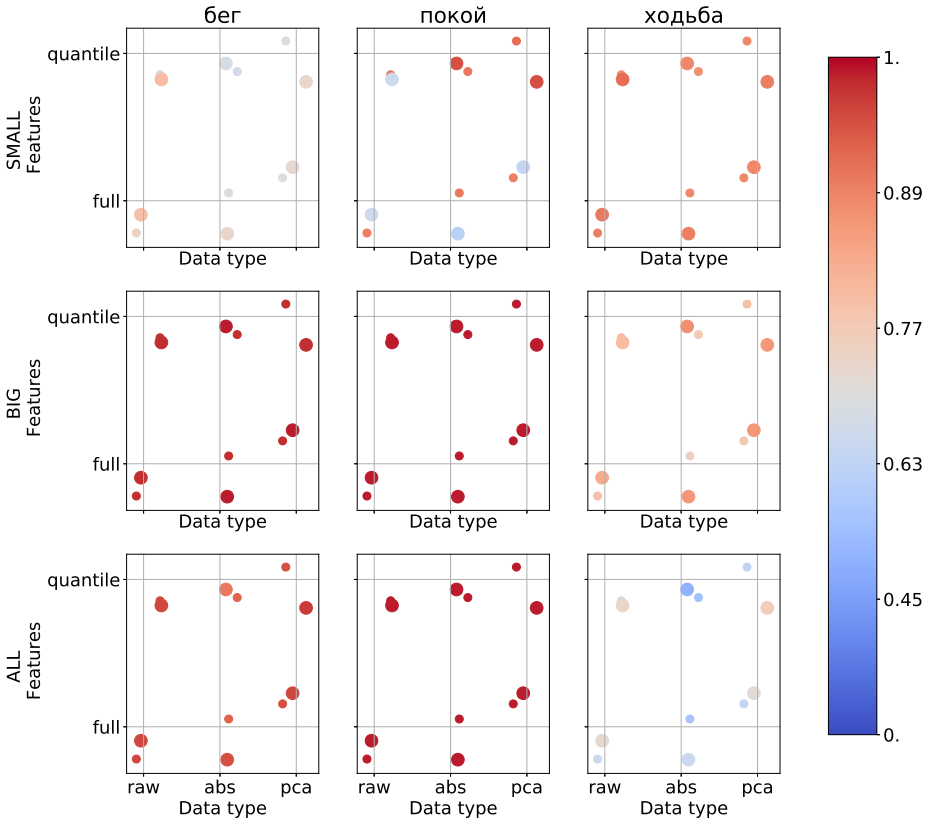


Fig. 1. Quality comparison for different feature space selection strategies

space, the analysis of quality behaviour was carried out on windows of different lengths (2 and 5 seconds - shown in Fig. 1 with small and large circles) and on dogs of different sizes (small (SMALL), large (BIG) and all (ALL)).

Figure 1 shows the values of the F1 metric for this combination of parameters in colour. Segments 2 seconds long are more resistant to changes in other structural parameters.

Also, the evaluation of the significance of the parameters led to the fact that quantile features and features based on the main components were the most infor-

mative for boosting models on almost all data and dogs. The final best results of the F1 measure of Fig. 1 are collected in Table 1.

Table 1. The best F1 values by class for all dog sizes

	running	rest	walking	average
SMALL	0.81	0.95	0.92	0.85
BIG	0.99	0.99	0.88	0.95
ALL	0.97	0.99	0.77	0.90

As a result, in the course of the research, the optimal feature space for dogs of various sizes was determined while maintaining high-quality classification.

The research was carried out using the infrastructure of the shared research facilities ≪High Performance Computing and Big Data≫ of FRC CSC RAS (CKP ≪Informatics≫).

This research is funded by RFBR, grant 19-07-00885.

[1] *Uijl I., Gómez C., Bartram D., Dror Y., Holland R., Cook A.* External validation of a collar-mounted triaxial accelerometer for second-by-second monitoring of eight behavioural states in dogs // PLoS One, 2017.

[2] *Preston T., Baltzer W., Trost S.* Accelerometer validity and placement for detection of changes in physical activity in dogs under controlled conditions on a treadmill // Res Vet Sci, 2012.

Разработка системы детекции аномалий с целью автоматизации визуального контроля поверхности листового металлопроката

Мортин Константин Владимирович[1] kvmortin@mail.ru

[1]Муром, Муромский институт федерального государственного бюджетного образовательного учреждения высшего образования Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых

На металлургических предприятиях страны остро стоит вопрос о выпуске продукции листового металлопроката без существенных дефектов или с оптимальными параметрами допуска без отбраковки. На таких предприятиях создаются отделы или лаборатории по техническому зрению. Процент дефектов, видимых на поверхности листового проката, который видят в лаборатории контроля по металлу составляет 84 процента, а из-за деформаций 16 процентов. Цель таких лабораторий - исследовать цифровые изображения листового металлопроката с различными видами поверхностных дефектов, своевременно их обнаружить и принять соответствующие мероприятия по ликвидации или минимизации их возникновения. Дефекты поверхности готового проката могут возникнуть при плавке металла и прокатке, а также вследствие нарушения технологии отделки проката. Площадь сталеплавильных дефектов при прокатке увеличивается прямо пропорционально общей вытяжке раската. Особенно велики потери на переделах производства специальных сталей и сплавов, где только при зачистке потери годного металла составляют 20 и более процентов. Изыскание методов борьбы с возникновением дефектов поверхности металла является предметом постоянного изучения металлургических лабораторий и по мере решения проблем снижения объема дефектного слоя должно вестись в тесной связи с металлообрабатывающими лабораториями, роль которых постоянно вырастает по мере снижения черновых операций и роста чистовой финишной обработки поверхности проката, вплоть до его полирования при массовом производстве. В заводской практике определение дефектов ведется визуальным способом на основе опыта персонала. Практически на каждом металлургическом заводе существует своя специфика определения дефектов и терминология, что отрицательно влияет на организацию научно обоснованной технологии отделки листового металлопроката. А также в заводских условиях уже применяются следующие подходы к обнаружению дефектов на листовом металлопрокате на основе фильтра Габора, вейвлет- преобразований, систем нечеткой логики, различных методов сегментации и бинаризации, алгоритмов адаптивного усиления, оптико- электронные информационно измерительные системы (ОЭИИС) оценки качества продукции, методы неразрушающего контроля, алгоритм SURF. Рассмотренные методы обработки цифровых изображений, применяемые для обнаружения дефектов на листовом металлопрокате, показывают быструю обработку изображений. Но их применение к дефектам листового металлопро-

ката имеет ряд недостатков: ограниченность функциональных возможностей и недостаточная эффективность при выделении контура дефекта из-за загрязнения исследуемой области на изображении, сглаживание изображения дефекта с помощью медианной фильтрации приводит к устранению границ дефекта и его зашумления,яркость фона изображения листового металлопроката занимает максимально допустимый диапазон, а яркость самого дефекта и его важных участков занимает узкий диапазон, что как следствие ухудшает контраст всего исследуемого изображения. Исходя из вышеперечисленных фактов, существует необходимость в разработке методов и алгоритмов улучшения и быстроты распознавания дефектов на изображениях листового металлопроката, что порождает необходимость выполнения локальных преобразований исследуемых изображений путем комплексного сегментирования, фильтрации, шумоподавления и использования комбинированного математического аппарата. Для решения поставленных задач проектируется система предварительной обработки полутоновых дефектоскопических изображений с чередованием линейной свертки, краевой фильтрации и яркостной сегментации. Такая архитектура позволяет обнаруживать дефект и удалять его фон, делая попиксельно яркие контура обнаруженного дефекта на изображении. Разработанная система состоит из следующей последовательности: 1. Входное дефектоскопическое изображение; 2. Построение вертикальной проекции полутонового изображения; 3. Линейная свертка (фильтр подчеркивание краев); 4. Линейная свертка с двумя масками; 5. Линейная свертка (фильтр высокая частота); 6. Краевая фильтрация; 7. Яркостная сегментация (спектр разделимости); 8. Построение вертикальной проекции изображения; 9. Попиксельно исключающее И (краевая фильтрация, операция xor, линейная свертка); 10. Выделенный дефект с контуром без фона; 11. Запись промежуточных результатов в базу данных. Такая система позволит повысить качество цифровой обработки дефектоскопических изображений в системах технического зрения за счет разработки новых методов, основанных на использовании комплексной свертки совместно с семантической сегментацией и теорией распознавания и детектирования полутоновых изображений. Предлагаемая новая модель предварительного улучшения дефектоскопических изображений, основанная на многопоточной свертке, семантической сегментации и препарировании, позволит разработать новые алгоритмы обработки дефектоскопических изображений, отличающихся от существующих с возможностью использования на производстве листового металлопроката и накоплением базы данных исходных дефектов и их актуального анализа. Экономическая целесообразность проекта – внедрение быстродействующего алгоритма в предобработку дефектоскопических изображений перед классификатором. Полученный метод можно применять в качестве детектора аномалий для построения датасетов, чтобы не пропустить не один дефект поверхности листового металлопроката. Эффективность от такого подхода будет заключаться в снижении затрат и времени на покупку или создание систем распознавания и детектирования. Прак-

тическую апробация планируется проводить на действующих производствах АО ВМЗ в рамках лаборатории компьютерного зрения.

[1] *Мортин К.В.,Привезенцев Д.Г.* Разработка сверточного слоя нейронной сети для обнаружения дефектов листового металлопроката на дефектоскопических изображениях // Международная конференция по мягким вычислениям и измерениям, 2021. С. 210–212.

Development of an anomaly detection system for the purpose of automating visual inspection of the surface of sheet metal

Mortin Konstantin[1] kvmortin@mail.ru

[1] Murom, Murom institute of the federal state budgetary educational institution of higher education Vladimir university named after Alexander Grigoryevich and Nikolai Grigoryevich Stoletov

At the metallurgical enterprises of the country, the issue of the production of sheet metal products without significant defects or with optimal tolerance parameters without rejection is acute. Departments or laboratories for technical vision are created at such enterprises. The percentage of defects visible on the surface of sheet metal, which is seen in the metal control laboratory is 84, and due to deformations 16. The purpose of such laboratories is to investigate digital images of sheet metal with various types of surface defects, detect them in a timely manner and take appropriate measures to eliminate or minimize their occurrence. Defects in the surface of the finished rolled product may occur during metal melting and rolling, as well as due to violations of the finishing technology of rolled products. The area of steelmaking defects during rolling increases in direct proportion to the total extraction of the roll. Losses are especially high at the processing stages of the production of special steels and alloys, where only during stripping losses of usable metal amount to 20 percent or more. The search for methods to combat the occurrence of metal surface defects is the subject of constant study of metallurgical laboratories and, as the problems of reducing the volume of the defective layer are solved, it should be carried out in close connection with metalworking laboratories, whose role is constantly growing as roughing operations decrease and the growth of finishing finishing of the rolled surface, up to its polishing in mass production. In factory practice, defects are determined visually based on the experience of the staff. Almost every metallurgical plant has its own specifics of determining defects and terminology, which negatively affects the organization of a scientifically based technology for finishing sheet metal. And also in factory conditions, the following approaches are already being used to detect defects on sheet metal based on the Gabor filter, wavelet transformations, fuzzy logic systems, various segmentation and binarization methods, adaptive amplification algorithms, optoelectronic information measuring systems (OEIIS) for product quality assessment, non-destructive testing methods, the SURF algorithm. The considered methods of digital image processing used to detect defects on rolled metal sheets show fast image processing. But their application to sheet metal defects has a number of disadvantages: -limited functionality and insufficient efficiency in highlighting the contour of the defect due to contamination of the area under study in the image; -smoothing the defect image using median filtering leads to the elimination of the boundaries of the defect and its noise; -the brightness of the background image of sheet metal occupies the maximum allowable range, and the brightness of the defect itself and its important areas

occupies a narrow range, which consequently worsens the contrast of the entire image under study. Based on the above facts, there is a need to develop methods and algorithms to improve and speed the recognition of defects in images of sheet metal, which creates the need to perform local transformations of the studied images by complex segmentation, filtering, noise reduction and the use of a combined mathematical apparatus. To solve the tasks set, a system of preprocessing of halftone flaw detection images with alternating linear convolution, edge filtering and brightness segmentation is being designed. This architecture allows you to detect a defect and remove its background, making pixel-by-pixel bright contours of the detected defect in the image. The developed system consists of the following sequence: 1. Input flaw detection image; 2. Building a vertical projection of a halftone image; 3. Linear convolution (edge underlining filter); 4. Linear convolution with two masks; 5. Linear convolution (high frequency filter); 6. Edge filtering; 7. Brightness segmentation (separability spectrum); 8. Building a vertical projection of the image; 9. Pixel-by-pixel exclusive And (edge filtering, xor operation, linear convolution); 10. Highlighted defect with contour without background; 11. Recording intermediate results in the database. Such a system will improve the quality of digital processing of flaw detection images in technical vision systems by developing new methods based on the use of complex convolution together with semantic segmentation and the theory of recognition and detection of halftone images. The proposed new model of preliminary improvement of flaw detection images, based on multithreaded convolution, semantic segmentation and preparation, will allow the development of new algorithms for processing flaw detection images that differ from existing ones with the possibility of using in the production of sheet metal and the accumulation of a database of initial defects and their current analysis. The economic feasibility of the project is the introduction of a high-speed algorithm in the preprocessing of flaw detection images before the classifier. The resulting method can be used as an anomaly detector for building datasets, so as not to miss more than one defect on the surface of sheet metal. The effectiveness of this approach will be to reduce the cost and time for the purchase or creation of recognition and detection systems. Practical testing is planned to be carried out at the existing facilities of JSC VMZ within the computer vision laboratory.

[1] *Mortin K., Privezentsev D.* Development of a convolutional layer of a neural network for detecting defects of sheet metal on flaw detection images // International Conference on Soft Computing and Measurements, 2021. Pp. 210–212.

Исследование применимости использования информации о состоянии канала передачи данных для организации позиционирования внутри помещений

Астафьев Александр Владимирович[1]⋆ Alexandr.Astafiev@mail.ru
Жизняков Аркадий Львович[1] Lvovich1975@mail.ru
Демидов Антон Александрович[1] AADemidov@list.ru
Кондрушин Илья Евгеньевич[1] IEKondrushin@list.ru

[1]Владимир, Владимирский государственный университет имени Александра Григорьевича и Николая Григорьевича Столетовых

Позиционирование и навигация плотно вошли в нашу повседневную жизнь: навигаторы, такси, разведка, беспилотные аппараты и многое другое. Проблема позиционирования и навигации на открытом пространстве решены за счёт технологий ГЛОНАСС и GPS, а вот системы позиционирования и навигации внутри помещений все еще имеют большую погрешность вычислений, которые могут составлять более 3 метров. Повышение точности позиционирования внутри помещений позволит создавать высокоточные системы навигации. К таким системам могу относится системы управления беспилотными аппаратами, роботизированными платформами и группами роботов, системы автоматизации навигации и эвакуации стратегических и развлекательных объектов и т.д. Исходя из этого совершенствование фундаментальных основ построения систем автономной навигации внутри помещений в режиме реального времени является актуальной научно-технической задачей.

Целью работы является исследование применимости использования информации о состоянии канала передачи данных для организации позиционирования внутри помещений.

Несмотря на большое количество исследований в области навигации в закрытых помещениях всё равно остаётся большое количество нерешенных вопросов. Необходима разработка более точных, быстрых и масштабируемых алгоритмов. Точность зависит от среды распространения сигнала, наличия физических преград и интерференции. В данном направлении исследования направлены на использование шумоподавляющих фильтров, искусственных нейронных сетей и подходов дообучения в процессе эксплуатации. Перспективным направлением повышение точности является исследования радиосигнала по каналам channel state information (CSI).

CSI – это, информация которая описывает, как сигнал распространяется от передатчика к приемнику. В беспроводных сетях, использующих OFDM модуляцию, CSI представляет собой комплексно число, в котором содержится информация об амплитуде и фазе сигнала. Пакет данных CSI содержит набор комплексных чисел, соответствующей каждой поднесущей.

В настоящем исследовании рассматриваются беспроводные сети с частотой пропускания в 20 МГц. Пакет данных CSI в этом случае представляет собой

сложный массив данных MxNx56, где M-количество передающих антенн, N – количество принимаемых антенн, а 56 – количество поднесущих для полосы пропускания.

Работа с CSI пакетами является перспективным направлением научных исследований, т.к. позволяет получать описание канала связи не одним целым числом, как это принято в методе RSSI, а целым набором параметров. Количество анализируемых параметров в одном пакете CSI составляет 448.

В ходе исследования планируется использование информации об изменении амплитуды и фазы сигнала по отдельным поднесущим для определения удалённости считывающего устройства, а также построения более точных системы позиционирования внутри помещений.

Работа поддержана грантом РНФ No. 21-71-00133.

[1] *Астафьев А. В., Титов Д. В., Жизняков А. Л., Демидов А. А.* Метод позиционирования мобильного устройства с использованием сенсорной сети BLE-маяков, аппроксимации значений уровней сигналов RSSI и искусственных нейронных сетей // Компьютерная оптика, 2021. Т. 45(2). С. 277–285.

Investigation the applicability of using channel state information for indoor positioning

Astafiev Alexandr[1]⋆ Alexandr.Astafiev@mail.ru
Zhiznyakov Arkady[1] Lvovich1975@mail.ru
Demidov Anton[1] AADemidov@list.ru
Kondrushin Ilya[1] IEKondrushin@list.ru
[1]Vladimir, Vladimir State University

Positioning and navigation have become part of our daily life: navigators, taxis, reconnaissance, drones and much more. The problem of positioning and navigation in open space has been solved due to GLONASS and GPS technologies, but indoor positioning and navigation systems still have a large calculation error, which can be more than 3 meters. Improving indoor positioning accuracy will allow the creation of high-precision navigation systems. Such systems can include control systems for unmanned vehicles, robotic platforms and groups of robots, automation systems for navigation and evacuation of strategic and entertainment facilities, etc. Proceeding from this, improving the fundamental foundations of building autonomous navigation systems inside buildings in real time is an urgent scientific and technical task.

The aim of the work is to study the applicability of using channel state information for indoor positioning.

Despite the large amount of research in the field of indoor navigation, there are still a large number of unresolved issues. The development of more accurate, faster and scalable algorithms is needed. Accuracy depends on signal propagation environment, physical obstructions and interference. In this direction, research is aimed at using noise-canceling filters, artificial neural networks and additional training approaches during operation. A promising direction for improving the accuracy is the study of the radio signal via channel state information (CSI) channels.

CSI is information that describes how a signal travels from a transmitter to a receiver. In wireless networks using OFDM modulation, CSI is a complex number that contains information about the amplitude and phase of the signal. The CSI data packet contains a set of complex numbers corresponding to each subcarrier.

This study considers wireless networks with a bandwidth of 20 MHz. The CSI data packet in this case is a complex data array MxNx56, where M is the number of transmit antennas, N is the number of received antennas, and 56 is the number of subcarriers for the bandwidth.

Working with CSI packages is a promising area of scientific research, because allows you to get a description of a communication channel not with one integer, as is customary in the RSSI method, but with a whole set of parameters. The number of analyzed parameters in one CSI package is 448.

In the course of the study, it is planned to use information about the change in the amplitude and phase of the signal for individual subcarriers to determine the

remoteness of the reading device, as well as to build more accurate indoor positioning systems.

This research is funded by Russian Science Foundation, grant 21-71-00133.

[1] *Astafiev A, Titov D, Zhiznyakov A, Demidov A* A method for mobile device positioning using a sensor network of BLE beacons, approximation of the RSSI value and artificial neural networks. // Computer Optics, 2021. Vol. 45(2). Pp. 277–285.

Использование мезоскопической модели для моделирования транспортных потоков на МКАД и управления въездами

Старожилец Всеволод Михайлович[1]⋆ starvsevol@gmail.com
Чехович Юрий Викторович[1] chehovich@forecsys.ru

[1]Вычислительный центр им. А. А. Дородницына ФИЦ ИУ РАН, Москва, ул. Вавилова, 40, Россия, 119333

В работе рассматривается проблема математического моделирования автомобильных транспортных потоков в рамках транспортных систем значительных масштабов. Непосредственно рассматривается задача применения математической модели, описанной авторами в [1], к моделированию конкретной транспортной магистрали — Московской кольцевой автомобильной дороги (МКАД).

На сегодняшний день моделирование крупных транспортных сетей представлено в работах [2, 3] в виде примеров применения существующих программных пакетов таких как SUMO (Simulation of Urban Mobility), iTETRIS ("An Integrated Wireless and Traffic Platform for Real-Time Road Traffic Management Solutions") и других. Детальное описание подхода к моделированию автомагистрали в данных пакетах зачастую отсутствует.

Моделирование транспортных потоков на автомагистрали тесно связано с задачей оптимизации светофорного управления в транспортной сети. Однако, в большинстве работ посвящённых светофорному управлению на перекрестке, не ставится задача обеспечения максимальной пропускной способности выделенной автомагистрали, и как следствие все дороги считаются равнозначными. В данной работе же ставится как раз задача обеспечения наиболее свободного проезда на выделенной ключевой магистрали.

В [1] предложена математическая модель, свойства которой были исследованы на модельных элементарных фрагментах транспортной сети. Эта модель создавалась для расчетов пропускной способности в различных режимах работы транспортных графов значительного масштаба, включающих тысячи сегментов и имеющих протяженность десятки километров. Модель основана на оригинальном мезоскопическом подходе, оперирующем в качестве объектов моделирования группами автомобильных транспортных средств (АТС), объединяющими автомобили со сходными параметрами, находящимися на одном сегменте транспортного графа. Скорость групп автомобилей рассчитывается с помощью фундаментальной диаграммы поток-плотность на магистрали [4]. Такой подход позволяет быстро обсчитывать достаточно большие транспортные сети, в том числе такую магистраль, как МКАД, что необходимо для решения оптимизационных задач, для которых проводится моделирование.

Данный подход отличается от двух классических направлений к моделированию транспортных потоков, представленных микроскопическим подходом, основанным на моделировании движения каждого отдельного автомобиля [5], и

макроскопическим подходом, опирающимся на сходство движения АТС с жидкостью или газом [6].

В работе строится модель одной из сторон МКАД с крупными въездами и съездами с неё. Ввиду недостаточных объемов реальных данных в работе на въездах на автомагистраль используются модельные построенные на основе имеющихся данных с дорожных датчиков на некоторых из въездов и статистических данных ЦОДД. Поскольку утром большая часть автомобилистов хотят въехать в Москву, а вечером, соответственно, выехать из города, то мы получаем модельные данные двух типов — с утренней пиковой загрузкой и с вечерней. Проводится моделирование поведения автомагистрали с различным модельными данными на въездах и сравниваются результаты с контролем на въездах и без него. Для проверки гипотезы об эффективности управления въездами рассчитываются временные потери на проезд по МКАД за день, а также число автомобилей, проехавших по автомагистрали.

Работа поддержана грантом РФФИ No. 20-07-01057 А.

[1] *Старожилец В. М, Чехович Ю. В.* Об одном подходе к статистическому моделированию транспортных потоков // Журнал Вычислительной математики и математической физики, 2021.

[2] *Yuta A., Nobuyasu I., Hajime I., Tetsuo I., Uchitane T.* Traffic simulation of Kobe-city // Proceedings of the international conference on social modeling and simulation, plus Econophysics Colloquium 2014, 2015. Vol. 229. Pp. 255–264.

[3] *Bieker L., Krajzewicz D., Morra A., Michelacci C., Cartolano F.* Traffic simulation for all: a real world traffic scenario from the city of Bologna // Modeling Mobility with Open Data, 2015. Vol. 229. P. 47–60.

[4] *Алексеенко А. Е., Холодов Я. А., Холодов А. С., Горева А. И., Васильев М. О., Чехович Ю. В., Мишин В. Д., Старожилец В. М.* Разработка, калибровка и верификация модели движения трафика в городских условиях. Ч. I // Компьютерные исследования и моделирование, 2015. Т. 7(6). С. 1185–1203.

[5] *Гасников А. В. и др.* Введение в математическое моделирование транспортных потоков // М.: Litres, 2015. 89 с.

[6] *Whitham J.* Linear and nonlinear waves // Wiley, 1974. 656 p.

About use of the mesoscopic model for traffic flows modeling on the Moscow Ring Road and enters control

Starozhilets Vsevolod[1]⋆ starvsevol@gmail.com
Chekhovich Yury[1] chehovich@forecsys.ru
[1]Dorodnicyn Computing Centre FRC CSC RAS, Vavilov st. 40, 119333 Moscow, Russia

The paper deals with the problem of mathematical modeling of road traffic flows within transport systems of significant scales. In this work we apply the mathematical model described by the authors in [1] to modeling a specific transport highway — the Moscow Ring Road (MRR).

At this moment, modeling of large transport networks presented in the works [2, 3] in form of examples of using existing software packages such as SUMO (Simulation of Urban Mobility), iTETRIS ("An Integrated Wireless and Traffic Platform for Real-Time Road Traffic Management Solutions") and others. A detailed description of the highway modeling approach in these packages is often missing.

Traffic flow modeling on a highway is closely related to the task of optimizing traffic light control in a transport network. However, in most works devoted to traffic light control at an intersection, the task of ensuring the maximum throughput of a dedicated highway is not set, and as a result, all roads are considered equal. In this work, the task is precisely to ensure maximum traffic flow on a dedicated highway.

In [1] a mathematical model is proposed which properties were investigated on elementary fragments of the transport network. This model was created for calculating traffic flow in various modes of operation of transport graphs of a significant scale, including thousands of segments with a length of tens kilometers. The model is based on the original mesoscopic approach, which operates as objects of modeling by groups of automobile vehicles, combining cars with similar parameters located on the same segment of the transport graph. The speed of groups of cars is calculated using the fundamental flow-density diagram on the highway [4]. This approach makes it possible to quickly calculate sufficiently large transport networks, including such a highway as the Moscow Ring Road, which is necessary to solve optimization problems for which modeling is carried out.

This approach differs from the two classical approaches to modeling traffic flows, represented by a microscopic — based on modeling the movement of each individual car [5], and a macroscopic — based on the similarity of vehicle movement with a liquid or gas [6].

In this work we create a model of one of sides of the Moscow Ring Road with major entrances and exits from it. Due to the insufficient amount of real data at the entrances to the motorway we create model data for Moscow Ring Road entrances based on the available information from road traffic detectors at some of entrances and statistical data from Moscow Traffic management center. Since most of the people want to enter Moscow in the morning and, accordingly, leave the city in the evening, we get model data of two types — with morning peak load and with

evening one. Modeling the behavior of the highway with different model data at the entrances is carried out and the results are compared with and without control at the entrances. To test the hypothesis about the efficiency of entry traffic flow control, time losses for travel along the Moscow Ring Road per day are calculated, as well as the number of cars that have driven along the highway.

This research is funded by RFBR, grant 20-07-01057 A.

[1] *Starozhilets V., Chekhovich Y.* About one approach to traffic flows statistical modeling // Computational Mathematics and Mathematical Physics, 2021. Vol. 61(7).

[2] *Yuta A., Nobuyasu I., Hajime I., Tetsuo I., Uchitane T.* Traffic simulation of Kobecity // Proceedings of the international conference on social modeling and simulation, plus Econophysics Colloquium 2014, 2015. Vol. 229. Pp. 255–264.

[3] *Bieker L., Krajzewicz D., Morra A., Michelacci C., Cartolano F.* Traffic simulation for all: a real world traffic scenario from the city of Bologna // Modeling Mobility with Open Data, 2015. Vol. 229. P. 47–60.

[4] *Alekseenko A., Kholodov Y., Kholodov A., Goreva A., Vasiliev M., Chekhovich Yu., Mishin V., Starozhilets V.* Development, calibration and verification of the traffic movement model in urban environments // Computer research and modeling, 2015. Vol. 7(6). Pp. 1185–1203.

[5] *Gasnikov A. et al* Vvedenie v matematicheskoe modelirovanie transportnykh potokov // M.: Litres, 2015. 89 p.

[6] *Whitham J.* Linear and nonlinear waves // Wiley, 1974. 656 p.

Информативные образы нестационарных цифровых сигналов в диагностических системах

Грызлова Татьяна Павловна ktntpgryzlova@mail.ru
Рыбинск,РГАТУ им. П. А. Соловьева

Диагностика состояния сложной технической системы по цифровым диагностическим сигналам и распознавание цифровых сигналов выполняются, как правило, после их обработки, в результате которой длинные последовательности N отсчетов $\mathbf{s}_0^{N-1}$ редуцируются к одномерным или n-мерным образам $\mathbf{x}$.

$$\mathbf{s}_0^{N-1} \to \mathbf{x} = \left|\mathbf{f}_1\left(\mathbf{s}_0^{N-1}\right) \quad \ldots \quad \mathbf{f}_n\left(\mathbf{s}_0^{N-1}\right)\right|$$

Далее задачу можно решать известными методами математической теории распознавания, используя матрицу образов $\mathbf{X}$, дополненную характеристическим вектором χ. Выбор алгоритмов обработки сигналов можно сделать на основе оценок информативности $\mathbf{I}(\mathbf{X},\ \chi)$, учитывающих размещение образов в пространстве образов. В качестве критерия потенциальной точности диагностики выбрано отношение среднего межклассового расстояния к среднему внутриклассовому. Ошибки диагностики часто связаны с обработкой нестационарных сигналов как стационарных, без учета изменчивости их образов на разных временных интервалах. Общие методы обработки нестационарных сигналов отсутствуют, для их анализа требуется разработка специальных методов.

Первый подход, представленный в работе – обработка сигналов большими технологическими блоками $G_m, m = 1..M$, соизмеримыми с предполагаемыми размерами сегментов или равными известному размеру цикла регистрации B. На его основе модифицированы системы диагностики подшипников трансмиссии ГТД на базе ИВУ-1М, что повысило информативность с 0,69 до 1,1. Второй подход – метод анализа полуволн (HW-анализ), предложенный и исследованный автором при решении задач диагностики, сегментации и распознавания последовательности состояний сложного источника.

При типовой блочной обработке сигнал раскладывается на $K = N/b$ технологических блоков g_k длины b:

$$\mathbf{s}_0^{N-1} = \mathbf{s}_0^{b-1} \cdot \mathbf{s}_b^{2b-1} \cdot \ldots \cdot \mathbf{s}_{(K-1)b}^{Kb-1}$$

Размер блоков не бывает очень большим, обычно его подбирают в каждой конкретной задаче обнаружения локальных неоднородностей, блочного кодирования или сегментации. Сигнал раскладывается на технологические блоки:

$$\mathbf{s}_0^{N-1} = g_0 \cdot \ldots \cdot g_{K-1}, k = 0..(K-1), N \gg K \gg M$$

Каждый технологический блок определен на соответствующем технологическом интервале на оси времени: $[(m-1)B, mB-1]$ или $[kb, (k+1)b-1]$.

Принципиально иное решение – разложение сигнала на его естественные образующие. Это последовательности отсчетов сигнала одного знака – положительные hw_i и отрицательные hwn_i полуволны, их количество заранее неизвестно. Пусть $t0_i$ – момент i-того пересечения сигналом нулевого уровня в положительном направлении, DT_i, DTn_i – длительности положительных и отрицательных полуволн, соответственно. Положительные полуволны определены на физических интервалах $[t0_i, t0_i + DT_i - 1]$, отрицательные - $[t0_i + DT_i = t0n_i, t0n_i + + DTn_i - 1]$. Полуволны можно сортировать по длительности и подобию. На рис. 1 показаны полуволны и блоки представителей классов диагностических сигналов кондиционных (**C**), кондиционных, необоснованно снятых с эксплуатации, (**N**) и неисправных (**B**) подшипников трансмиссии ГТД. Видно, что

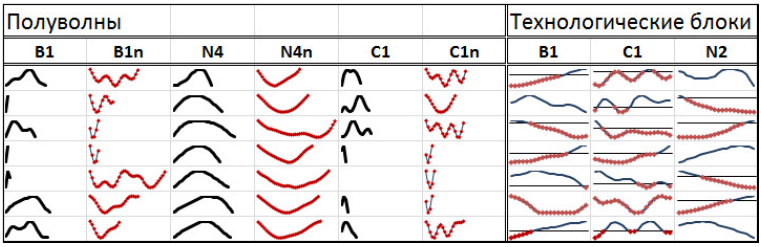


Рис. 1. Полуволны и блоки сигналов подшипников трансмиссии ГТД

представление в виде полуволн сохраняет индивидуальные характерные свойства и особенности сигнала, а формальное по сущности представление в виде блоков эту информацию нивелирует. Метод реализован как комплекс алгоритмов кластеризации элементов сигнала (полуволн) как для постобработки, так и для реального времени. Развитие метода идет по направлению анализа статистики полуволн и их характеристик как по сигналу, так и по технологическим интервалам. В простых случаях стабильных условий измерений характеристики эмпирических распределений полуволн по кластерам полуволн одной длительности и даже просто количество полуволн имеют информативность, близкую к 3. В задаче диагностики подшипников трансмиссии ГТД условия измерений неблагоприятны (аэродром), необходим этап кластеризации полуволн по подобию и анализ потока новых эталонов, или динамический анализ размеров кластеров, или ранговая фильтрация последовательностей длительностей полуволн. В последнем случае мы имеем последовательность трехмерных образов, привязанных к моментам появления полуволн, Δ - апертура фильтров, $\mathbf{\Phi}_x$ - выход максимального фильтра, $\mathbf{\Phi}_y$- выход минимального фильтра.

$$\mathbf{x}^* = \left| t0_i \quad \mathbf{\Phi}_x \left(DT_i^{i+\Delta-1} \right) \quad \mathbf{\Phi}_y \left(DT_i^{i+\Delta-1} \right) \right|^*$$

На рис. 2 показана последовательность образов сигнала «_mama_».

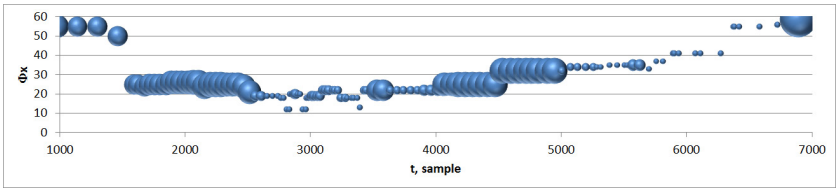


Рис. 2. Образы полуволн сигнала «_mama_». Размер шариков - $\mathbf{\Phi}_y$

Разработаны критерии для сегментации и распознавания состояний сложных источников на основе локальной ранговой фильтрации со скользящей апертурой последовательностей длительностей полуволн, позволяющие решать задачи сегментации, а при ограниченном алфавите состояний выполнять распознавание состояния, в котором такие полуволны могли бы появиться.

The informative Images of the non-stationary digital signals in the systems of diagnostics

Gryzlova Tatyana ktntpgryzlova@mail.ru
Rybinsk, RSATU named after P. A. Soloviev

As a rule State's diagnostics of complex technical system via digital diagnostic signals and signal's recognition are being performed after their processing, as a result of which long sequences of N samples $\mathbf{s}_0^{N-1}$ are reduced to one-dimensional or n-dimensional patterns $\mathbf{x}$:

$$\mathbf{s}_0^{N-1} \to \mathbf{x} = \left|\mathbf{f}_1\left(\mathbf{s}_0^{N-1}\right) \quad \ldots \quad \mathbf{f}_n\left(\mathbf{s}_0^{N-1}\right)\right|$$

Further, the problem can be solved by known methods of mathematic theory of the recognition with using pattern matrix $\mathbf{X}$, supplemented by a characteristic vector χ. The reasonable choice of signals processing algorithm can be done on the base of the estimates of informativeness $\mathbf{I}(\mathbf{X},\ \chi)$ that take into account the placement of images in the image space. As criteria of the potential quality of diagnostics with a fixed signal processing algorithm the ratio of the average inter-class distance to the average intra-class distance is used. Diagnostic errors are often associated with the processing of non-stationary signals as stationary without taking into account the variability of their images at different time intervals. A common method of signal processing in technical diagnostics and speech recognition systems is spectral analysis, although signals can be significantly non-stationary. The main problem is that there are no general methods for processing non-stationary signals, and the development of special methods is required for the analysis of non-stationary signals. The first approach presented in the paper is signal processing on large technological blocks $G_m, m = 1..M$ that are commensurate with the supposed size of the segments or equal to the e known size of the registration cycle B. Based on it, diagnostic systems of the GTE transmission bearing with IVU-1M were modified, that help to increase the informativeness from 0.69 to 1.1. Another approach - the method of half-wave analysis (HW- analysis) proposed and investigated by the author when solving problems of diagnostics, segmentation and recognition of a sequence of the states of a complex source.

The typical block processing is usually associated with forced partitioning into technological blocks. The signal is decomposed into $K = N/b$ technological blocks g_k with block length b:

$$\mathbf{s}_0^{N-1} = \mathbf{s}_0^{b-1} \cdot \mathbf{s}_b^{2b-1} \cdot \ldots \cdot \mathbf{s}_{(K-1)b}^{Kb-1}$$

The size of the blocks is not very large, because it is a technological block for detecting local inhomogeneities, block coding, segmentation. Usually, the sizes of b blocks are selected for each specific task.

$$\mathbf{s}_0^{N-1} = g_0 \cdot \ldots \cdot g_{K-1}, k = 0..(K-1), N \gg K \gg M$$

Each technological unit is defined on the corresponding technological interval on the time axis $[(m-1)B, mB-1]$ or $[kb, (k+1)b-1]$.

The fundamentally basic solution is to decompose the signal into elements naturally included in it (splitting of a signal into its natural generators). These elements are sequences of samples of a signal of one sign – positive (hw_i) and negative (hwn_i) half-waves, their number is unknown in advance. Let $t0_i$ – the point of the i-th crossing of the zero level signal in the positive direction, DT_i, DTn_i – the durations of positive and negative half-waves, respectively. Positive half-waves are defined at physical intervals $[t0_i, t0_i + DT_i - 1]$, and negative half-waves are defined at $[t0_i + + DT_i = t0n_i, t0n_i + DTn_i - 1]$. Half-waves can be sorted by duration and similarity. Fig. 1 shows half-waves and blocks of representatives of classes of diagnostic signals of serviceable (**C**), serviceable, unreasonably decommissioned bearings, (**N**) and faulty GTE transmission bearings (**B**). It can be seen that the representation in the

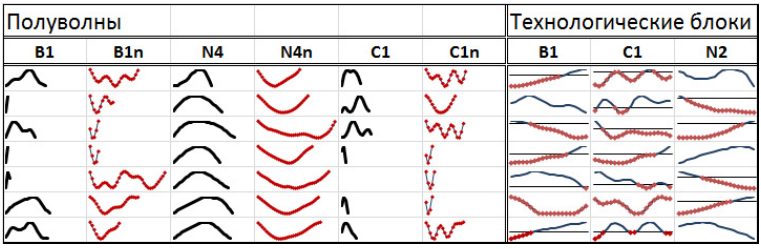


Fig. 1. The half-waves and the technological blocks of the signals of transmission bearing

form of half-waves preserves the individual characteristic properties and features of the signal, and the formal representation in the form of blocks in essence levels out this information. The method is implemented as a set of clustering algorithms for signal elements (half-waves) for both post-processing and real-time. The method is being developed in the direction of analyzing the statistics of half-waves and their characteristics both by signal and by technological intervals. In simple cases of stable measurement conditions, the characteristics of empirical half-wave distributions over clusters of half-waves of the same duration and even just the number of half-waves have an informativeness close to 3. In problem of GTE transmission bearing diagnostics, the measurement conditions are unfavorable (airfield), a stage of clustering of half-waves by similarity is necessary and analysis of the flow of new patterns, or dynamic analysis of cluster sizes, or rank filtering of sequences of half-wave durations. in the latter case, we have a sequence of three-dimensional images tied to the moments of the appearance of half-waves, Δ - is filter aperture, $\mathbf{\Phi}_x$ - maximum filter output, $\mathbf{\Phi}_y$- minimum filter output.

$$\mathbf{x}^* = \left| t0_i \quad \mathbf{\Phi}_x \left(DT_i^{i+\Delta-1} \right) \quad \mathbf{\Phi}_y \left(DT_i^{i+\Delta-1} \right) \right|^*$$

Fig. 2 shows the sequence of images of the signal "_mama_"

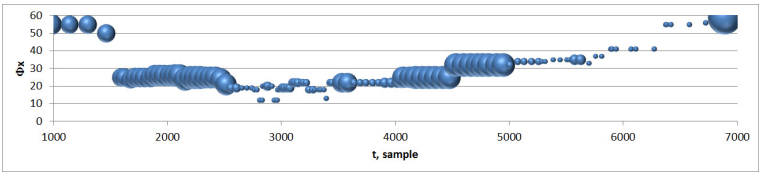


Fig. 2. Half-wave images of the signal ≪_mama_≫. The size of the balls - $\mathbf{\Phi}_y$

Criteria for segmentation and recognition of states of complex sources based on local rank filtering with a sliding aperture of sequences of half-wave durations have been developed, which allow solving segmentation problems, and with a limited alphabet of states, perform recognition of the state in which such half-waves could appear.

Анализ фаз огибающих ЭМГ мышц-антагонистов у пациентов с нейродегенеративными заболеваниями

Сушкова Ольга Сергеевна[1]⋆ o.sushkova@mail.ru
Морозов Алексей Александрович[1] morozov@cplire.ru
Габова Александра Васильевна[2] agabova@yandex.ru
Кершнер Иван Андреевич[1] ivan.kershner@gmail.com
Чигалейчик Лариса Анатольевна[3] chigalei4ick.lar@yandex.ru
Карабанов Алексей Вячеславович[3] doctor.karabanov@mail.ru
[1]Москва, ИРЭ им. В.А. Котельникова РАН
[2]Москва, ИВНД и НФ РАН
[3]Москва, ФГБНУ «Научный центр неврологии»

Одним из перспективных подходов к диагностике таких нейродегенеративных заболеваний, как болезнь Паркинсона (БП) и эссенциальный тремор (ЭТ), является исследование мгновенной фазы мышц-антагонистов. Исследован подход к диагностике БП и ЭТ на основе анализа гистограмм разности мгновенных фаз огибающих электромиограмм (ЭМГ) мышц-антагонистов, а именно, мышц-разгибателей и мышц-сгибателей рук пациентов с БП и ЭТ. Гистограммы разности фаз показывают статистическое распределение разности мгновенной фазы, наблюдаемой на двух исследуемых сигналах. Были исследованы диапазоны частот 1-12 и 20-28 Гц. В данных частотных диапазонах наблюдались статистически значимые отличия (тест Манна-Уитни, альфа-уровень 0.05). Проведено сравнение двух методов вычисления мгновенной фазы. Первый метод основан на вычислении комплексных хребтов вейвлет-спектрограмм и четырёхквадрантного арктангенса. Второй метод основан на использовании полосовой фильтрации, преобразования Гильберта и четырёхквадрантного арктангенса. В качестве метрики для сравнения групп пациентов и контрольной группы испытуемых были исследованы абсолютных значений медиан разности мгновенных фаз огибающих ЭМГ мышц-антагонистов. Обнаруженные статистически значимые отличия абсолютных значений медиан в диапазоне 1-12 Гц соответствуют результатам когерентного анализа, описанным в литературе. Выявленные статистически значимые отличия в диапазоне 20-28 Гц являются новыми результатами. Сравнение методов вычисления мгновенной фазы показало, что комплексные вейвлет-хребты и преобразование Гильберта не имеют явных преимуществ друг перед другом по количеству выявленных статистически значимых отличий, однако время обработки сигналов при использовании преобразования Гильберта меньше (примерно в 10 раза).

Работа выполнена в рамках государственного задания.

[1] *Сушкова О. С., Морозов А. А., Габова А. В., Кершнер И. А., Чигалейчик Л. А., Карабанов А. В.* Анализ фаз огибающих электромиограмм мышц-антагонистов у пациентов с нейродегенеративными заболеваниями // Сборник трудов по материалам ИТНТ-2021, 2021.

[2] *Sushkova O., Morozov A., Kershner I., Petrova N., Gabova A., Chigaleichik L., Karabanov A.* Investigation of distribution laws of the phase difference of the envelopes of electromyograms of antagonist muscles in Parkinson's disease and essential tremor patients // RENSIT, 2020. Vol. 12(3). Pp. 415–428.

[3] *Sushkova O., Morozov A., Gabova A., Karabanov A., Illarioshkin S.* A statistical method for exploratory data analysis based on 2D and 3D area under curve diagrams: Parkinson's disease investigation // Sensors, 2021.

Phase analysis of EMG envelopes of antagonist muscles in patients with neurodegenerative diseases

Sushkova Olga[1]⋆ o.sushkova@mail.ru
Morozov Alexei[1] morozov@cplire.ru
Gabova Alexandra[2] agabova@yandex.ru
Kershner Ivan[1] ivan.kershner@gmail.com
Chigaleichik Larisa[3] chigalei4ick.lar@yandex.ru
Karabanov Alexei[3] doctor.karabanov@mail.ru
[1]Moscow, Kotel'nikov IRE RAS
[2]Moscow, IHNA&NPh RAS
[3]Moscow, FSBI "Research Center of Neurology"

The investigation of the instantaneous phase of antagonist muscles is a promising approach to the diagnosis of neurodegenerative diseases such as Parkinson's disease (PD) and essential tremor (ET). An approach to the diagnosis of PD and ET was investigated based on the analysis of histograms of the difference between instantaneous phases of electromyogram (EMG) envelopes of antagonist muscles, namely, extensor muscles and flexor muscles of the arms of patients with PD and ET. The phase difference histograms indicate the statistical distribution of the instantaneous phase difference observed on the two signals. The frequency ranges from 1 to 12 Hz and from 20 to 28 Hz were investigated. Statistically significant differences were observed in these frequency ranges (the Mann-Whitney test, the alpha level was 0.05). A comparison of two methods for calculating the instantaneous phase was carried out. The first method was based on calculating complex ridges of wavelet spectrograms and four-quadrant arctangent. The second method was based on the use of bandpass filtering, Hilbert transform, and four-quadrant arctangent. The absolute values of the medians of the difference between the instantaneous phases of the EMG envelopes of the antagonist muscles were investigated as a metric for comparing the groups of patients and the control group of subjects. The obtained statistically significant differences in the absolute values of the medians in the 1-12 Hz range correspond to known results of the coherent analysis. The revealed statistically significant differences in the 20-28 Hz range are new results. The comparison of the methods for calculating the instantaneous phase demonstrated that complex wavelet ridges and the Hilbert transform do not have clear advantages over each other in the quantity of revealed statistically significant differences. However, the signal processing time is much less when using the Hilbert transform (by about 10 times).

The work was carried out within the framework of the state task.

[1] *Sushkova O. S., Morozov A. A., Gabova A. V., Kershner I. A., Chigaleichik L. A., Karabanov A. V.* Analiz faz ogibayushchikh elektromiogramm myshts-antagonistov u patsiyentov s neyrodegenerativnymi zabolevaniyami // Sbornik statey ITNT–2021, 2021.

[2] *Sushkova O., Morozov A., Kershner I., Petrova N., Gabova A., Chigaleichik L., Karabanov A.* Investigation of distribution laws of the phase difference of the envelopes of electromyograms of antagonist muscles in Parkinson's disease and essential tremor patients // RENSIT, 2020. Vol. 12(3). Pp. 415–428.

[3] *Sushkova O., Morozov A., Gabova A., Karabanov A., Illarioshkin S.* A statistical method for exploratory data analysis based on 2D and 3D area under curve diagrams: Parkinson's disease investigation // Sensors, 2021.

Топологическая теория анализа хемографов как перспективный подход к имитационному моделированию квантово-механических свойств молекул

Торшин Иван Юрьевич[1]⋆ tiy1357@yandex.ru
Рудаков Константин Владимирович[1] rudakov@ccas.ru
[1]Москва, ФИЦ ИУ РАН

На основе теории топологического анализа данных и теории хемографов разработана проблемно-ориентированная теория для оценочных вычислений квантово-механических свойств молекул по структурной формуле. Показано, что результаты, получаемые, в рамках разработанного формализма, соответствуют решению одноэлектронного уравнения Шредингера на фрагментах молекул с учётом перекрывания фрагментов, аддитивной схеме расчета электронной плотности в теории функционала плотности и учету интегралов перекрывания в теории молекулярных орбиталей (МО). Проведена апробация алгоритмов на выборке из 134000 молекул, для которых значения HOMO (энергия самой высокой занятой МО), LUMO (энергия самой низкой занятой МО), щель HOMO–LUMO, константы вращения, изотропная поляризуемость и других свойств были рассчитаны посредством высокоточной схемы квантовомеханических расчётов B3LYP/6-31G(2df,p). Кросс-валидационное тестирование линейных и нелинейных моделей позволило получить ранговые корреляции между рассчитанными и экспериментальными величинами в диапазоне 0.67-0.85. Скорость вычислений посредством разработанных алгоритмов превышала скорость высокоточных квантово-механических вычислений на 8-9 порядков. Разработанные алгоритмы могут использоваться для крупномасштабных скринингов молекул в рамках решения задач молекулярной фармакологии и материаловедения.

[1] *Torshin I., Rudakov K.* Topological Chemograph Analysis Theory As a Promising Approach to the Simulation Modeling of the Quantum-Mechanical Properties of Molecules. Part I: On the Generationof Feature Descriptions of Molecules // Pattern Recognition and Image Analysis, 2021, Vol. 31(4). Pp. 873–883.

Topological theory of chemograph analysis as a promising approach to simulation modeling of quantum mechanical properties of molecules

Torshin Ivan[1]⋆ tiy1357@yandex.ru
Rudakov Konstantin[1] rudakov@ccas.ru
[1]Moscow, FIC IU RAS

Based on the theory of topological data analysis and the theory of chemographs, a problem-oriented theory has been developed for evaluating the quantum-mechanical properties of molecules by the structural formula. It is shown that the results obtained within the framework of the developed formalism correspond to the solution of the one-electron Schrödinger equation on fragments of molecules taking into account the overlapping of fragments, an additive scheme for calculating the electron density in the theory of the density functional and taking into account the overlap integrals in the theory of molecular orbitals (MO). The algorithms based on the developed formalism were tested on a sample of 134000 molecules, for which the values of HOMO (energy of the highest occupied MO), LUMO (energy of the lowest occupied MO), HOMO – LUMO gap, rotation constants, isotropic polarizability, and other properties were calculated using the high-precision quantum mechanical calculation scheme B3LYP / 6-31G (2df, p). Cross-validation testing of linear and nonlinear models made it possible to obtain rank correlations between the calculated and experimental values in the range of 0,67–0,85. At the same time, the speed of calculations by means of the developed algorithms exceeded the speed of high-precision quantum mechanical calculations by 8-9 orders of magnitude. The developed algorithms can be used for large-scale screening of molecules in the framework of solving problems of molecular pharmacology and materials science.

[1] *Torshin I., Rudakov K.* Topological Chemograph Analysis Theory As a Promising Approach to the Simulation Modeling of the Quantum-Mechanical Properties of Molecules. Part I: On the Generationof Feature Descriptions of Molecules // Pattern Recognition and Image Analysis, 2021, Vol. 31(4). Pp. 873–883.

Распознавание опасных аритмий для выявления рисков осложнений сердечно-сосудистых заболеваний

Манило Людмила Алексеевна* lmanilo@yandex.ru
Немирко Анатолий Павлович apn-bs@yandex.ru
Алексеев Борис Эдуардович boris123z@yandex.ru

Санкт-Петербург, Санкт-Петербургский государственный электротехнический университет "ЛЭТИ"

Одним из факторов риска развития сердечно-сосудистых заболеваний является наличие у пациента сердечных аритмий. Своевременное распознавание этих нарушений в ходе мониторного контроля ЭКГ позволяет врачу оценить величину риска и предупредить возможное появление тяжелых последствий. Особенно важным становиться обнаружение опасных форм аритмий, требующих принятия срочных реанимационных мер, а также распознавание их предвестников. К опасным аритмиям относят фибрилляцию желудочков (ФЖ) сердца, а к её предвестникам желудочковую тахикардию (ЖТ), имеющую несколько форм проявления. Наибольшую угрозу возникновения ФЖ представляет пируэтная форма ЖТ.

Цель данного исследования заключена в оценке возможности эффективного распознавания жизнеугрожающих аритмий по коротким записям ЭКГ сигнала. В частности, рассматриваются алгоритмы классификации опасных аритмий по 2 с фрагментам ЭКГ с использованием спектрального описания сигнала. Выбор такого подхода диктуется требованием мгновенного обнаружения нарушения в момент его появления, а также выраженными отличиями частотных свойств ЭКГ для разных классов аритмий.

В работе исследуются алгоритмы классификации, реализованные на базе 4-х известных методов распознавания, а также с использованием нейронных сетей [1]. Они включают взвешенный метод k ближайших соседей (kNN), метод ближайшей выпуклой оболочки [2], линейный дискриминантный анализ, метод опорных векторов (SVM) с несколькими вариантами ядер: линейным, квадратичным, кубическим и с ядром, основанным на функции Гаусса.

В качестве первичного источника ЭКГ записей выбрана база PhysioNet «The MIT-BIH Malignant Ventricular Ectopy Database» (MVED), которая содержит достаточно полный набор необходимых для исследования нарушений сердечного ритма. На её основе создана собственная база коротких фрагментов ЭКГ, в которой выделено 6 классов аритмий [3]. Группировка аритмий по классам и ранжирование классов проведены с учётом степени опасности нарушений для жизни пациента и риска возможных осложнений. В результате анализа данных MVED было вручную вырезано 1000 фрагментов 2с ЭКГ и путём тщательного отбора сформированы выборки по 90 объектов в каждом классе. В качестве признаков выбраны сглаженные отсчёты спектральной плотности мощности в

области частот 0 – 15 Гц. Они получены с использованием алгоритма быстрого преобразования Фурье и применением периодограммной оценки Даньелла.

Учитывая особенности группировки этих классов в признаковом пространстве, задача распознавания решалась в два этапа: на первом этапе распознавались группы нарушений (опасные/неопасные), а на втором этапе классифицировался вид опасного нарушения. К опасным нарушениям отнесены: ФЖ, ЖТ и пируэтная форма ЖТ. Альтернативную группу составили различные формы неопасных желудочковых и наджелудочковых нарушений, а также синусовый ритм. Для оценки качества работы алгоритмов использовались показатели: чувствительность, специфичность и общая точность.

Сравнение результатов распознавания двух групп аритмий показало, что наибольшую точность (94,8%) обеспечивает алгоритм на основе кубического SVM. Установлено также, что основные ошибки связаны с резкой нестационарностью ЭКГ при фибрилляции желудочков, особенно в момент возникновения пируэтной формы ЖТ. Увеличение верхней границы частотного диапазона с 15 Гц до 50 Гц не дало существенного выигрыша в точности классификации. На втором этапе классификации, т.е. при распознавании ФЖ на фоне ЖТ, наилучшие результаты показал алгоритм kNN (чувствительность – 82,2%, точность – 77,2%). Как показали эксперименты, повышение чувствительности алгоритма можно достичь за счёт незначительного снижения специфичности. Такую коррекцию решающей функции целесообразно применить ввиду особой важности задачи надежного обнаружения ФЖ.

С целью сравнения качества работы классических и нейросетевых методов проведено обучение нейронных сетей двух типов (полносвязная нейронная сеть и рекуррентная нейронная сеть). Установлено, что нейронные сети не смогли повысить качество классификации наилучшего алгоритма на основе кубического SVM. Близкий по точности результат показал алгоритм на основе полносвязной нейронной сети (93,8%), а для рекуррентной нейронной сети точность оказалась ниже (85,4%).

Таким образом, задача обнаружения опасных для жизни нарушений ритма может быть эффективна решена. В то же время дифференциация вида опасной аритмии вызывает определенные трудности, связанные в основном с большим разнообразием форм сигнала и резкой его изменчивостью даже в пределах короткого 2 с фрагмента. Это требует проведения дальнейших исследований, поскольку задача оценки риска возможных осложнений по результатам анализа аритмий остаётся актуальной.

Работа поддержана грантами РФФИ No. 19-07-00475 и No. 19-29-01009.

[1] *Nemirko A., Manilo L., Alekseev B., Sokolova A., Yuldashev Z.* The Comparison of Algorithms for Life-threatening Cardiac Arrhythmias Recognition // SCITEPRESS - Science and Technology Publications, 2021. Vol. 1. Pp. 402–407.

[2] *Manilo L., Nemirko A., Evdakova E., Tatarinova A.* ECG Database for Evaluating the Efficiency of Recognizing Dangerous Arrhythmias // IEEE CSGB, 2021. Pp. 120–125.

[3] *Manilo L., Nemirko A., Evdakova E.* Recognition of Dangerous Rhythm Disturbances from Short ECG Fragments // IEEE USBEREIT, 2021. Pp. 41–44.

Recognition of dangerous arrhythmias to identify the risks of complications of cardiovascular diseases

Manilo Liudmila* lmanilo@yandex.ru
Nemirko Anatoly apn-bs@yandex.ru
Alekseev Boris boris123z@yandex.ru
Saint Petersburg, Saint Petersburg Electrotechnical University "LETI"

One of the risk factors for the development of cardiovascular diseases is the presence of cardiac arrhythmias in a patient Timely recognition of these disorders during ECG monitoring allows the doctor to assess the magnitude of the risk and prevent the possible occurrence of severe consequences. It becomes especially important to detect dangerous forms of arrhythmias requiring urgent resuscitation measures, as well as to recognize their precursors. Dangerous arrhythmias include ventricular fibrillation (VF) of the heart, and its precursor is ventricular tachycardia (VT), which has several forms of manifestation. The pirouette form of VT poses the greatest threat to the onset of VF.

The purpose of this study is to evaluate the possibility of effective recognition of life-threatening arrhythmias from short recordings of the ECG signal. In particular, algorithms for classifying dangerous arrhythmias by 2 s ECG fragments using spectral description of the signal are considered. The choice of such an approach is dictated by the requirement for instant detection of the violation at the moment of its occurrence, as well as the pronounced differences in the frequency properties of the ECG for different classes of arrhythmias.

The paper investigates classification algorithms implemented on the basis of 4 well-known recognition methods, as well as using neural networks [1]. These include the weighted k nearest neighbors (kNN) method, the nearest convex hull method [2], linear discriminant analysis, support vector machine (SVM) with multiple kernels: linear, quadratic, cubic, and Gaussian-based kernel.

The PhysioNet database "The MIT-BIH Malignant Ventricular Ectopy Database" (MVED) was selected as the primary source of ECG records, which contains a fairly complete set of necessary for the study of cardiac arrhythmias. Based on it, a proprietary database of short ECG fragments has been created, in which 6 classes of arrhythmias have been identified [3]. The grouping of arrhythmias by classes and the ranking of classes were carried out taking into account the degree of danger of violations for the patient's life and the risk of possible complications. As a result of the analysis of the MVED data, 1000 2s ECG fragments were manually cut out and samples of 90 objects in each class were formed by careful selection. Smoothed samples of the spectral power density in the frequency range $0 - 15$ Hz were selected as features. They are obtained using the Fast Fourier Transform algorithm and Daniell's periodogram estimate.

Taking into account the peculiarities of grouping these classes in the feature space, the recognition problem was solved in two stages: at the first stage, groups of

violations (dangerous/non-dangerous) were recognized, and at the second stage, the type of dangerous violation was classified. Dangerous disorders include: VF, VT and the pirouette form of VT. The alternative group consisted of various forms of non-dangerous ventricular and supraventricular disorders, as well as sinus rhythm. To assess the quality of the algorithms, the following indicators were used: sensitivity, specificity and overall accuracy.

Comparison of the recognition results of two groups of arrhythmias showed that the highest accuracy (94.8%) is provided by an algorithm based on cubic SVM. It was also found that the main errors are associated with a sharp non-stationarity of the ECG during ventricular fibrillation, especially at the time of the occurrence of the pirouette form of VT. Increasing the upper limit of the frequency range from 15 Hz to 50 Hz did not give a significant gain in classification accuracy. At the second stage of classification, i.e., when recognizing VF against the background of VT, the kNN algorithm showed the best results (sensitivity - 82.2%, accuracy - 77.2%). Experiments have shown that an increase in the sensitivity of the algorithm can be achieved at the expense of a slight decrease in specificity. It is advisable to apply such a correction of the decisive function due to the special importance of the task of reliable detection of VF.

In order to compare the quality of classical and neural network methods, two types of neural networks (fully connected neural network and recurrent neural network) were trained. It was found that neural networks could not improve the classification quality of the best algorithm based on cubic SVM. An algorithm based on a fully connected neural network showed a close result in accuracy (93.8%), and for a recurrent neural network the accuracy was lower (85.4%).

Thus, the problem of detecting life-threatening rhythm disturbances can be effectively solved. At the same time, the differentiation of the type of dangerous arrhythmia causes certain difficulties, mainly due to a large variety of signal shapes and its sharp variability even within a short 2s fragment. This requires further research, since the task of assessing the risk of possible complications based on the results of the analysis of arrhythmias remains relevant.

This research is funded by RFBR, grant No 19-07-00475 and No 19-29-01009.

[1] *Nemirko A., Manilo L., Alekseev B., Sokolova A., Yuldashev Z.* The Comparison of Algorithms for Life-threatening Cardiac Arrhythmias Recognition // SCITEPRESS - Science and Technology Publications, 2021. Vol. 1. Pp. 402–407.

[2] *Manilo L., Nemirko A., Evdakova E., Tatarinova A.* ECG Database for Evaluating the Efficiency of Recognizing Dangerous Arrhythmias // IEEE CSGB, 2021. Pp. 120–125.

[3] *Manilo L., Nemirko A., Evdakova E.* Recognition of Dangerous Rhythm Disturbances from Short ECG Fragments // IEEE USBEREIT, 2021. Pp. 41–44.

Применение хребтов вейвлет спектров в выделении диагностических признаков ЭЭГ: Длительный мониторинг эпилепсии

Обухов Юрий Владимирович[1] yuvobukhov@mail.ru
Кершнер Иван Андреевич[1]⋆ ivan_kershner@mail.ru
Синкин Михаил Владимирович[2,3] mvsinkin@gmail.com

[1]Москва, Институт радиотехники и электроники имени В.А. Котельникова РАН
[2]Москва, Научно-исследовательский институт скорой помощи имени Н.В. Склифосовского
[3]Москва, Московский государственный медико-стоматологический университет имени А. И. Евдокимова

Межканальная синхронизация ЭЭГ, а также ее нарушение - важный диагностический признак ряда заболеваний. В частности, при эпилептическом приступе синхронизация может наблюдаться как на одной паре ЭЭГ каналов, так и на нескольких парах при генерализованном приступе. В рамках подхода к анализу межканальной синхронизации ЭЭГ по хребтам вейвлет-спектров решена задача сегментации данных длительного мониторинга ЭЭГ с поверхности головы, содержащего различные артефакты и фрагменты, характерные для эпилептических приступов, с целью сокращения общей длительности анализируемых врачом фрагментов. Исходная ЭЭГ была сегментирована на фрагменты, в которых было синхронизировано не менее двух каналов, а методом адаптивного порога - на фрагменты с высоким значением спектральной плотности мощности. Обнаружены временные интервалы, соответствующие пересечению синхронизированных по времени фрагментов с фрагментами с высокими значениями спектральной плотности мощности. В результате общая длительность обработки фрагментов для анализа врачом сократилась более чем в 60 раз.

[1] *Obukhov Yu. et al.* Wavelet Ridges in EEG Diagnostic Features Extraction: Epilepsy Long-Time Monitoring and Rehabilitation after Traumatic Brain Injury // Sensors, 2021. Vol. 21(18). Pp. 59–89.

Wavelet Ridges in EEG Diagnostic Features Extraction: Epilepsy Long-Time Monitoring

Obukhov Yury[1] yuvobukhov@mail.ru
Kershner Ivan[1]⋆ ivan_kershner@mail.ru
Sinkin Mikhail[2,3] mvsinkin@gmail.com

[1]Moscow,Kotelnikov Institute of Radio Engineering and Electronics of RAS
[2]Moscow, N.V. Sklifosovsky Research Institute for Emergency Medicine of Moscow Healthcare Department
[3]Moscow, A.I. Yevdokimov Moscow State University of Medicine and Dentistry

Interchannel EEG synchronization, as well as its violation, is an important diagnostic sign of a number of diseases. In particular, during an epileptic seizure, such synchronization occurs starting from some pairs of channels up to many pairs in a generalized seizure. Within the framework of the approach to the analysis of interchannel EEG synchronization, the problem of long-term EEG segmentation has been solved. The initial data of long-term monitoring of the scalp EEG containing various artefacts were reduced by analyzing the wavelet spectrum ridges to several fragments. Overlapping in time synchronized fragments with fragments of high spectral power density was determined. As a result, the total duration of the fragments for analysis by the doctor was reduced by more than 60 times.

[1] *Obukhov Yu. et al.* Wavelet Ridges in EEG Diagnostic Features Extraction: Epilepsy Long-Time Monitoring and Rehabilitation after Traumatic Brain Injury // Sensors, 2021. Vol. 21(18). Pp. 59–89.

Перспективы использования обобщённого спектрального анализа в решении задач распознавания паттернов вариабельности ритма сердца

Махортых Сергей Александрович[1]⋆ makh@impb.ru
Москаленко Андрей Витальевич[1] physiome@mail-avm.ru
[1]Пущино, ИМПБ РАН — филиал ИПМ им. М. В. Келдыша РАН

Автоматизация массового доврачебного обследования населения с целью выявления скрытых нарушений здоровья остаётся актуальной задачей современности. Необходимость такой автоматизации обусловлена принципиальной невозможностью обеспечения такой численности врачей-специалистов, какая требуется для оказания врачебных услуг каждому в них нуждающемуся гражданину РФ. Ранее было показано в многочисленных исследованиях, что наиболее удобным и дешевым неинвазивным методом выявления скрытых нарушений здоровья следует признавать анализ вариабельности сердечного ритма, ВСР, основанный на исследовании ритмограмм (хронокардиограмм). Анализ ВСР в сочетании с функциональными пробами применяется для проведения дифференциальной диагностики и с целью решения задач медицинского прогнозирования; с 1960-х годов этот метод прочно зарекомендовал себя в космической медицине. Около четверти века назад специалистами по анализа ритма сердца и обработке биосигналов были сформулированы два варианта рекомендаций по использованию методов анализа ВСР, международные и российские. Следует отметить, что благодаря работам Р. М. Баевского в российских рекомендациях содержатся также и методы распознавания образов. А в последние годы многие исследователи пришли к выводам о необходимости поиска и распознавания паттернов хронокардиограм, характерных для тех или иных психофизиологических состояний человека.

Вместе с тем, стандарты методов анализа ВСР отсутствуют до настоящего времени, хотя необходимость их разработки и принятия была отмечена в обоих текстах рекомендаций. Как указано во многих работах, нерешёнными остаются также и проблемы спектрального анализа ВСР, исследования нестационарных (переходных) физиологических процессов, использования методов нелинейной динамики и распознавания образов.

Для успешного решения этих проблем, нам представляется полезным применение обобщённого спектрально-аналитического метода (ОСАМ) в сочетании с классическими методами распознавания образов и современными технологиями искусственного интеллекта. Обоснованность такого подхода нами была прежде показана в ряде теоретических работ [1]. Отметим также, что ОСАМ был успешно использован для решения многих медико-биологических задач. Новый подход к анализу ВСР основывается на эквивалентной модели основного ритма сердца, для идентификации параметров которой и будет использован ОСАМ.

[1] *Москаленко А. В.* Базовые механизмы аритмий сердца // М.: ИД Медпрактика-М, 2021.

Perspectives for using of generalized spectral analysis in solving problems of recognizing patterns of heart rate variability

Makhortykh Sergey[1]⋆ makh@impb.ru
Moskalenko Andrey[1] physiome@mail-avm.ru
[1]Pushchino, IMPB RAS — Branch of KIAM RAS

Automation of mass pre-medical inspection of the population in order to identify hidden health disorders remains an actual task of our time. The necessity for such automation is conditioned by the fundamental impossibility of providing such a number of medical specialists, which is required for providing adequate medical services to every citizen of the Russian Federation, who is in need of them. Earlier it was shown in numerous studies that the analysis of heart rate variability (HRV), which based on the study of rhythmograms (chronocardiograms), should be recognized as the most convenient cheap non-invasive method for detecting latent health disorders. Analysis of HRV in combination with functional tests is used for differential diagnostics and for solving problems of medical prediction; since the 1960s, this method has firmly established itself in space medicine. About for a quarter of a century ago, specialists in the analysis of heart rhythm and processing of biosignals have formulated and proposed two versions of recommendations for the use of methods of analysis of HRV, international and Russian. It should be noted that thanks to the works of R. M. Baevsky, Russian recommendations also contain methods for pattern recognition. In recent years, researchers have come to the conclusion about necessity of searching and recognizing patterns of chronocardiograms, that are characteristic for certain psychophysiological states of a person.

At the same time, the standards for methods of HRV analysis, the need for the development and acceptance of which was indicated in both texts of the recommendations, are still absent. As indicated in a number of publications, some difficulties of spectral analysis of HRV, non-stationary (transient) physiological processes studies, and the use of methods of nonlinear dynamics and pattern recognition also remain unsolved.

It is supposed that the generalized spectral-analytical method (GSAM) can be useful to solve these problems successfully, when it will be combined with classical methods of pattern recognition as well as with modern technologies of artificial intelligence. The validity of such an approach was previously shown by us in a number of theoretical works [1]. We also note that GSAM has already been successfully applied for solving a number of biomedical problems. The new approach to the analysis of HRV is supposed to be based on an equivalent model of the basic rhythm of the heart, for the identification of the parameters of which the GSAM will be used.

[1] *Moskalenko A. V.* Basic mechanisms of cardiac arrhythmias // M.: Medpraktika-M, 2021.

Использование методов кластерного анализа в исследовании эпидемических процессов COVID-19 в странах мира.

Сенько Олег Валентинович[1,2]⋆ senkoov@mail.ru
Кузнецова Анна Викторовна[3,2] azforus@yandex.ru
Добролюбова Ольга Анатольевна[2,4] dbrl.olga@gmail.com
Воронин Евгений Михайлович[2] emvoronin@yandex.ru
Акимкин Василий Геннадьевич[2] crie@pcr.ru
Плоскирева Антонина Александровна[2] antoninna@mail.ru

[1]Москва, ФИЦ "Информатика и управление"РАН
[2]Москва, ФБУН ЦНИИ Эпидемиологии Роспотребнадзора
[3]Москва, Институт биохимической физики им. Н.М.Эмануэля РАН
[4]Москва, МГУ имени М.В.Ломоносова

Целью исследования является изучения различных факторов на динамику заболеваемости COVID-19 в различных странах мира. В число таких факторов входят демографические, социально-экономические и климатические факторы. Наиболее полно информация об эпидпроцессе COVID-19 содержится в кривых динамики заболеваемости, описывающих суточный прирост числа новых случаев заболевания. Исследования проводились по набору кривых динамики заболеваемости COVID-19 для 106 стран, описывающих развитие эпидпроцесса за период с 29.01.2020 по 05.08.2021. Использовался способ исследования, основанный на выделении нескольких групп сходных по своей форме эпидемиологических кривых (эпидкривых) с использованием методов кластерного анализа. Далее соответствующие группы стран сравнивались между собой по набору факторов перечисленных типов с выявлением статистически значимых различий.

Использовался метод иерархической агломеративной кластеризации. В качестве меры сходства двух кривых использовался коэффициент корреляции Пирсона между суточными приростами числа новых случаев заболевания. Для более точного учёта сходства/различий между двумя эпидкривыми коэффициент корреляции ρ вычислялся при различных величинах лага, то есть сдвига между датами наблюдений. Мерой сходства $\varrho(C_i, C_j)$ между двумя эпидкривыми C_i и C_j считалась максимальная величина коэффициента корреляции на множестве значений лага из интервала от 0 до 19 суток. В качестве меры сходства двух групп эпидкривых $G^{'}$ и $G^{''}$ использовалось среднее значение меры сходства между эпидкривыми из разных групп: $P(G^{'}, G^{''}) = \frac{1}{m^{'}m^{''}} \sum_{i=1}^{m^{'}} \sum_{j=1}^{m^{''}} \varrho(C_i, C_j)$. Процесс слияния кластеров прекращался, если мера сходства P между любыми двумя кластерами в текущей кластеризацией не окажется ниже 0.5. В результате использования указанного процесса было получено 4 кластера, каждый из которых включал не менее 10 стран: кластер I - 11 географически удалённых стран, включая США, Великобританию, РФ, Мексику, ЮАР и др.; кластер II - 39 преимущественно европейских стран; кластер III - 17 стран преимуществен-

но Азии и Северной Африки; кластер IV - 13 стран, включая 5 стран Южной Америки, Индии и ряда азиатских стран.

Важнейшей составляющей исследований кластерной структуры является её верификация. В задачах кластеризации эпидкривых и других временных рядов убедительным свидетельством объективности кластеризации является сохранение выявленной кластерной структуры на временных интервалах, находящихся за пределами интервала, по которому она была получена. Было проведено исследование кластеров, полученных на контрольном временном интервале от 05.08.2021 до 01.11.2021. С этой целью проводился визуальный анализ эпидкривых содержащих суточные приросты числа новых случаев заболевания для всех стран, вошедших в каждый отдельный кластер. Данный анализ подтвердил существенное различие форм эпидкривых на контрольном интервале. Производилось также сравнение средней величины упомянутой меры $\varrho(C_i, C_j)$ по всем парам эпидкривых из объединённой группы. Средняя мера сходства по объединённой группе составила 0.248, в то время как внутрикластерные меры сходства составляют 0.295, 0.345, 0.533, 0.458 для кластеров I, II, III, IV соответственно. Таким образом, вариабельность кривых внутри кластеров на контрольном временном интервале заметно ниже вариабельности эпидкривых в объединённой группе.

Для оценки статистической значимости различий между двумя кластерами G_i и G_j использовался перестановочный тест. Значение критерия $F = \frac{P(G_i,G_i)+P(G_j,G_j)-P(G_i,G_j)}{P(G_i,G_j)}$ для двух найденных кластеров сравнивается со значениями критерия F для пар случайных групп G_i^r и G_j^r , имеющих тот же самый размер, что и исходные группы. Группы G_i^r и G_j^r генерируются из исходных кластеров G_i и G_j с помощью случайных перестановок номеров кривых в группе G_i, G_j . Перестановочный тест позволил выявить значимые на уровне $p<0.01$ различия между кластером II и кластерами I, III, IV.

Оценка различий между кластерами производилась по 116 факторам. Попарное сравнение кластеров проводилось с использованием нескольких методов машинного обучения, а также метода оптимальных достоверных разбиений (ОДР). При этом ОДР использовался для изучения различий как по отдельным факторам, так и по парным сочетаниям факторов. В ходе исследования была выявлена статистически значимая связь полученной кластеризации с такими социально-экономическими и демографическими факторами как коэффициент Джини, ВВП на душу населения, величина экспорта и импорта товаров и услуг, ожидаемая продолжительностью жизни, коэффициент рождаемости, коэффициент детской смертности. Использование методов машинного обучения подтверждает наличие существенных различий между кластерами по анализируемым факторам: величина ROC AUC изменялась от 0.648 при попытке взаимного распознавания кластеров I и III до 0.807 при попытке взаимного распознавания кластеров II и IV.

Таким образом полученные результаты свидетельствует о связи эпидпроцесса COVID-19 с социально-экономическими и демографическими показателями.

[1] *Kuznetsova A., Kostomarova I., Senko O.* Modification of the method of optimal valid partitioning for comparison of patterns related to the occurrence of ischemic stroke in two groups of patients // Pattern recognition and image analysis (advances in mathematical theory and applications), 2014. Vol. 24(1). Pp. 5–25.

[2] *Брико Н., Онищенко Г., Покровскийо В.* Руководство по эпидемиологии инфекционных болезней // Москва, издательство «МИА», 2019. Т. 1, С. 72—75.

Clustering methods for COVID-19 epidemic analytics in countries across the world

Senko Oleg[1,2]⋆ senkoov@mail.ru
Kuznetsova Anna[3,2] azforus@yandex.ru
Dobrolyubova Olga[2,4] dbrl.olga@gmail.com
Voronin Evgeniy[2] emvoronin@yandex.ru
Akimkin Vasiliy[2] crie@pcr.ru
Ploskireva Antonina[2] antoninna@mail.ru

[1]Moscow, Federal Research Center Computer Science and Control RAS
[2]Moscow, Central Research Institute of Epidemiology (Rospotrebnadzor)
[3]Moscow, Institute of Biochemical Physics, Russian Academy of Sciences
[4]Moscow, Lomonosov Moscow State University

The study aims to study various factors on the dynamics of the incidence of COVID-19 in different countries of the world. These factors include demographic, socio-economic and climatic factors. The complete information about the epidemiological process of COVID-19 is contained in the incidence dynamics curves describing the daily increase in the number of new cases of the disease. This research considers a COVID-19 epidemiological dataset for 106 periods from 01/29/2020 to 08/05/2021. A study of epidemic curves groups (epi curves) is selecting several groups by cluster analysis methods. Further, we compared the corresponding groups of countries according to a set of factors of the listed types to identify statistically significant differences.

We used the method of hierarchical agglomerative clustering. As a measure of the similarity of the two curves, Pearson's correlation coefficient between daily increases in the number of new cases of diseases. The correlation coefficient ρ was calculated for different lag values - time lag between the observation dates - to account for the similarities/differences between the two epi curves more accurately. A measure of the similarity $\varrho(C_i, C_j)$ between two epi curves C_i and C_j calculated maximum correlation coefficient value on the set of lag values from the interval from 0 to 19 days. We used the average value of the measure between epi curves $G^{'}$ and $G^{''}$ from different groups to measure the similarity between two groups of epi curves: $P(G^{'}, G^{''}) = \frac{1}{m^{'} m^{''}} \sum_{i=1}^{m^{'}} \sum_{j=1}^{m^{''}} \varrho(C_i, C_j)$. We terminated the process of merging clusters if the similarity measure P between any two groups in the current clustering is not lower than 0.5. As a result, 4 clusters were obtained, each of which included at least 10 countries: cluster I - 11 geographically distant countries, including the USA, Britain, Russia, Mexico, South Africa, etc.; Cluster II - 39 predominantly European countries; Cluster III - 17 countries, mainly in Asia and North Africa; cluster IV - 13 countries, including 5 countries in South America, India and some Asian countries.

The essential component of cluster structure research is its verification. The task of clustering epi curves and other time intervals is an obvious indicator of clustering objectivity - preserving a new cluster structure on time intervals outside

the interval. Clusters validation was doing a control time interval from 08/05/2021 to 11/01/2021. For this purpose, a visual analysis of the daily number of disease new cases was carried out for all countries in every cluster. This analysis showed a significant difference in the shape of the epi curves in the control interval. We also made a comparison of the average measure $\varrho(C_i, C_j)$ for all pairs of epi curves from the combined group . The combined group included all countries from 4 clusters. The average similarity score for the combined group was 0.248, while the within-cluster similarity measures were 0.295, 0.345, 0.533, 0.458 for clusters I, II, III, IV, respectively.

The within-clusters variability of the curves in the control time interval is noticeably lower than that of the epi curves in the combined group. A permutation test was used to assess the statistical significance of the differences between the two clusters- G_i and G_j. The test value $F = \frac{P(G_i,G_i)+P(G_j,G_j)-P(G_i,G_j)}{P(G_i,G_j)}$ for the two found clusters is compared with test value F of random groups pairs G_i^r and G_j^r of the same size as the original groups. We generated G_i^r and G_j^r groups from the initial clusters G_i and G_j by random permutations of the curve index in the group G_i, G_j .The permutation test revealed significance at the $\rho < 0.01$ level between cluster II and clusters I, III, IV.

Also, we assessed the differences between the clusters for 116 factors. Pairwise comparison of clusters was carried out using several machine learning methods and the method of optimal valid partitioning (OVP). In this case, we used the OVP to study differences between individual factors and factor paired combinations. The study revealed a statistically significant relationship between the obtained clustering by such socio-economic and demographic factors as the Gini coefficient, GDP per capita, the exports value and imports of goods and services, life expectancy, fertility rate, infant mortality rate. The use of machine learning methods confirms the presence of significant differences between the clusters in terms of the analyzed factors: the ROC AUC value varied from 0.648 when attempting to recognize clusters I and III to acknowledge each other to 0.807 when trying to identify clusters II and IV.

Consequently, there is evidence of a link between the COVID-19 epidemic process in various world countries with socio-economic and demographic indicators.[1]

[1] *Kuznetsova A., Kostomarova I., Senko O.* Modification of the method of optimal valid partitioning for comparison of patterns related to the occurrence of ischemic stroke in two groups of patients // Pattern recognition and image analysis (advances in mathematical theory and applications), 2014. Vol. 24(1). Pp. 5–25.

[2] *Briko N., Onischenko G., Pokrovsky V.* Guideline for epidemiology of infectious diseases // Moscow, publishing house ¡¡MIA¿¿, 2019. Vol. 1, P. 72–75.

Оценка восстановления межканальных фазовых связей электроэнцефалограмм при когнитивных тестах у пациентов с черепно-мозговой травмой средней тяжести до и после реабилитации

Обухов Юрий Владимирович[1] yuvobukhov@mail.ru
Толмачева Рената Алексеевна[1]⋆ tolmatcheva@ya.ru
Жаворонкова Людмила Алексеевна[2] lzhavoronkova@hotmail.com
[1]Москва, ИРЭ им. В.А. Котельникова РАН
[2]Москва, ИВНД и НФ РАН

Одним из методов оценки нарушения связей между различными областями мозга пациентов в результате черепно-мозговой травмы является определение межканальной связанности электроэнцефалограмм. После черепно-мозговой травмы происходит разрушение межнейронных связей, что приводит к нарушению межканальной синхронизации при выполнении моторных или когнитивных тестов. В рамках подхода к анализу межканальной синхронизации ЭЭГ с использованием хребтов вейвлет-спектров была решена задача оценки восстановления после реабилитации когнитивных функций у пациентов с черепно-мозговой травмой средней степени тяжести. Межканальная фазовая связанность сигналов электроэнцефалограмм определялась в ходе когнитивного теста до и после реабилитации. Использовались счетно-логический и пространственно-образный когнитивные тесты. Положительная динамика реабилитации определялась при инициализации межполушарных связей и связей в лобной коре головного мозга.

[1] *Obukhov Yu. et al.* Wavelet Ridges in EEG Diagnostic Features Extraction: Epilepsy Long-Time Monitoring and Rehabilitation after Traumatic Brain Injury // Sensors, 2021. Vol. 21(18). Pp. 59-89.

[2] *Толмачева Р. А. Обухов Ю. В. Жаворонкова Л. А.* Оценка восстановления межканальных фазовых связей электроэнцефалограмм при когнитивных тестах у пациентов с черепно-мозговой травмой до и после реабилитации //Радиотехника и электроника, 2021. Т. 66(10). С. 1004-1010.

Estimation of recovery of interchannel phase connections of electroencephalograms during cognitive tests in patients with moderate traumatic brain injury before and after rehabilitation

Obukhov Yury[1] yuvobukhov@mail.ru
Tolmacheva Renata[1]⋆ tolmatcheva@ya.ru
Zhavoronkova Ludmila[2] lzhavoronkova@hotmail.com
[1]Moscow, Kotel'nikov IRE RAS
[2]Moscow, IHNA&NPh RAS

One of the methods for estimation the disruption of connections between different areas of the brain of patients as a result of traumatic brain injury is to determine the interchannel EEG connectivity. After traumatic brain injury, the destruction of interneuronal connections occurs, which leads to a violation of interchannel synchronization when performing motor or cognitive tests. Within the framework of a approach to the analysis of interchannel EEG synchronization using the ridges of wavelet spectra the task of assessment of recovery after rehabilitation of cognitive functions in patients with moderate traumatic brain injury was solved. The interchannel phase connectivity EEG channels was determined during the cognitive test before and after rehabilitation. Calculation-logical and spatial-pattern cognitive tests were used. The positive dynamics of rehabilitation was determined during the initialization of interhemispheric connections and connections in the frontal cortex of the brain.

[1] *Obukhov Yu. et al.* Wavelet Ridges in EEG Diagnostic Features Extraction: Epilepsy Long-Time Monitoring and Rehabilitation after Traumatic Brain Injury // Sensors, 2021. Vol. 21(18). Pp. 59-89.

[2] *Tolmacheva R. A., Obukhov Yu. V., Zhavoronkova L. A.* Otsenka vosstanovleniya mezhkanal'nykh fazovykh svyazey elektroentsefalogramm pri kognitivnykh testakh u patsiyentov s cherepno-mozgovoy travmoy do i posle reabilitatsii // Radiotekhnika i elektronika, 2021. Vol. 66(10). Pp. 1004-1010.

Оценка корреляций между компартментами мозга при синдроме 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

[1]Пущино, ИМПБ РАН – филиал ИПМ им. М.В. Келдыша РАН

Предложен новый метод для изучения корреляции между частями головного мозга человека по данным магнитной энцефалографии. Временные ряды для корреляционного анализа генерируются при помощи метода виртуальных электродов. На первом этапе многоканальные магнитоэнцефалограммы субъектов с подтвержденным диагнозом дефицита внимания и гиперактивности преобразуются в функциональные томограммы – пространственное распределение источников магнитного поля на дискретной сетке. Это достигается решением обратной задачи для всех элементарных осцилляций, выделяемых с помощью преобразования Фурье. Каждой частоте соответствует узел трехмерной сетки, в котором располагается токовый диполь, осциллирующий с данной частотой. Виртуальный электрод представляет собой область пространства, активность которой предполагается изучать. Временная зависимость этой активности получается суммированием спектральной мощности всех источников, попадающих в область виртуального электрода.

В работе анализировались две магнитных энцефалограммы (МЭГ) и две магнитно-резонансных томограммы (МРТ) из базы OMEGA. Для анализа корреляций были выбраны следующие полосы частот, или ритмы: тета — 4-8 Гц, альфа — 8-13 Гц, бета — 13-35 Гц, гамма — 35-50 Гц. Магнитная энцефалограмма в полосе частот дельта ритма обусловлена работой сосудов, в основном расположенных вне мозга. Поэтому дельта ритм не рассматривается при анализе корреляций.

Были рассчитаны корреляции между следующими компартментами мозга: лобная доля (левая и правая), височная доля (левая и правая), теменная доля (левая и правая), затылочная доля (левая и правая), диэнцефалон, мозжечок.

Для характеристики функциональной связанности между областями мозга i и j был использован квадрат коэффициента корреляции Пирсона. Сначала вычисляются временные ряды $P_i(t)$ и $P_j(t)$ на дискретной временной сетке с частотой регистрации 2400 Гц, где t изменяется от 0 до 300 секунд (продолжительность эксперимента). Ряды $P_i(t)$ и $P_j(t)$ разбиваются на эпохи длительностью 1 секунда, на каждой эпохе номер l вычисляется квадрат коэффициента корреляции Пирсона по формуле:

$$\left(r_{ij}^{l}\right)^2 = \frac{cov^2(x,y)}{s_x^2 s_y^2},$$

где выборки $x = (P_i(t_1^l), ..., P_i(t_m^l)$ и $y = (P_j(t_1^l), ..., P_j(t_m^l)$,$s_x^2$ и s_y^2 - выборочные дисперсии, l – номер эпохи, m – размер выборки. Затем проводится усреднение по эпохам:

$$c_{ij} = \frac{1}{300} \sum_{i=1}^{300} \left(r_{ij}^l\right)^2$$

Коэффициент c_{ij} отображается на квадратных матрицах, характеризующих корреляции между областями i и j в различных частотных диапазонах. На рисунках показаны матрицы корреляции для одного из субъектов в частотных диапозонах $\theta, \alpha, \beta, \gamma$.

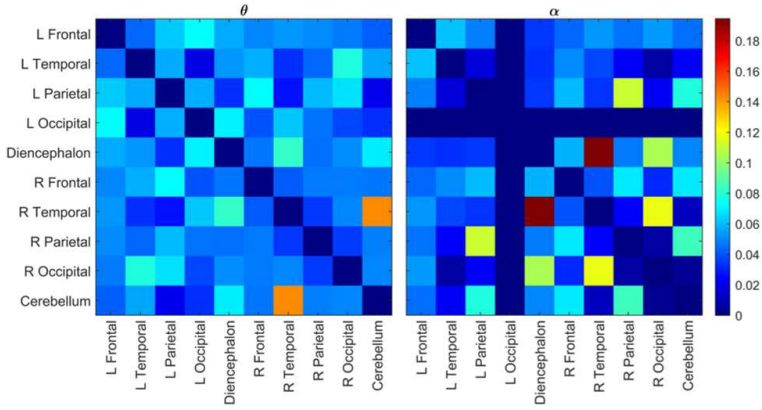


Рис. 1. Матрицы корреляций между компартментами мозга для субъекта sub-0108 в диапазонах частот тета и альфа.

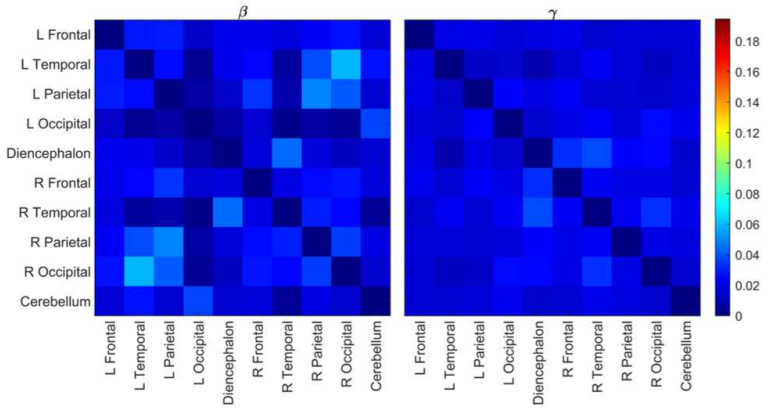


Рис. 2. Матрицы корреляций между компартментами мозга для субъекта sub-0108 в диапазонах частот бета и гамма.

Работа поддержана грантами РФФИ No. 19-07-00964, 20-07-00733, 20-07-00842.

[1] *Устинин М. Н., Рыкунов С. Д., Бойко А. И.* Корреляция между компартментами мозга при синдроме дефицита внимания и гиперактивности, рассчитанная методом виртуальных электродов по данным магнитной энцефалографии // Математическая биология и биоинформатика, 2020. Т. 15(2). С. 471–486.

Evaluation of correlations between brain compartments in attention deficit hyperactivity disorder using the method of virtual electrodes

Rykunov Stanislav[1]⋆ rykunov@impb.ru
Ustinin Mikhail[1] u_m_n@mail.ru
Boyko Anna[1] a.boyko@list.ru

[1]Pushchino, IMPB RAS - Branch of KIAM RAS

New method to study the correlation of the human brain compartments based on the magnetic encephalography data analysis was proposed. The time series for the correlation analysis are generated by the method of virtual electrodes. First, the multichannel time series of the subject with confirmed attention deficit and hyperactivity disorder are transformed into the functional tomogram - spatial distribution of the magnetic field sources structure on the discrete grid. This structure is provided by the inverse problem solution for all elementary oscillations, found by the Fourier transform. Each frequency produces the elementary current dipole located in the node of the 3D grid. The virtual electrode includes the part of space, producing the activity under study. The time series for this activity is obtained by the summation of the spectral power of all sources, covered by the virtual electrode.

In this work, two magnetic encephalograms (MEG) and two magnetic resonance imaging (MRI) from the OMEGA database were analyzed. For the analysis of correlations, the following frequency bands, or rhythms, were selected: theta - 4-8 Hz, alpha - 8-13 Hz, beta - 13-35 Hz, gamma - 35-50 Hz. Magnetic encephalogram in the delta rhythm frequency band is caused by the work of blood vessels, mainly located outside the brain. Therefore, the delta rhythm is not considered in the analysis of correlations.

Correlations were calculated between the following brain compartments: frontal lobe (left and right), temporal lobe (left and right), parietal lobe (left and right), occipital lobe (left and right), diencephalon, cerebellum.

To characterize the functional connectivity between the brain regions i and j, the square of the Pearson's correlation coefficient was used. First, the time series $P_i(t)$ and $P_j(t)$ are calculated on a discrete time grid with a recording frequency of 2400 Hz, where t varies from 0 to 300 seconds (the duration of the experiment). The series $P_i(t)$ and $P_j(t)$ are divided into epochs with a duration of 1 second, at each epoch the number l is calculated by the square of the Pearson's correlation coefficient by the formula:

$$\left(r_{ij}^{l}\right)^2 = \frac{cov^2(x,y)}{s_x^2 s_y^2},$$

where the samples $x = (P_i(t_1^l), ..., P_i(t_m^l)$ and $y = (P_j(t_1^l), ..., P_j(t_m^l)$, s_x^2 and s_y^2 - sample variances, l - epoch number, m - sample size. Then averaging over epochs is

performed:

$$c_{ij} = \frac{1}{300} \sum_{i=1}^{300} \left(r_{ij}^{l}\right)^2$$

The c_{ij} coefficient is displayed on square matrices characterizing the correlations between the i and j regions in different frequency ranges. The figures show the correlation matrices for one of the subjects in the frequency bands $\theta, \alpha, \beta, \gamma$.

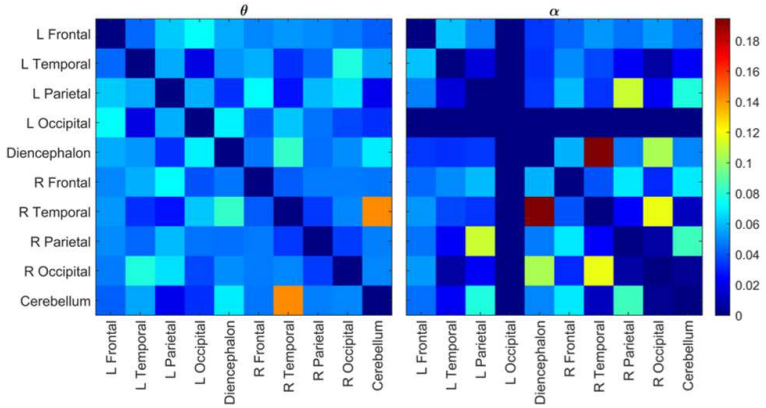


Fig. 1. Correlation matrices between brain compartments for subject sub-0108 in the theta and alpha frequency bands.

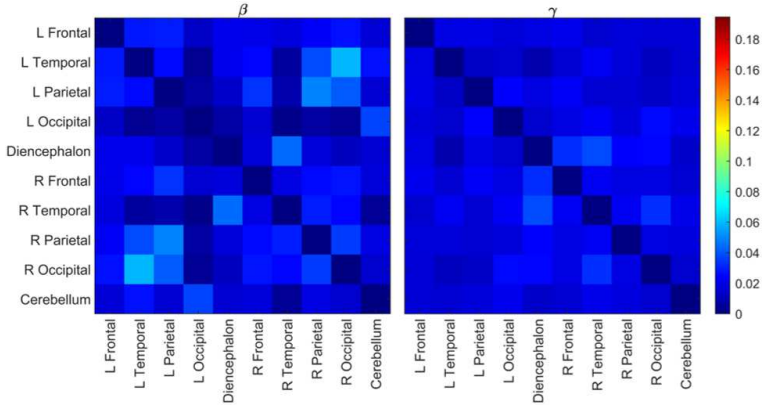


Fig. 2. Correlation matrices between brain compartments for subject sub-0108 in the beta and gamma frequency bands.

This research is funded by RFBR, grant 19-07-00964, 20-07-00733, 20-07-00842.

[1] *Ustinin M., Rykunov S., Boyko A.* Correlation of the Brain Compartments in the Attention Deficit and Hyperactivity Disorder Calculated by the Method of Virtual Electrodes from Magnetic Encephalography Data // Mathematical Biology and Bioinformatics, 2020. Vol. 15(2). Pp. 471–486.

Программный комплекс для компьютерного моделирования магнитной и электрической активности головного мозга человека и интеллектуального анализа модельных данных

Бойко Анна Ивановна[1]⋆ a.boyko@list.ru
Рыкунов Станислав Дмитриевич[1] rykunov@impb.ru
Устинин Михаил Николаевич[1] u_m_n@mail.ru

[1] Пущино, ИМПБ РАН – филиал ИПМ им. М.В. Келдыша РАН

Создан программный комплекс для компьютерного моделирования магнитной и электрической активности головного мозга человека и интеллектуального анализа модельных данных. Программный комплекс реализован в виде кроссплатформенного приложения на языке Python с использованием свободно распространяемых библиотек Numpy, hdf5storage и matplotlib. Программный комплекс обладает следующим функционалом: -интерактивное размещение источников поля в пространстве головы и редактирование амплитудно-временной зависимости; -пакетная загрузка большого количества источников; -моделирование шумов из существующих записей спонтанной активности и редактора внутри программной системы; -моделирование малоканальных планарных магнитометров различных порядков, с заданием формы прибора, количества датчиков и их параметров(расстояние между датчиками, количество и направление витков). Программный комплекс поддерживает следующие типы магнитометров: фиксированный 275-канальный градиометр первого порядка, фиксированный 148-канальный магнитометр, 7-канальный градиометр второго порядка с возможностью задания взаимного расположения модели головы и модели прибора. Также возможно моделирование данных электроэнцефалографии для приборов с различным числом и размещением датчиков. При помощи модельных данных будут совершенствоваться алгоритмы фильтрации и обработки экспериментальных данных. Разработанная программная система обладает следующими возможностями и характеристиками: - моделирование магнитных полей, порождаемых большим количеством источников; - моделирование потенциалов электрического поля на поверхности головы, порождаемых одним или множеством источников; - интерактивный процесс моделирования; - интуитивно понятный интерфейс. Структурно программа состоит из трех модулей: модуля ввода-вывода, расчетного модуля и модуля визуализации. Модуль ввода-вывода отвечает за загрузку моделей приборов, моделей головного мозга и параметров источников поля (расположение, направление и амплитудно-временная зависимость). После проведения расчетов данные о суммарной магнитоэнцефалограме или электроэнцефалограмме записываются в файл при помощи того же модуля. Расчетный модуль отвечает за непосредственный расчет поля и преобразование координат между индексной системой и системой головы. Модуль визуализации отвечает за отрисовку модели мозга, положения источников поля, графического представления амплитудно-временной зависимости источников поля и рассчи-

танных значений суммарного поля. Программный комплекс был использован при моделировании магнитных энцефалограмм для тестирования метода расчета корреляции между компартментами мозга. Два дипольных источника размещались в различных компартментах мозга. Задавались их временные зависимости и рассчитывалась корреляция между ними. После этого моделировалась магнитная энцефалограмма, производимая этими диполями на 275-канальном градиометре первого порядка производства VSM MedTech. Анализ модельной энцефалограммы помощью метода функциональной томографии показал корреляцию, равную найденной по временным рядам диполей. После этого были рассчитаны матрицы корреляций реальных магнитоэнцефалограмм, опубликованные в работе [1]. Программный комплекс также использовался для усреднения откликов на речевые стимулы и для совместного показа функциональных томограмм откликов и магнитно-резонансных томограмм в работе [2].

Работа поддержана грантами РФФИ No. 19-07-00964, 20-07-00733, 20-07-00842.

[1] *Устинин М. Н., Рыкунов С. Д., Бойко А. И.* Корреляция между компартментами мозга при синдроме дефицита внимания и гиперактивности, рассчитанная методом виртуальных электродов по данным магнитной энцефалографии // Математическая биология и биоинформатика, 2020. Т. 15(2). С. 471–486.

[2] *Устинин М. Н., Рыкунов С. Д., Бойко А. И., Тарасов Е. Ф., Журавлев И. В., Поликарпов М. А., Рябов Т. А., Филатов И. А., Юреня А. Ю., Панченко В. Я.* Изучение восприятия письменной речи методом функциональной томографии по данным электроэнцефалографии // Математическая биология и биоинформатика, 2021. Т. 16(1). С. 1–14.

Software for computer modeling of magnetic and electrical activity of the human brain and intelligent analysis of model data

Boyko Anna[1]⋆ a.boyko@list.ru
Rykunov Stanislav[1] rykunov@impb.ru
Ustinin Mikhail[1] u_m_n@mail.ru
[1]Pushchino, IMPB RAS - Branch of KIAM RAS

A software package for computer modeling of magnetic and electrical activity of the human brain and intelligent analysis of model data has been created. The software package is implemented as a cross-platform application in Python using the free Numpy, hdf5storage and matplotlib libraries. The software package has the following functionality: -interactive placement of field sources in the head space and editing of the amplitude-time dependence; -batch loading of a large number of sources; -simulation of noise from existing records of spontaneous activity and the editor inside the software system; -simulation of low-channel planar magnetometers of various orders, specifying the shape of the device, the number of sensors and their parameters (the distance between the sensors, the number and direction of turns). The software package supports the following types of magnetometers: fixed 275-channel first-order gradiometer, fixed 148-channel magnetometer, 7-channel second-order gradiometer with the ability to specify the relative position of the head and the device model. It is also possible to simulate electroencephalographic data for devices with different numbers and placement of sensors. Algorithms for filtering and processing experimental data will be improved with the help of model data. The developed software system has the following capabilities and characteristics: - modeling of magnetic fields generated by a large number of sources; - modeling the potentials of the electric field on the surface of the head, generated by one or more sources; - interactive modeling process; - intuitive interface. Structurally, the program consists of three modules: an input-output module, a calculation module, and a visualization module. The I/O module is responsible for loading device models, brain models and field source parameters (location, direction and amplitude-time dependence). After the calculations, the data on the total magnetoencephalogram or electroencephalogram are written to a file using the same module. The calculation module is responsible for directly calculating the field and transforming coordinates between the index system and the head system. The visualization module is responsible for drawing the brain model, the position of the field sources, a graphical representation of the amplitude-time dependence of the field sources and the calculated values of the total field. The software package was used to simulate magnetic encephalograms to test the method for calculating the correlation between brain compartments. Two dipole sources were located in different compartments of the brain. Their time dependencies were set and the correlation between them was calculated. The magnetic encephalogram produced by these dipoles was then simulated on a 275-channel first-order gradiometer manufactured by VSM MedTech. Analysis of the model encephalogram using

the functional tomography method showed a correlation equal to that found from the time series of dipoles. After that, the correlation matrices of real magnetoencephalograms were calculated, published in the paper [1]. The software package was also used for averaging responses to speech stimuli and for joint display of functional tomograms of responses and magnetic resonance imaging in the paper [2].

This research is funded by RFBR, grant 19-07-00964, 20-07-00733, 20-07-00842.

[1] *Ustinin M.,Rykunov S., Boyko A.* Correlation of the Brain Compartments in the Attention Deficit and Hyperactivity Disorder Calculated by the Method of Virtual Electrodes from Magnetic Encephalography Data // Mathematical Biology and Bioinformatics, 2020. V. 15(2). Pp. 471–486.

[2] *Ustinin M., Rykunov S., Boyko A., Tarasov E., Zhuravlev I., Polikarpov M., Ryabov T., Filatov I., Yurenya A., Panchenko V.* Study of the Perception of Written Speech Using Functional Tomography Based On Electroencephalography Data // Mathematical Biology and Bioinformatics, 2021. Vol. 16(1). Pp. 1–14.

Изучение особенностей пространственного распределения откликов на различные речевые стимулы по данным электроэнцефалографии

Устинин Михаил Николаевич[1]⋆ u_m_n@mail.ru
Бойко Анна Ивановна[1] a.boyko@list.ru
Рыкунов Станислав Дмитриевич[1] rykunov@impb.ru

[1] Пущино, ИМПБ РАН – филиал ИПМ им. М.В. Келдыша РАН

Изучались спектральные и пространственные характеристики электроэнцефалограмм, регистрируемых при восприятии письменной речи. Для экспериментального исследования было сформировано четыре группы, содержащих по 100 слов: слова с положительной эмоциональной оценкой, слова с отрицательной эмоциональной оценкой, слова с конкретными значениями, и слова с абстрактными значениями. Для каждой группы с испытуемыми проводился отдельный эксперимент. Слова были представлены белым текстом на черном фоне, каждое слово предъявлялось в течение 1000 мс, после предъявления стимула следовала пауза длительностью 500 мс. Активность мозга регистрировалась при помощи электроэнцефалографа с 19 отведениями, расставленными по схеме 10–20. Для детального количественного анализа этой активности использовался метод функциональной томографии мозга по данным электроэнцефалографии. Этот метод опирается на преобразование Фурье многоканальных данных энцефалографии и локализацию отдельных спектральных компонент. Метод позволяет с высокой точностью выделить и устойчиво локализовать в пространстве различные спектральные особенности активности мозга, изучаемой в экспериментах по исследованию речи. Анализировалась полоса частот от 8 до 30 Гц, для всех спектральных компонент в этой полосе была решена обратная задача в приближении эквивалентного токового диполя в однослойном сферическом проводнике, без каких-либо ограничений положения источника. В результате были построены трехмерные карты активности – функциональные структуры мозга. Представление этих функциональных структур на магнитно-резонансной томограмме позволяет изучать частотные и пространственные особенности откликов на различные речевые стимулы.

Работа поддержана грантами РФФИ No. 19-07-00964, 20-07-00733, 20-07-00842.

[1] *Устинин М. Н., Рыкунов С. Д., Бойко А. И., Тарасов Е. Ф., Журавлев И. В., Поликарпов М. А., Рябов Т. А., Филатов И. А., Юреня А. Ю., Панченко В. Я.* Изучение восприятия письменной речи методом функциональной томографии по данным электроэнцефалографии // Математическая биология и биоинформатика, 2021. Т. 16(1). С. 1–14.

Study of the features of spatial distribution of responses to various speech stimuli based on the electroencephalography

Ustinin Mikhail[1]⋆ u_m_n@mail.ru
Boyko Anna[1] a.boyko@list.ru
Rykunov Stanislav[1] rykunov@impb.ru
[1]Pushchino, IMPB RAS - Branch of KIAM RAS

The spectral and spatial characteristics of the electroencephalograms recorded during the perception of written speech were studied. For the experimental study, four groups were formed, each containing 100 words: words with a positive emotional rating, words with a negative emotional rating, words with concrete meanings, and words with abstract meanings. A separate experiment was conducted for each group with the subjects. Words were represented by white text on a black background, each word was presented for 1000 ms, after the presentation of the stimulus there was a pause of 500 ms. Brain activity was recorded using an electroencephalograph with 19 leads, arranged according to the 10–20 scheme. For detailed quantitative analysis of this activity, method of functional tomography of the brain, based on electroencephalography data, was used. This method is based on the Fourier transform of multichannel encephalographic data and the localization of individual spectral components. The method makes it possible to single out and stably localize in space various spectral features of the brain activity studied in experiments on speech research. The frequency band from 8 to 30 Hz was analyzed; for all spectral components in this band, the inverse problem was solved in the approximation of an equivalent current dipole in a single-layer spherical conductor, without any restrictions on the position of the source. As a result, three-dimensional maps of activity were built - the functional structures of the brain. The presentation of these functional structures on magnetic resonance imaging allows one to study the frequency and spatial characteristics of responses to various speech stimuli.

This research is funded by RFBR, grant 19-07-00964, 20-07-00733, 20-07-00842.

[1] *Ustinin M. N., Rykunov S. D., Boyko A. I., Tarasov E. F., Zhuravlev I. V., Polikarpov M. A., Ryabov T. A., Filatov I. A., Yurenya A. Yu., Panchenko V. Ya.* Study of the Perception of Written Speech Using Functional Tomography Based On Electroencephalography Data // Mathematical Biology and Bioinformatics, 2021. Vol. 16(1). Pp. 1–14.

Алгоритм точного обучения частотного фильтра для задачи малоракурсной компьютерной томографии

Ямаев Андрей Викторович[1,2]⋆ rewin1996@gmail.com
Чуличков Алексей Иванович[2] achulichkov@gmail.com
[1]Москва, ООО "Смарт Энжинс Сервис"
[2]Москва, Московский государственный университет имени М. В. Ломоносова

Рентгеновская компьютерная томография - метод, позволяющий неразрушающим способом изучить внутреннюю структуру объекта. Это достигается путем регистрации проекций, определяющих степень поглощения объектом проходящего сквозь него под различными углами рентгеновского излучения. Такой метод часто применяется в медицине. Однако рентгеновское излучение является ионизирующим, поэтому стремятся минимизировать получаемую пациентом дозу поглощенного рентгеновского излучения. Это возможно сделать одним из двух способов: уменьшением времени экспозиции одной проекции или уменьшением числа проекций. В первом случае проекции будут содержать зашумленные недостоверные данные. Во-втором случае задача восстановления является недоопределенной, но мы точно можем определить изображение структуры объекта под определенным углом. Существуют различные алгоритмические и нейросетевые методы решения задачи малоракурсной томографии, использующие возможность достоверного знания структуры объекта под определенными углами. В работе [1] предлагается использовать SIRT - итеративный метод, основанный на градиентном спуске. Метод показывает хорошие результаты реконструкции в случае малого числа углов. Однако количество итераций, необходимых для качественной реконструкции, может составлять несколько сотен, а время расчета одной реконструкции на обычном компьютере исчисляться часами. Существуют различные нейросетевые методы, модифицирующие SIRT-схему так, чтобы уменьшить число необходимых итераций [2]. Однако время расчетов все равно остается слишком большим. В работах [4, 5] предлагается модифицировать этап свертки в алгоритме Filtered Back Projection (FBP) [3] таким образом, чтобы повысить качество малоракурсной реконструкции за счет учета дискретности задачи и ограниченного числа проекций. Эти методы позволяют за малое по сравнению с SIRT время получить реконструкции с нулевой ошибкой репроецирования; ошибка определяется l2 расстоянием между регистрируемыми проекциями и проекциями, рассчитанными из реконструкции. Однако такие подходы лишь интерполируют частотный фильтр. В данной работе предлагается алгоритм точного попиксельно подбора частотного фильтра для этапа свертки алгоритма FBP. В алгоритме частотный фильтр представляется в виде набора обучаемых коэффициентов в рамках библиотеки машинного обучения Pytorch. Выбор этих коэффициентов был выполнен с помощью набора фантомов типа foam. Показано, что эти же коэффициенты применимы для фантомов другого типа. Продемонстрированы результаты работы алгоритма на реальных данных

и показано, что предложенный метод обеспечивает меньшую ошибку репроекции, чем аналогичные методы [4, 5].

Работа поддержана грантами РФФИ No. 18-29-26020 и No. 19-01-00790.

[1] *Trampert J., Leveque J.* Simultaneous iterative reconstruction technique: physical interpretation based on the generalized least squares solution // Journal of Geophysical Research: Solid Earth, 1990. Vol. 95(B8). Pp. 12553–12559.

[2] *Adler J., Öktem O.* Learned primal-dual reconstruction // IEEE transactions on medical imaging, 2018. Vol. 37(6). Pp. 1322–1332.

[3] *Lauritsch G., Härer W.* Theoretical framework for filtered back projection in tomosynthesis // Medical Imaging 1998: Image Processing, 1998. Vol. 3338. Pp. 1127–1137.

[4] *Ganguly P. et al* Improving reproducibility in synchrotron tomography using implementation-adapted filters // arXiv:2103.08288, 2021.

[5] *Batenburg K., Plantagie L.* Fast approximation of algebraic reconstruction methods for tomography // IEEE Transactions on Image Processing, 2012. Vol. 21(8). Pp. 3648–3658.

Algorithm of exact frequency filter finding for few view tomography problem

Yamaev Andrei[1,2]⋆ rewin1996@gmail.com
Chulichkov Alexey[2] achulichkov@gmail.com
[1]Moscow, Smart Engines Services Ltd
[2]Moscow, Moscow State University

X-ray computed tomography is a method that allows non-destructive examination of the internal structure of an object. This is accomplished by registration projections which contain data of X-rays absorption by object by various angles. This method is often used in medicine. However, X-ray radiation is the ionizing ray, therefore efforts are made to minimize the dose of absorbed X-ray radiation received by the patient. This can be done in one of two ways: by decreasing the exposure time of one projection or by decreasing the number of projections. In the first case, the projection will indicate noisy unreliable data. In the second case, the problem of reconstruction is undefined, but we can accurately determine the image of the object's structure at a certain angle. There are various algorithmic and neural network methods for solving the problem of small-angle tomography, using reliable knowledge of the structure of an object at projections under certain angles. For example, work [1] suggests using SIRT, an iterative method based on gradient descent. This method shows good results of reconstruction in the case of a small number of angles. However, the number of iterations required for a high-quality reconstruction can be several hundred, and the reconstruction time for one reconstruction on an ordinary computer can be calculated in hours. There are various neural network methods [2] that modify the SIRT scheme to reduce the number of iterations required. However, the calculation time is still too long. In the works [4, 5] there is changed the convolution stage in the Filtered Back Projection (FBP) [3] algorithm in a way as to improve the quality of discrete low-angle reconstruction. These methods allow obtaining reconstructions with zero reprojection error in a short working time compared to SIRT; error in determining the distance l2 between the recorded projections and the projections calculated from the reconstruction. However, such filter approaches only interpolate the frequency filter. In this paper, the algorithm of the FBP algorithm. In the algorithm, the frequency filter is represented as a set of trainable coefficients in the Pytorch machine learning library. The selection of these coefficients was performed using a set of foam-type phantoms. It is shown that the same coefficients are applicable for phantoms of a different type. The results of the algorithm for working on real data are demonstrated and it is shown that the proposed method provides a smaller reprojection error than similar methods [4, 5].

This research is funded by RFBR, grants 18-29-26020 and 19-01-00790.

[1] *Trampert J., Leveque J.* Simultaneous iterative reconstruction technique: physical interpretation based on the generalized least squares solution // Journal of Geophysical Research: Solid Earth, 1990. Vol. 95(B8). Pp. 12553–12559.

[2] *Adler J., Öktem O.* Learned primal-dual reconstruction // IEEE transactions on medical imaging, 2018. Vol. 37(6). Pp. 1322–1332.

[3] *Lauritsch G., Härer W.* Theoretical framework for filtered back projection in tomosynthesis // Medical Imaging 1998: Image Processing, 1998. Vol. 3338. Pp. 1127–1137.

[4] *Ganguly P. et al* Improving reproducibility in synchrotron tomography using implementation-adapted filters // arXiv:2103.08288, 2021.

[5] *Batenburg K., Plantagie L.* Fast approximation of algebraic reconstruction methods for tomography // IEEE Transactions on Image Processing, 2012. Vol. 21(8). Pp. 3648–3658.

Новый метод определения скорости тяжелого иона, основанный на математическом формализме субъективного моделирования

Фаломкина Олеся Владимировна[1]⋆ olesya.falomkina@gmail.com
Пытьев Юрий Петрович[1] yuri.pytyev@physics.msu.ru
Чуличков Алексей Иванович[1] achulichkov@gmail.com
Пятков Юрий Васильевич[2]⋆ yvp_nov@mail.ru
Жучко Владимир Евгеньевич[3] zhuchko@jinr.ru
Каманин Дмитрий Владимирович[3] kamanin@jinr.ru
Горяйнова Зоя Игоревна[3] zoyag@yandex.ru

[1]Москва, МГУ им. М. В. Ломоносова
[2]Москва, НИЯУ МИФИ
[3]Дубна, ОИЯИ

В докладе рассмотрен новый метод и алгоритм решения задачи определения скорости тяжелого иона с помощью полупроводникового детектора (PIN диода) [1], [2], использующие математический формализм субъективного моделирования (МФСМ) [3], [4], [5], позволяющий математически сформулировать как субъективную модель исследуемого объекта, так и субъективную математическую модель его измерений и их субъективной интерпретации. С помощью МФСМ можно математически формализовать субъективные суждения модельера-исследователя[1] (м-и) о физических свойствах исследуемого объекта и о средствах его (м-и) измерений, о математических свойствах шума и т. п.; вся подобная субъективная информация основана на научном опыте м-и и на его интуиции учёного.

В экспериментальной практике скорость иона измеряют «по времени пролета», т. е. измеряется время пролета ионом некоторого известного расстояния, называемого пролетной базой. Для измерения времени пролета необходимо получить временные отметки «старт» и «стоп», соответствующие моментам начала и окончания движения иона вдоль пролетной базы. Для получения таких отметок используют временные детекторы. Отметку «стоп» часто получают с полупроводникового детектора, например, с так называемого PIN диода. При попадании иона в диод на выходе диода появляется сигнал (импульс напряжения), который можно представить как сумму собственно импульса напряжения, вызванного регистрируемым ионом, и аддитивного вероятностного шума. Физика взаимодействия тяжелого иона с полупроводником такова, что форма сигнала представляет сначала медленно растущую функцию, график которой *неизвестен*, выходящую потом на почти линейную зависимость (длина этого участка также *неизвестна*). Требуется определить момент времени, когда ион

[1]Модельер-исследователь владеет современными методами математического моделирования, включая математический формализм субъективного моделирования, и знаком с предметной областью как её исследователь.

попал в детектор («абсолютную временную привязку») – т. е. собственно начало сигнала, при том, что начальная часть фронта импульса лежит внутри области с высоким уровнем шума («шумовой дорожки»).

Для решения задачи определения скорости тяжелого иона разработан и реализован алгоритм, основанный на математическом формализме субъективного моделирования, позволяющий восстановить неизвестную форму фронта сглаживающим сплайном со следующим специальным условием: начальная часть сплайна (слева) задается уравнением параболы, а вершина этой параболы должна лежать на усредненной шумовой линии, поскольку в отсутствие шума фронт начинает расти с нулевой линии. Для определения оптимального сглаживающего фактора сплайна использован *субъективный критерий оптимальности* [5].

Метод решения задачи определения скорости тяжелого иона с помощью полупроводникового детектора (PIN диода) прошел апробацию в Лаборатории ядерных реакций Объединенного института ядерных исследований (Дубна) на реальных ядернофизических данных.

Работа поддержана грантами РФФИ No. 18-07-00424а, No. 19-29-09044мк.

[1] *Pyatkov Yu., Kamanin D., Oertzen W. et al.* Collinear cluster tri-partition of 252Cf (sf) and in the 235U(nth, f) reaction // Eur. Phys. J. A., 2010. Vol. 45(1). Pp. 29–37.

[2] *Pyatkov Yu., Kamanin D., Strekalovsky A. et al.* New approaches to determination of the heavy ion's mass in measurements with PIN diodes // Bull. Russ. Acad. Sci.: Phys., 2018. Vol. 82(6). Pp. 804–807.

[3] *Pyt'ev Y.* Modeling of subjective judgments made by a researcher-modeler about the model of the research object // Mathematical Models and Computer Simulations, 2013. Vol. 5(6). Pp. 538–557.

[4] *Пытьев Ю. П.* Вероятность, возможность и субъективное моделирование в научных исследованиях // Математические и эмпирические основы, 2018. 268 с.

[5] *Pyt'ev Yu., Falomkina O., Shishkin S.* Subjective Restoration of Mathematical Models for a Research Object, Its Measurements, and Measurement-Data Interpretation // Pattern Recognition and Image Analysis, 2019. Vol. 29(4). Pp. 577–591.

A new method for determining the velocity of a heavy ion based on the mathematical formalism of subjective modeling

Falomkina Olesya[1]⋆ olesya.falomkina@gmail.com
Pyt'ev Yuri[1] yuri.pytyev@physics.msu.ru
Chulichkov Alexey[1] achulichkov@gmail.com
Pyatkov Yuri[2]⋆ yvp_nov@mail.ru
Zhuchko Vladimir[3] zhuchko@jinr.ru
Kamanin Dmitry[3] kamanin@jinr.ru
Goryaynova Zoya[3] zoyag@yandex.ru

[1]Moscow, Lomonosov Moscow State University
[2]Moscow, National Research Nuclear University MEPHI
[3]Dubna, JINR

The report discusses a new method and algorithm for solving the problem of determining the velocity of a heavy ion using a semiconductor detector (PIN diode) [1], [2], using the mathematical formalism of subjective modeling (MFSM) [3], [4], [5], which allows to mathematically formulate both a subjective model of the object under study and a subjective mathematical model of its measurements and their subjective interpretation. With the help of the MFSM, it is possible to mathematically formalize the subjective judgments of a researcher–modeler[1] (r-m) about the physical properties of the object under study and about the means of its (r-m) measurements, about the mathematical properties of noise, etc.; all such subjective information is based on the scientific experience of r-m and on his scientific intuition.

In experimental practice, the ion velocity is measured by "time-of-flight", i.e., the time of flight by an ion of a certain known distance, called the span base, is measured. To measure the time-of-flight, it is necessary to obtain the timestamps "start" and "stop" corresponding to the moments of the beginning and end of the ion movement along the flight path. Time pick-off detectors are used to obtain timestamps. The "stop" timestamp is often obtained from a semiconductor detector, for example, from a so-called PIN diode. When an ion hits the diode, a signal (voltage pulse) appears at the output of the diode, which can be represented as the sum of the actual voltage pulse caused by the recorded ion and additive probabilistic noise. The physics of the interaction of a heavy ion with a semiconductor is such that the waveform first represents a slowly growing function, the graph of which is *unknown*, then comes out to an almost linear dependence (the length of this section is also *unknown*). It is required to determine the moment of time when the ion hits the detector ("absolute time reference") - i.e., the actual beginning of the signal, despite the fact that the initial part of the pulse leading edge lies inside the area with a high noise level.

[1]The researcher–modeleris familiar with modern methods of mathematical modeling, including the mathematical formalism of subjective modeling, and is familiar with the subject area as its researcher.

To solve the problem of determining the velocity of a heavy ion, an algorithm based on the mathematical formalism of subjective modeling has been developed and implemented, which allows to restore the unknown shape of the pulse leading edge by a smoothing spline with the following special condition: the initial part of the spline (on the left) is given by the parabola equation, and the vertex of this parabola should lie on the averaged noise line, since in the absence of noise the leading edge begins to grow from the zero line. To determine the optimal smoothing factor of the spline, *subjective optimality criterion* [5] was used.

Correctness of new time pick-off algorithm was tested in experiment at the accelerator in the Laboratory of Nuclear Reactions of the Joint Institute for Nuclear Research (Dubna).

This research is funded by RFBR, grants 18-07-00424, 19-29-09044.

[1] *Pyatkov Yu., Kamanin D., Oertzen W. et al.* Collinear cluster tri-partition of 252Cf (sf) and in the 235U(nth, f) reaction // Eur. Phys. J. A., 2010. Vol. 45(1). Pp. 29–37.

[2] *Pyatkov Yu., Kamanin D., Strekalovsky A. et al.* New approaches to determination of the heavy ion's mass in measurements with PIN diodes // Bull. Russ. Acad. Sci.: Phys., 2018. Vol. 82(6). Pp. 804–807.

[3] *Pyt'ev Y.* Modeling of subjective judgments made by a researcher-modeler about the model of the research object // Mathematical Models and Computer Simulations, 2013. Vol. 5(6). Pp. 538–557.

[4] *Pyt'ev Yu.* Probability, possibility and subjective modeling in scientific research // Mathematical and empirical foundations, 2018. 268 p.

[5] *Pyt'ev Yu., Falomkina O., Shishkin S.* Subjective Restoration of Mathematical Models for a Research Object, Its Measurements, and Measurement-Data Interpretation // Pattern Recognition and Image Analysis, 2019. Vol. 29(4). Pp. 577–591.

Использование классификации при решении регрессионной обратной задачи разведочной геофизики как способ повышения устойчивости решения к шумам в данных

Исаев Игорь Викторович[1,2]⋆ isaev_igorl@mail.ru
Оборнев Иван Евгеньевич[1] o_ivano@mail.ru
Оборнев Евгений Александрович[3] eugenyo@mail.ru
Родионов Евгений Александрович[3] evgeny_980@list.ru
Шимелевич Михаил Ильич[3] Shimelevich-M@yandex.ru
Доленко Сергей Анатольевич[1] dolenko@srd.sinp.msu.ru

[1]Москва, НИИ ядерной физики имени Д.В. Скобельцына
[2]Москва, Институт радиотехники и электроники им. В.А.Котельникова РАН
[3]Москва, Российский государственный геологоразведочный университет имени Серго Орджоникидзе

Обратные задачи разведочной геофизики заключаются в восстановлении пространственного распределения свойств среды в толще Земли по геофизическим полям, измеренным на ее поверхности. В данной работе рассматриваются обратные задачи гравиметрии, магнитометрии и магнитотеллурического зондирования, а также их комплексирование, т. е. одновременное использование нескольких геофизических полей для восстановления желаемого распределения.

Для реализации комплексирования использовалась 4-х слойная 2D-модель, где обратная задача заключалась в определении нижних границ слоев, а каждый слой характеризовался переменными значениями глубины нижней границы по разрезу и фиксированными значениями плотности, намагниченности и удельного электрического сопротивления как для слоя, так и для всего набора данных.

Набор данных для реализации нейросетевого решения обратной задачи был сгенерирован путем решения прямой задачи, где для каждого примера распределение значений глубины слоя задавалось случайным образом в заданном диапазоне и с заданным шагом, т.е. использовались дискретные значения из определенного набора.

В данной работе рассматривался подход, предполагающий использование нейронных сетей для решения задачи мультиклассовой классификации, где метки классов соответствуют дискретным значениям определяемых глубин слоев. Результаты использования такого подхода сравнивались с результатами нейросетевого решения той же обратной задачи в регрессионной постановке. Исследовалась устойчивость нейросетевого решения к шумам в данных, как для регрессионной постановки задачи, так и для классификационной. Рассматри-

вались аддитивные и мультипликативные шумы различных уровней, имеющие нормальное или равномерное распределение.

Показано, что использование подхода, основанного на решении задачи классификации, позволяет повысить устойчивость результатов к шумам в данных относительно решения рассматриваемой задачи как задачи регрессии. Данный эффект наблюдается для большинства рассматриваемых геофизических методов и их сочетаний, а также для всех рассматриваемых типов и статистик шума.

Исследование выполнено за счёт гранта Российского Научного фонда, проект No. 19-11-00333.

Using classification in solving regression inverse problem of exploration geophysics as a way to improve the resilience of the solution to noise in data

Isaev Igor[1,2]⋆ isaev_igorl@mail.ru
Obornev Ivan[1] o_ivano@mail.ru
Obornev Eugeny[3] eugenyo@mail.ru
Rodionov Eugeny[3] evgeny_980@list.ruu
Shimelevich Mikhail[3] Shimelevich-M@yandex.ru
Dolenko Sergey[1] dolenko@srd.sinp.msu.ru

[1]Moscow, D.V. Skobeltsyn Institute of Nuclear Physics
[2]Moscow, Kotelnikov Institute of Radioengineering and Electronics, Russian Academy of Sciences
[3]Moscow, Sergo Ordjonikidze Russian State University for Geological Prospecting

The inverse problems of exploration geophysics are to reconstruct the spatial distribution of the properties of the medium in the Earth's thickness from the geophysical fields measured on its surface. In this paper, we consider the inverse problems of gravimetry, magnetometry, and magnetotelluric sounding, as well as their integration, i.e., simultaneous use of several geophysical fields to restore the desired distribution.

To implement the integration, a 4-layer 2D model was used, where the inverse problem was to determine the lower boundary of the layers, and each layer was characterized by variable values of the depth of the lower boundary along the section and fixed values of density, magnetization, and resistivity, both for the layer and for the entire data set.

The data set for the implementation of the neural network solution of the inverse problem was generated by solving the direct problem, where for each pattern, the distribution of layer depth values was set randomly in a given range and with a given step, i.e. discrete values from a certain set were used.

In this paper, an approach was considered that involves the use of neural networks to solve the problem of multiclass classification, where class labels correspond to the discrete values of the determined layer depths. The results of using this approach are compared with the results of the neural network solution of the same inverse problem within the formulation of the regression problem. The resilience of the neural network solution to noise in the data was investigated, both for the regression formulation of the problem and for the classification one. Additive and multiplicative noise of various levels with normal or uniform distribution was considered.

It is shown that the use of the approach based on the solution of the classification problem allows improving the resilience of the solution to noise in data compared to the solution of the problem under consideration as a regression problem. This effect

is observed for most of the considered geophysical methods and their combinations, as well as for all considered types and statistics of noise.

This study has been performed at the expense of the grant of the Russian Science Foundation (project no. 19-11-00333).

Восстановление решений уравнений типа Урысона

Белозуб Владимир Антонович[1] disstroier@mail.ru
Козлова Маргарита Геннадьевна[1] art-inf@mail.ru
Лукъяненко Владимир Андреевич[1]⋆ art-inf@yandex.ru

[1]Симферополь, ФГАОУ ВО «Крымский федеральный университет имени В. И. Вернадского»

Нелинейные уравнения типа Урысона с дельтаобразным или осциллирующим ядром и их дискретные аналоги представляют теоретический интерес, имеют приложения в качестве моделей восстановления поверхности по результатам косвенных измерений (дистанционное зондирование, сейсмо-, магниторазведка).

В фиксированный момент сканирующее устройство (антенна) регистрирует дельтаобразный импульс, который отражается от изучаемой поверхности. Характер отражения может быть заданным или искомым. Задача решения таких уравнений является некорректно поставленной. Ситуация мало упрощается даже в случае поиска характерных точек поверхности (блестящих), вклад которых в правую часть наибольший. Данное исследование является продолжением работы [1], где приведен достаточно подробный обзор модельных уравнений, методов и регуляризирующих алгоритмов.

В одномерном случае уравнение представляет строку двумерной поверхности, т. е. имеем близкие уравнения. Решение для одного уравнения является приближенным для итерационного алгоритма построения решения близкого уравнения. Соответствующие теоремы, шаги алгоритма и оценки погрешности приведены в работе [1] для уравнения Урысона вида

$$A(f,\tau) \equiv \int\limits_{\mathbb{R}} f(\xi - s)n(t - \tau(s))ds = u(t,\xi), \quad t,\xi \in \mathbb{R}$$

с ядром $n(t)$ в виде дельтаобразной функции. Здесь f – функция, характеризующая свойства отражения от поверхности, τ – искомая функция, зависящая от поверхности $h(\xi)$. Функция f может зависеть от разности аргументов $(\xi - s)$. В этом случае преобразование Фурье правой части $v = \mathcal{F}u$ будет зависеть от (ξ, w). Искомыми могут быть как функция f, так и τ. Неизвестная функция $\tau(s)$ характеризует время прохождения двойного расстояния импульсом от точки наблюдения до искомой поверхности и обратно. Для единственности решения необходима система таких уравнений или регистрация отраженного сигнала в процессе движения точки наблюдения по заданной траектории.

Применяя к уравнению преобразование Фурье, приходим к уравнению Урысона с осциллирующим ядром. Рассмотрим частный случай уравнения Урысона

с осциллирующим ядром:

$$A(f,\tau) \equiv \int_a^b f(s)e^{iw\tau(s)}ds = v(w), \quad c \leqslant w \leqslant d. \tag{1}$$

Дискретный вариант имеет вид

$$\sum_{k=1}^{p} A_k f_k e^{i[c+(n-1)h_w]\tau_k} = x[n], \quad n = \overline{1,p}, \tag{2}$$

$$w_n = c + (n-1)\frac{d-c}{p-1} = c + (n-1)h_w, \;\; n = \overline{1,p},$$
$$s_k = a + (k-1)\frac{b-a}{p-1} = a + (k-1)h_s, \;\; k = \overline{1,p},$$
$$\tau_k = \tau(s_k), \;\; x[n] = v(w_n), \;\; f_k = f(s_k).$$

Обозначая

$$h_k = A_k f_k e^{ic\tau_k}, \quad z_k = e^{ihw\tau_k}, \tag{3}$$

приходим к нелинейному уравнению

$$\sum_{k=1}^{p} h_k z_k^{n-1} = x[n], \quad n = \overline{1,p} \tag{4}$$

с искомыми величинами $h_1, h_2, ..., h_p, z_1, z_2, ..., z_p$.

Если $z = (z_1, z_2, ..., z_p)$ и $h = (h_1, h_2, ..., h_p)$ найдены, то τ_k и f_k находятся по формулам (3).

Решение уравнения (4) находится методом Прони в случае точно заданных величин $x[n]$.

Входящие в (4) p уравнений, где $1 \leqslant n \leqslant p$, можно записать в матричной форме с матрицей Вандермонда.

Вектор $x[n]$ является решением однородного линейного разностного уравнения с постоянными коэффициентами.

$$\sum_{m=0}^{p} a[m]x[n-m] = 0. \tag{5}$$

Полином

$$\varphi(z) = \sum_{m=0}^{p} a[m]z^{p-m}, \tag{6}$$

где $a[m]$ – комплексные коэффициенты ($a[0] = 1$), ассоциированный с этим линейным разностным уравнением, называется характеристическим. Матрица в

уравнении (5) имеет структуру матрицы Тёплица, поэтому решение может быть получено с помощью известных программ.

Используя корни z_k, можно определить значения искомой функции τ_k и f_k по формулам (3). Так как задача является некорректной, то на всех этапах алгоритма необходимо проводить регуляризацию. Например, для уравнения (5) в операторной форме $XA = U$ регуляризирующий алгоритм может иметь вид $\alpha X + X^*XA = X^*U$, где α – параметр регуляризации.

Развитием метода Прони является метод матричных пучков [2], [3], который направлен на повышение точности нахождения z_k. Метод применим к уравнениям типа Урысона.

[1] *Belozub V., Kozlova M., Lukianenko V.* Approximated solution algorithms for Urysohn-type equations // Journal of Physics: Conference Series, 2021.

[2] *Hua Y., Sarkar T.* Matrix Pencil Method for Estimating Parameters of Exponentially Damped/Undamped Sinusoids in Noise // IEEE Transactions on Acoustics, Speech and Signal Proc, 1990. Vol. 38(5). Pp. 814–824.

[3] *Larin V.* The use of matrix pencils in an identification problem // Journal of Automation and Information Sciences, 1996. Vol. 28(384). Pp. 53-62.

Recovery of solutions of Urysohn-type equations

Belozub Vladimir[1] disstroier@mail.ru
Kozlova Margarita[1] art-inf@mail.ru
Lukianenko Vladimir[1]⋆ art-inf@yandex.ru
[1]Simferopol, V. I. Vernadsky Crimean Federal University

Nonlinear Urysohn-type equations with a delta-shaped or oscillating core and their discrete analogues are of theoretical interest, have applications as models for surface restoration based on the results of indirect measurements (remote sensing, seismic, magnetic exploration).

At a fixed moment, the scanning device (antenna) registers a delta-shaped pulse that is reflected from the studied surface. The nature of the reflection can be set or desired. The problem of solving such equations is incorrectly posed. The situation is a little simplified even in the case of searching for characteristic points of the surface (shiny), whose contribution to the right side is the greatest. This study is a continuation of the work [1], which provides a fairly detailed overview of model equations, methods and regularizing algorithms.

In the one-dimensional case, the equation represents a string of a two-dimensional surface, i.e. we have close equations. The solution for one equation is approximate for an iterative algorithm for constructing a solution to a close equation. The corresponding theorems, algorithm steps and error estimates are given in the paper [1] for the equation Urysohn view

$$A(f,\tau) \equiv \int\limits_{\mathbb{R}} f(\xi - s)n(t - \tau(s))ds = u(t,\xi), \quad t,\xi \in \mathbb{R}$$

with the kernel $n(t)$ in the form of a delta-like function. Here f is a function characterizing the properties of reflection from the surface, τ is the desired function depending on the surface $h(\xi)$. The function f may depend on the difference of the arguments $(\xi - s)$. In this case, the Fourier transform of the right side of $v = \mathcal{F}u$ will depend on (ξ, w). The desired function can be either f or τ. Unknown function $\tau(s)$ characterizes the time of passage of a double distance by a pulse from the observation point to the desired surface and back. For the uniqueness of the solution, a system of such equations or registration of the reflected signal during the movement of the observation point along a given trajectory is necessary.

Applying the Fourier transform to the equation, we come to the Urysohn equation with an oscillating core. Consider a special case of the Urysohn equation with an oscillating kernel:

$$A(f,\tau) \equiv \int\limits_{a}^{b} f(s)e^{iw\tau(s)}ds = v(w), \quad c \leqslant w \leqslant d. \tag{1}$$

The discrete variant has the form

$$\sum_{k=1}^{p} A_k f_k e^{i[c+(n-1)h_w]\tau_k} = x[n], \quad n = \overline{1,p}, \tag{2}$$

$w_n = c + (n-1)\dfrac{d-c}{p-1} = c + (n-1)h_w, \;\; n = \overline{1,p},$

$s_k = a + (k-1)\dfrac{b-a}{p-1} = a + (k-1)h_s, \;\; k = \overline{1,p},$

$\tau_k = \tau(s_k), \;\; x[n] = v(w_n), \;\; f_k = f(s_k).$

Denoting

$$h_k = A_k f_k e^{ic\tau_k}, \quad z_k = e^{ihw\tau_k}, \tag{3}$$

we come to the nonlinear equation

$$\sum_{k=1}^{p} h_k z_k^{n-1} = x[n], \quad n = \overline{1,p} \tag{4}$$

with the required values $h_1, h_2, ..., h_p, z_1, z_2, ..., z_p$.

If $z = (z_1, z_2, ..., z_p)$ and $h = (h_1, h_2, ..., h_p)$ are found, then τ_k and f_k are found by the formulas (3).

The solution of the equation (4) is found by the Prony method in the case of exactly specified quantities $x[n]$.

The p equations included in (4), where $1 \leqslant n \leqslant p$, can be written in matrix form with the Vandermond matrix.

The vector $x[n]$ is the solution of a homogeneous linear difference equation with constant coefficients.

$$\sum_{m=0}^{p} a[m]x[n-m] = 0. \tag{5}$$

The polynomial

$$\varphi(z) = \sum_{m=0}^{p} a[m]z^{p-m}, \tag{6}$$

where $a[m]$ are complex coefficients ($a[0] = 1$) associated with this linear difference equation is called characteristic. The matrix in the equation (5) has the structure of the Toeplitz matrix, so the solution can be obtained using well-known programs.

Using the roots of z_k, you can determine the values of the desired function τ_k and f_k by the formulas (3). Since the problem is incorrect, it is necessary to carry out regularization at all stages of the algorithm. For example, for the equation (5) in the operator form $XA = U$, the regularization algorithm can have the form $\alpha X + X^*XA = X^*U$, where α is the regularization parameter.

The development of the Proni method is the matrix beam method [2], [3], which aims to improve the accuracy of finding z_k. The method is applicable to Urysohn type equations.

[1] *Belozub V., Kozlova M., Lukianenko V.* Approximated solution algorithms for Urysohn-type equations // Journal of Physics: Conference Series, 2021.

[2] *Hua Y., Sarkar T.* Matrix Pencil Method for Estimating Parameters of Exponentially Damped/Undamped Sinusoids in Noise // IEEE Transactions on Acoustics, Speech and Signal Proc, 1990. Vol. 38(5). Pp. 814–824.

[3] *Larin V.* The use of matrix pencils in an identification problem // Journal of Automation and Information Sciences, 1996. Vol. 28(384). Pp. 53-62.

Интеллектуализация обработки информации социальных сетей

Германчук Мария Сергеевна[1] m.german4uk@yandex.ru
Козлова Маргарита Геннадьевна[1]⋆ art-inf@mail.ru
Руденко Людмила Ивановна[1] domlir@yandex.ru
[1]Симферополь, ФГАОУ ВО «Крымский федеральный университет имени В. И. Вернадского»

Социальные сети обладают огромным ресурсом неиспользуемой информации. Визуальная информация (изображения) сопровождается текстом и разными активностями (лайки, комментарии и пр.), отображающими динамику восприятия агентами-потребителями этой информации. Кластеризация потребителей, оценка системы влияния, управления влиянием для достижения коммерческих, политических и других целей являются актуальными задачами интеллектуализированного подхода обработки информации социальных сетей. Для решения этих задач необходимо иметь инструменты поиска, извлечения, обработки и сравнения объектов в неорганизованной коллекции изображений социальных сетей, а также процессов распространения. Данная работа является продолжением работ авторов [1]-[4].

Исследование процессов распространения информации в сложных сетях конкретизируется на распространении мемов в социальных сетях. Существенную роль играют структурные факторы сети, которые способствуют или препятствуют распространению мемов. Обходы кластеров сообществ социальных сетей (маршрутизация) определяют пути распространения мемов по сообществам. Субъекты сети, как правило, входят в тесно связанные группы, которые способствуют вирусному распространению близких к ним мемов. Таким образом, важными для управления и прогнозирования распространения мемов являются структурные параметры кластеризации, локальные меры близости к влиятельным источникам распространения. Выявление агентов, принадлежащих к активным сообществам, может служить признаком распространения мемов. Агент (вершина сети), степень которого наибольшая, может быть активным распространителем вирусной информации, но агент, являющийся членом пересекающихся кластеров (сообществ) имеет больше возможностей для распространения мемов. Одной из целей в сетях является предсказание эффективности процесса распространения, т.е. насколько быстро и насколько широко заражаются вершины сети (узлы-агенты). Обычно исследуется, какие узлы наиболее влияют на процесс распространения информации, что позволяет их контролировать (иммунизация, атака на эти узлы). Для выявления перекрывающихся кластеров (чередующихся, вложенных) [5] используется алгоритм линк-сообществ [6].

Способность узла (агента) к распространению информации соответствует множеству его взаимосвязей с кластерами (количество кластеров, с которыми он связан). Для обнаружения таких свойств используются меры центральности

узла или k-оболочки – индексы узлов, что позволяет исследовать сценарии распространения информации в социальных сетях. Важно, насколько затратными являются соответствующие алгоритмы.

В работе [7] на примере реальных и сгенерированных сетей показана важность перекрытия сообществ на процесс распространения, а также важность локальной информации о непосредственном соседстве узлов, которая отражает (соответствует) глобальную структуру сообществ сетей. Использование локальной информации является менее затратной.

Разработан прототип модульного комплекса ПО Memometrix с функционалом автоматического сбора мемов, их структурированного хранения, автоматического тегирования и визуализации. На вход интеллектуализированной системы Memometrix подаются запросы на мемы, отвечающие актуальным событиям, которые отражаются в заголовках новостных лент. Естественно считать, что журналисты (эксперты) следят за новостями, пишут тексты и в заголовках концентрируются наиболее характерные словосочетания (броские заголовки), которые и являются источником мемов. Для формирования запросов необходим онто-семантический анализ, в результате которого строится сеть из набора графов с частотными характеристиками (вершины) и силой связи (ребра). На основании запросов Memometrix выдает ранжированный набор мемов и соответствующие им показатели по социальным сетям. На основании матрицы показателей строятся интегральные показатели влияния интернет-мемов на русскоязычных пользователей сети Интернет, та также индексы $I(t), I^*(t), t = \overline{1,T}$, отвечающие моментам получения данных.

Данный подход позволяет отслеживать жизненный цикл конкретного мема (или группы мемов). По графикам индекса I^* (нормированы показатели по границам безопасности) можно: выяснять их безопасность; отсекать от рассмотрения мемы с незначительными значениями показателей (индексов) I, I^*; проводить кластеризацию мемов; проводить анализ запросов на выявление новых мемов и их распространение.

Работа поддержана грантом РФФИ No. 21-011-31733.

[1] *Gabrielyan O., Lukyanenko V., Kozlova M., Gasparyan M., Gabrielyan T.* Intellectualization Of The Sociometric Data Processing Of Internet Memes Within Virtual Communication Structure // The European Proceedings of Social & Behavioural Sciences, 2021. Vol. 102. Pp. 274–278.

[2] *Kozlova M., Lukianenko V., Germanchuk M.* Development of the toolkit to process the Internet memes meant for the modelling, analysis, monitoring and management of social processes // In book «Recognition and Perception of Images. Fundamentals and Applications», 2021. Pp. 189–220.

[3] *Germanchuk M., Kozlova M., Lukianenko V.* Some features of design of intelligent systems for processing the internet memes flow // CEUR Workshop Proceedings, 2021. Vol. 2834. Pp. 148–158.

[4] *Germanchuk M., Kozlova M., Lukianenko V.* Identification and Prediction of an Internet Meme Flow Lifecycle // 2021. Vol. 2914. Pp. 112–123.

[5] *Xie J., Kelley S., Szymansky B.* Overlapping community detectioon in networks: the state-of-the-art and comparative study // ACM Computing Surveys, 2011.

[6] *Ahn Y.-Y., Bagrow J. P., Lehmann S.* Link communities revel multiscale complexity in networks // Nature, 2010. Vol. 466(7303). Pp. 761–764.

[7] *Krukowski S., Hecking T.* Global and local community memberships for estimating spreading capability of nodes in social networks // Applied Network Science, 2021.

Intellectualization of social network information processing

Germanchuk Mariia[1] m.german4uk@yandex.ru
Kozlova Margarita[1]⋆ art-inf@mail.ru
Rudenko Ludmila[1] domlir@yandex.ru
[1]Simferopol, V. I. Vernadsky Crimean Federal University

Social networks have a huge resource of unused information. Visual information (images) is accompanied by text and various activities (likes, comments, etc.) that reflect the dynamics of perception by consumer agents of this information. Clustering of consumers, evaluation of the system of influence, influence management to achieve commercial, political and other goals are the actual tasks of the intellectualized approach to information processing of social networks. To solve these problems, it is necessary to have tools for searching, extracting, processing and comparing objects in an unorganized collection of images of social networks, as well as distribution processes. This work is a continuation of the authors' papers [1]-[4].

The study of the processes of information dissemination in complex networks is concretized on the dissemination of memes in social networks. A significant role is played by the structural factors of the network that contribute to or hinder the spread of memes. Traversals of social network community clusters (routing) determine the ways memes are distributed across communities. The subjects of the network, as a rule, belong to closely related groups that contribute to the viral spread of memes close to them. Thus, structural clustering parameters and local proximity measures to influential distribution sources are important for managing and predicting the spread of memes. The identification of agents belonging to active communities can serve as a sign of the spread of memes. The agent (the top of the network), whose degree is the greatest, can be an active distributor of viral information, but an agent who is a member of overlapping clusters (communities) has more opportunities for spreading memes. One of the goals in networks is to predict the effectiveness of the propagation process, i.e. how quickly and how widely the network nodes (agent nodes) are infected. It is usually investigated which nodes most influence the process of information dissemination, which allows them to be controlled (immunization, attack on these nodes). To identify overlapping clusters (alternating, nested) [5] uses the link community algorithm [6].

The ability of a node (agent) to disseminate information corresponds to the set of its relationships with clusters (the number of clusters with which it is connected). To detect such properties, measures of node centrality or k-shells are used – node indexes, which allows us to investigate scenarios for the dissemination of information in social networks. It is important how expensive the corresponding algorithms are.

At paper [7] on the example of real and generated networks, the importance of overlapping communities on the distribution process is shown, as well as the importance of local information about the immediate neighborhood of nodes, which

reflects (corresponds to) the global structure of network communities. Using local information is less costly.

A prototype of a modular Memometrix software package with the functionality of automatic collection of memes, their structured storage, automatic tagging and visualization has been developed. Requests for memes corresponding to current events, which are reflected in the headlines of news feeds, are submitted to the input of the intelligent Memometrix system. It is natural to assume that journalists (experts) follow the news, write texts and the most characteristic phrases (catchy headlines) are concentrated in the headlines, which are the source of memes. To form queries, an onto-semantic analysis is necessary, as a result of which a network is built from a set of graphs with frequency characteristics (vertices) and connection strength (edges). Based on requests, Memometrix provides a ranked set of memes and their corresponding indicators for social networks. Based on the matrix of indicators, integral indicators of the influence of Internet memes on Russian-speaking Internet users are constructed, as well as indices $I(t), I^*(t), t = \overline{1,T}$ corresponding to the moments of data acquisition.

This approach allows you to track the life cycle of a particular meme (or group of memes). According to the graphs of the index I^* (normalized indicators for security boundaries), you can: find out their security; cut off from consideration memes with insignificant values of indicators (indices) I, I^*; conduct clustering of memes; analyze requests for identifying new memes and their distribution.

This research is funded by RFBR, grant 21-011-31733.

[1] *Gabrielyan O., Lukyanenko V., Kozlova M., Gasparyan M., Gabrielyan T.* Intellectualization Of The Sociometric Data Processing Of Internet Memes Within Virtual Communication Structure // The European Proceedings of Social & Behavioural Sciences, 2021. Vol. 102. Pp. 274–278.

[2] *Kozlova M., Lukianenko V., Germanchuk M.* Development of the toolkit to process the Internet memes meant for the modelling, analysis, monitoring and management of social processes // In book ¡¡Recognition and Perception of Images. Fundamentals and Applications¿¿, 2021. Pp. 189–220.

[3] *Germanchuk M., Kozlova M., Lukianenko V.* Some features of design of intelligent systems for processing the internet memes flow // CEUR Workshop Proceedings, 2021. Vol. 2834. Pp. 148–158.

[4] *Germanchuk M., Kozlova M., Lukianenko V.* Identification and Prediction of an Internet Meme Flow Lifecycle // 2021. Vol. 2914. Pp. 112–123.

[5] *Xie J., Kelley S., Szymansky B.* Overlapping community detectioon in networks: the state-of-the-art and comparative study // ACM Computing Surveys, 2011.

[6] *Ahn Y.-Y., Bagrow J. P., Lehmann S.* Link communities revel multiscale complexity in networks // Nature, 2010. Vol. 466(7303). Pp. 761–764.

[7] *Krukowski S., Hecking T.* Global and local community memberships for estimating spreading capability of nodes in social networks // Applied Network Science, 2021.

Радиус устойчивости и робастное расписание на примере задачи железнодорожного планирования

Гришин Егор Максимович[1] grishin.em16@physics.msu.ru

[1]Москва, Институт Проблем Управления РАН

В настоящее время во многих исследовательские работы по теории расписаний и дискретной оптимизации посвящены работе с детерминированными входными данными. Даже в прикладных задачах, в которых могут происходить сбои, возникать ошибки и неточности, предполагается, что входные данные известны заранее и не меняются со временем.

В задачах железнодорожного планирования могут происходить нештатные ситуации (например, упавшее дерево на один из путей), при которых необходимо производить перепланирование с минимизацией отклонений от базового расписания. Это обусловлено как высокой нагрузкой железнодорожной сети, так и обязательствами компаний по доставке грузов в срок с уплатой штрафов за задержки.

Для создания расписания, которое будет оставаться оптимальным (субоптимальным) при некоторых вариациях входных данных, разработаны методы робастного планирования [1]. В том числе существуют алгоритмы для решения различных практических железнодорожных задач, представленные в [2] и [3].

Кроме того, для оценки допустимых вариаций входных параметров (по отдельности или в совокупности) были предложены методы оценки радиуса устойчивости [4] – минимаксного значения допустимых интервалов, на которых лежат входные данные, при которых полученное расписание (последовательность операций) остается оптимальным. При этом значение целевой функции и времена начала (окончания) операций могут изменяться.

В данной работе предложен метод построения робастного расписания и оценки радиуса устойчивости для одной практической задачи железнодорожного планирования. Предложенная задача представима в виде графовой модели так как определение радиуса устойчивости наиболее часто рассматривается для графовых задач теории расписаний [5].

Работа частично поддержана грантом РФФИ, НТУ «Сириус», ОАО «РЖД» и Образовательного Фонда «Талант и успех» No. 20-38-51010.

[1] *Kouvelis P., Yu G.* Robust Scheduling Problems. In: Robust Discrete Optimization and Its Applications. // Nonconvex Optimization and Its Applications, 1997. Pp. 241–289.

[2] *Ran C., Leishan Z., Yixiang Y., Jinjin T., Chao L.* The integrated optimization of robust train timetabling and electric multiple unit circulation and maintenance scheduling problem. // Advances in Mechanical Engineering, 2018. Pp. 89–103.

[3] *Maróti G.* A branch-and-bound approach for robust railway timetabling. // Public Transp, 2017. Vol. 9. Pp. 73–94.

[4] *Sotskov Y., Tanaev V., Werner F.* Stability Radius of an Optimal Schedule: A Survey and Recent Developments. // Industrial Applications of Combinatorial Optimization. Applied Optimization, Boston: Springer, 1998. Pp. 72–108.

[5] *Braunj O., La T., Schmidt G., Sotskov Y.* Stability of Johnson's schedule with respect to limited machine availability. // Int. J. Prod. Res, 2002. Vol. 40. Pp. 4381–4400.

Stability radius and robust scheduling in a rail planning problem as an example

Grishin Egor[1] grishin.em16@physics.msu.ru

[1]Moscow, Institute of Control Science RAS

At present, in many research papers on scheduling theory and discrete optimisation, the topic is dealing with determined input data. Even in applicable problems in which failures, errors, and inaccuracies may occur, it is assumed that the input data is known in advance and does not change over time.

In railway planning problems, unexpected situations can happen (e.g. a fallen tree on one of the tracks), in which it is necessary to carry out rescheduling with minimised deviations from the baseline timetable. This is caused both by the high load on the railway network and by the companies' obligations to deliver goods on time and pay penalties for delays.

In order to create a schedule that will remain optimal (suboptimal) under some variations of the input data, robust scheduling methods have been developed [1]. Including algorithms for solving various practical railway problems, presented in [2] and [3].

In addition, in order to estimate acceptable variations of input parameters (individually or in complex), there were proposed methods to estimate the stability radius [4] – the minimax value of possible intervals, in which the input data lie, at which the obtained schedule (sequence of operations) remains optimal. At the same time, the value of the objective function and start (completion) times of operations may vary.

In this paper we propose the method of robust scheduling and stability radius estimation for one practical problem of railroad planning. The proposed problem is represented in the form of graph model since the definition of stability radius is most often considered for graph problems of scheduling theory [5].

This work was partly supported by a grant from the RFBR, NTU Sirius, Russian Railways and the Talent and Success Educational Foundation No. 20-38-51010.

[1] *Kouvelis P., Yu G.* Robust Scheduling Problems. In: Robust Discrete Optimization and Its Applications. // Nonconvex Optimization and Its Applications, 1997. Pp. 241–289.

[2] *Ran C., Leishan Z., Yixiang Y., Jinjin T., Chao L.* The integrated optimization of robust train timetabling and electric multiple unit circulation and maintenance scheduling problem. // Advances in Mechanical Engineering, 2018. Pp. 89–103.

[3] *Maróti G.* A branch-and-bound approach for robust railway timetabling. // Public Transp, 2017. Vol. 9. Pp. 73–94.

[4] *Sotskov Y., Tanaev V., Werner F.* Stability Radius of an Optimal Schedule: A Survey and Recent Developments. // Industrial Applications of Combinatorial Optimization. Applied Optimization, Boston: Springer, 1998. Pp. 72–108.

[5] *Braunj O., La T., Schmidt G., Sotskov Y.* Stability of Johnson's schedule with respect to limited machine availability. // Int. J. Prod. Res, 2002. Vol. 40. Pp. 4381–4400.

Специфика задач маршрутизации в условиях локальных преобразований сети

Лукьяненко Владимир Андреевич[1]⋆ art-inf@yandex.ru
Германчук Мария Сергеевна[1] m.german4uk@yandex.ru
Макаров Олег Олегович[1] fantom2.00@mail.ru
[1]Симферополь, ФГАОУ ВО «Крымский федеральный университет имени В. И. Вернадского»

Рассматриваются задачи выбора наилучших маршрутов для многих агентов-коммивояжеров в случае структурных преобразований сложной инфраструктурной сети. Показано, что многоагентный подход в сочетании с упрощением структуры сети и кластеризацией сети позволяют получить предварительный набор маршрутов многих коммивояжеров, реоптимизация которых позволит найти новые маршруты в случае изменения структуры сети. Многоагентные задачи маршрутизации (Multiple Traveling Salesman Problems – mTSP) на сложных сетях существенно обобщают классическую задачу коммивояжера на конечных графах. Такие задачи являются важной частью прикладной теории графов. Разработка алгоритмов их решения связана с учетом сложности и специфики структуры сети, наличием ограничений, априорной информации, прецедентности. Задача усложняется, если сеть меняется с течением времени, либо необходимо синтезировать фрагменты сети [4]. Условия устойчивости решений по отношению к малым изменениям матрицы расстояний носят, в основном, теоретический характер. С практической точки зрения интересны случаи реоптимизации, т. е. когда при удалении (или добавлении) вершины или дуги новое решение может быть построено по предыдущему с помощью полиномиального алгоритма.

Многоагентный подход предполагает сочетание различных постановок задач, методов и алгоритмов, основанных на кластеризации; псевдобулевой оптимизации большой размерности с ограничениями в виде ДНФ условий; применением композиций метаэвристик [2, 3]. Многоагентные задачи маршрутизации на сложных сетях в условиях изменяющейся сети являются частью моделирования системы управления. Здесь сочетаются задачи выбора решения для лица, принимающего решения; распределения ресурсов; синтеза сети; задачи анализа потери устойчивости сети в зависимости от удаления вершин, дуг или части маршрута одного или нескольких агентов-коммивояжеров; кластеризации; обмена информацией между агентами; потоковые задачи; транспортные задачи; задачи прокладки кратчайших путей и замкнутых маршрутов с обязательным посещением выделенного множества вершин.

Выбор алгоритмов поиска оптимальных маршрутов основан на преобразовании сети и учет всей имеющейся информации. Задается исходная сеть $S=(G,C)$, где $G=(V,U)$, G – граф, V – множество вершин, U – множество дуг, C – весовая матрица, определенная на дугах $(i,j)\in U$, $i,j\in V$ с элементами $c_{ij}\geqslant 0$.

В преобразованиях сети $\pi_i : S \to S_i,\ i \in I$ используются: метрические характеристики сети; структурные составляющие сети: «висячие» вершины, мосты, сочленения, очевидные кластеры; априорная информация, запреты, предписания, прецеденты; кластеризация сети с помощью различных методов, согласованных с поиском маршрутов коммивояжеров: k-means, c-средних, max-cut, результату решения задачи и назначениях и др.; сопоставление более простой сети, удовлетворяющей свойствам, которые допускают полиномиальные алгоритмы нахождения наилучших маршрутов; учет количества агентов-коммивояжеров; информация о возможном изменении сети в результате ЧС и т.п.

В соответствии с такими преобразованиями формируется база знаний и набор вариантов сетей $S_i, i \in I$ для которых формулируются адекватные постановки задач и набор алгоритмов решения, предназначенных для исполнения МАС mTSP.

В качестве примера рассмотрим инфраструктурную сеть дорог и объектов водообеспечения в случае чрезвычайной ситуации (ЧС) [1]. Объекты, расположенные на карте ГИС, представлены в виде информации на разных слоях карты. Многослойность картографических данных позволяет получать более целостные и полезные, с практической точки зрения, результаты. В данном случае используется инфраструктура и координаты объектов водоснабжения Большой Ялты (гидранты, колодцы и т. п.). В случае ЧС остро стоит вопрос быстрого доступа к объектам водоснабжения. Моделирование ЧС предполагает изменение маршрутов в случае выхода из строя ряда водных объектов.

Выделяются два слоя. В наземном слое требуется в расчет брать существующие дороги и подъезды к объекту. Воздушный слой, конечно, предоставляет больше возможностей и формирует полный неориентированный граф между точками. Решение mTSP на такой сети затем проецируется на реальную сеть S. Для упрощения визуализции отображения построенных маршрутов используется проекция реальных точек с дорогами на преобразованную сеть $S_i, i \in I$. Здесь каждое ребро будет иметь вес, соответствующий кратчайшему расстоянию между объектами.

В критической ситуации требуется максимально быстрый расчет по предварительным, начальным, приближенным решениям задачи mTSP, основываясь на котором производятся уточнение, исключение и добавление объектов. Реализован подход, согласно которому исходной сети ставится сеть с графом облета сети (проекция на плоскость), в котором каждое ребро имеет вес, соответствующий кратчайшему расстоянию между объектами.

Получено решение задачи mTSP для Ялты с прилегающими территориями в случае нескольких агентов-коммивояжеров. Грубое решение на кластерах коммивояжеров улучшается с помощью локальных алгоритмов (2-Opt, алгоритм имитации отжига). Многоагентность данных позволяет планировать реализацию функционала взаимодействия в МАС с уже полученными данными для имитации режимов ЧС.

[1] *Артюхин В. В.* Прогнозирование чрезвычайных ситуаций с помощью дискретной оптимизации и современных программных средств // Технологии гражданской безопасности, 2014. Т. 11(1). С. 86–91.

[2] *Germanchuk M., Kozlova M., Lukianenko V.* Pseudo-Boolean Conditional Optimization Models for a Class of Multiple Traveling Salesmen Problems // Automation and Remote Control, 2021. Vol. 82(10). Pp. 1651–1667.

[3] *Germanchuk M., Lemtyuzhnikova D., Lukianenko V.* Metaheuristic Algorithms for Multiagent Routing Problems // Automation and Remote Control, 2021. Vol. 82(10). Pp. 1787–1801.

[4] *Овчинников В. А.* Графы и задачи анализа и синтеза структур сложных систем // М.: Изд-во МГТУ им. Н. Э. Баумана, 2014. 423 с.

The specifics of routing tasks in the context of local network transformations

Lukianenko Vladimir[1]⋆ art-inf@yandex.ru
Germanchuk Mariia[1] m.german4uk@yandex.ru
Makarov Oleg[1] fantom2.00@mail.ru
[1]Simferopol, V. I. Vernadsky Crimean Federal University

The problems of choosing the best routes for many traveling sales agents in the case of structural transformations of a complex infrastructure network are considered. It is shown that the multi-agent approach combined with the simplification of the network structure and clustering of the network make it possible to obtain a preliminary set of routes for many traveling salesmen, the reoptimization of which will allow finding new routes in the event of a change in the network structure. Multi-agent routing problems (Multiple Traveling Salesman Problems – mTSP) on complex networks significantly generalize the classical traveling salesman problem on finite graphs. Such problems are an important part of applied graph theory. The development of algorithms for their solution is related to the complexity and specificity of the network structure, the presence of restrictions, a priori information, precedent. The task becomes more complicated if the network changes over time, or it is necessary to synthesize fragments of the network [4]. The conditions for the stability of solutions with respect to small changes in the distance matrix are mainly theoretical in nature. From a practical point of view, cases of reoptimization are interesting, i.e. when removing (or adding) a vertex or an arc, a new solution can be constructed from the previous one using a polynomial algorithm.

The multi-agent approach involves a combination of various problem statements, methods and algorithms based on clustering; pseudo-boolean optimization of large dimension with constraints in the form of DNF conditions; the use of compositions of metaheuristics [2, 3]. Multi-agent routing tasks on complex networks in a changing network are part of the control system modeling. It combines the tasks of choosing a solution for the decision-maker; resource allocation; network synthesis; tasks of analyzing the loss of network stability depending on the removal of vertices, arcs, or part of the route of one or more traveling agents; clustering; information exchange between agents; streaming tasks; transport tasks; tasks of laying shortest paths and closed routes with mandatory visits to the selected set of vertices.

The choice of algorithms for finding optimal routes is based on the transformation of the network and taking into account all available information. The source network is set to $S=(G,C)$, where $G=(V,U)$, G – graph, V – set of vertices, U – set of arcs, C – weight matrix defined on arcs $(i,j)\in U$, $i,j\in V$ with elements $c_{ij}\geqslant 0$.

In network transformations $\pi_i : \S \to S_i,\ i\in I$ uses: metric characteristics of the network; structural components of the network: "hanging" vertices, bridges, joints, obvious clusters; a priori information, prohibitions, prescriptions, precedents; clustering of the network using various methods consistent with the search for traveling

salesmen routes: k-means, c-means, max-cut, the result of solving the problem and assignments, etc.; comparison of a simpler network satisfying properties that allow polynomial algorithms for finding the best routes; accounting for the number of traveling sales agents; information about possible network changes as a result of emergencies, etc.

In accordance with such transformations, a knowledge base and a set of network options are formed $S_i, i \in I$ for which adequate problem statements and a set of solution algorithms designed for the execution of MAC mTSP are formulated.

As an example, consider the infrastructure network of roads and water supply facilities in the event of an emergency (emergency) [1]. The objects located on the GIS map are presented as information on different layers of the map. The layering of cartographic data makes it possible to obtain more holistic and useful, from a practical point of view, results. In this case, the infrastructure and coordinates of the water supply facilities of the Big Yalta (hydrants, wells, etc.) are used. In case of an emergency, the issue of quick access to water supply facilities is acute. Emergency modeling involves changing routes in case of failure of a number of water bodies.

Two layers are highlighted. In the ground layer, it is required to take into account existing roads and entrances to the object. The air layer, of course, provides more possibilities and forms a complete undirected graph between points. The mTSP solution on such a network is then projected onto the real network S. To simplify the visualization of the display of the constructed routes, the projection of real points with roads on the transformed network $S_i, \ddot{u}sused \in I$. Here, each edge will have a weight corresponding to the shortest distance between objects.

In a critical situation, the fastest possible calculation is required based on preliminary, initial, approximate solutions of the mTSP problem, based on which refinement, exclusion and addition of objects are performed. An approach is implemented according to which a network with a network flyover graph (projection onto a plane) is placed in the initial network, in which each edge has a weight corresponding to the shortest distance between objects.

The solution of the mTSP problem for Yalta with adjacent territories in the case of several traveling sales agents is obtained. The rough solution on traveling salesman clusters is improved using local algorithms (2-Opt, simulated annealing algorithm). Multi-agent data allows you to plan the implementation of the interaction functionality in the MAC with already received data to simulate emergency modes.

[1] *Artyukhin V.* Forecasting of emergency situations with the help of discrete optimization and modern software tools // Civil security technologies, 2014. Vol. 11(1). Pp. 86–91.

[2] *Germanchuk M., Kozlova M., Lukianenko V.* Pseudo-Boolean Conditional Optimization Models for a Class of Multiple Traveling Salesmen Problems // Automation and Remote Control, 2021. Vol. 82(10). Pp. 1651–1667.

[3] *Germanchuk M., Lemtyuzhnikova D., Lukianenko V.* Metaheuristic Algorithms for Multiagent Routing Problems // Automation and Remote Control, 2021. Vol. 82(10). Pp. 1787–1801.

[4] *Ovchinnikov V.* Graphs and problems of analysis and synthesis of structures of complex systems // Moscow: Publishing House of the Bauman Moscow State Technical University, 2014. 423 p.

Метрический подход для задач на быстродействие

Букуева Елена Сергеевна[1]⋆ alena-bukueva@mail.ru
Кудинов Илья Дмитриевич[2,3] ilja@kdsli.ru
Лемтюжникова Дарья Владимировна[3] darabbt@gmail.com
[1]Москва, Московский государственный университет
[2]Москва, Московский авиационный институт
[3]Москва, Институт проблем управления им. В. А. Трапезникова

Одним из подходов, используемых при рассмотрения NP-трудных задач теории расписаний, является метрический подход. Он основан на введении метрик, с помощью которых по исходному примеру, не являющемуся полиномиально разрешимым, проекцированием на известную полиномиално разрешимую область, получается оптимальное или приближённое решение.

Рассматривается проблема минимизации задачи на быстродействие на двух параллельных приборах с использованием известных алгоритмов со сложностью $O(n^2)$ и $O(e + n\alpha(n))$ для получения полиномиальных решений при рассмотрении работ, выполняемых за равные промежутки времени с заданной последовательностью выполнения: $P2|prec,\ p_j = p|C_{max}$.

Для получения оценки абсолютной погрешности и нахождении приближённого решения для задач теории расписаний проводится эксперимент.

Работа поддержана грантом РФФИ No. 20-58-S52006.

[1] *Lazarev A., Lemtyuzhnikova D., Werner F.* A metric approach for scheduling problems with minimizing the maximum penalty // Oxford(Great Britain): Applied Mathematical Modelling, 2021. Vol. 89. Pp. 1163–1176.

Metric approach for minimization makespan problem

Bukueva Elena[1]⋆ alena-bukueva@mail.ru
Kudinov Ilya[2,3] ilja@kdsli.ru
Lemtyuzhnikova Daria[3] darabbt@gmail.com
[1]Moscow, Lomonosov Moscow State University
[2]Moscow, Moscow Aviation Institute National State University
[3]Moscow, V. A. Trapeznikov Institute of Control Sciences of RAS

One of the approaches used in considering NP-hard problems in the theory of schedules is the metric approach. It is based on the introduction of metrics, with the help of which, according to the original example, which is not polynomial solvable, projection onto a known polynomial solvable area, an optimal or approximate solution is obtained.

The problem of minimizing the performance problem on two parallel processors using known algorithms with complexity $O(n^2)$ and $O(e + n\alpha(n))$ is considered to obtain polynomial solutions when tasks are performed at equal intervals with a given sequence of execution: $P2|prec, p_j = p|C_{max}$.

To obtain an estimate of the absolute error and find an approximate solution for the problem of the theory of schedules, an experiment is carried out.

This research is funded by RFBR, grant 20-58-S52006.

[1] *Lazarev A., Lemtyuzhnikova D., Werner F.* A metric approach for scheduling problems with minimizing the maximum penalty // Oxford(Great Britain): Applied Mathematical Modelling, 2021. Vol. 89. Pp. 1163–1176.

Планирование операций с участием анестезиолога в операционных комнатах

Лазарев Александр Алексеевич[1] jobmath@mail.ru
Лемтюжникова Дарья Владимировна[1,2] darabbt@gmail.com
Сомов Михаил Львович[1]⋆ somovml1999@gmail.com

[1]Москва, Институт проблем управления им. В. А. Трапезникова Российской академии наук
[2]Москва, Московский авиационный институт

Разработка моделей и алгоритмов составления расписания приобрела особую актуальность в связи с пандемией Ковид-19. В частности, существует высокая потребность в планировании работы операционных залов. В данной работе рассматривается проблема составления расписания в операционные комнаты с учетом работы анестезиологов. Сформулируем данную проблему в терминах теории расписаний. Пусть даны операционные комнаты, каждая из которых имеет свой набор операций назначенных в эту операционную, а также набор анестезиологов на все операционные. Каждая операция состоит из двух частей, анестезии и хирургической операции. Даны длительности этих частей. Для выполнения всех операций по анестезии во всех операционных имеется определенное количество анестезиологов. Другими словами, в любой момент времени не может одновременно проводиться операций анестезии больше, чем анестезиологов. Задача заключается в составлении такого расписания операций, чтобы минимизировать время окончание последней операции, т.е. все назначенные операции должны завершится за наименьшее время.

Для получение оптимального решения была предложена модель целочисленного линейного программирования, которая решалась с помощью Gurobi. Данная проблема является NP-трудной и нахождение оптимального решения вычислительно очень сложная задача. В связи с этим были разработаны два эвристических алгоритма. Эксперименты были проведены на псевдо-реальных данных, которые были сгенерированны на основе данных, предоставленных Центром нейрохирургии им. Н.Н. Бурденко.

Работа поддержана грантом РФФИ No. 20-58-S52006.

Scheduling of Anaesthesia Operations in Operating Rooms

Lazarev Alexander[1] jobmath@mail.ru
Lemtyuzhnikova Darya[1,2] darabbt@gmail.com
Somov Mikhail[2]⋆ somovml1999@gmail.com
[1]Moscow, Institute of Control Sciences
[2]Moscow, Moscow Aviation Institute

Health scheduling is an essential component for the medicine automation. The development of scheduling models and algorithms has gained particular relevance in connection with the Covid-19 pandemic. In particular, there is a high demand for scheduling operating rooms. This paper addresses the problem of scheduling operating rooms with anesthesiologists in mind. Let us formulate this problem in terms of the theory of schedules. The problem is formerly defined as follows. Let there be operating rooms, each of which has a different set of surgeries assigned to that operating room, as well as a set of anesthesiologists for all operating rooms. Each operation consists of two parts, anaesthesia operation and surgical operation. The duration of these parts are given. There are a certain number of anesthesiologists in all operating rooms to perform all anesthesia operations. In other words, there cannot be more anesthesia operations than there are anesthesiologists at any one time. The problem is to determine a feasible anaesthesia schedule that has a minimum makespan, i.e., all operations are finished in the shortest time.

To obtain an optimal solution, we proposed an integer linear programming model, which was solved using Gurobi. This problem is NP-hard and finding the optimal solution is computationally very difficult. In this regard, two heuristic algorithms were developed. The experiments were performed on pseudo-real data, which were generated from data provided by The Burdenko Neurosurgical Center.

This research is funded by RFBR, grant No. 20-58-S52006.

Сравнение булевой и целочисленной моделей для задачи пункта перевалки морской порт – железная дорога

Морозов Николай Юрьевич[1]⋆ morozov.nikolay@physics.msu.ru
Гришин Егор Максимович[1] grishin.em16@physics.msu.ru
Коровкин Дмитрий Михайлович[2] korovkin.dm19@physics.msu.ru
[1]Москва, Институт Проблем Управления им. В. А. Трапезникова РАН
[2]Москва, Московский Государственный Университет им. М. В. Ломоносова

Работа посвящена комплексной задаче, возникающей в морском порту при транспортировке грузов на кораблях. В первую очередь необходимо для каждого корабля назначить причал, на котором он должен быть разгружен. Не все причалы идентичны, и разгрузчики на них могут разгружать только определенные типы грузов. Данная задача в научной литературе называется задачей распределения причалов (berth allocation problem, BAP). В [1] представлены различные постановки данной задачи. Кроме того, данная задача является NP-трудной [2].

После разгрузки грузов с кораблей, прибывающих в разное время, необходимо сформировать составы. Для каждого груза заданы пункт назначения, его объем (выраженный в двадцатифутовом эквиваленте, TEU) и важность. Целевой функцией является минимизация взвешенных затрат на доставку всех прибывающих грузов до пунктов их назначения. Это задача чрезвычайно важна для ОАО "РЖД"так как объемы грузоперевозок растут ежегодно [3].

Были составлены две математические модели: булева и целочисленная [4]. У каждой из них есть свои плюсы и недостатки. Так, например, в булевой модели создавалось большое количество переменных и вводились дополнительные ограничения на них. В целочисленной же модели количество переменных (и ограничений) значительно меньше, но при этом сами переменные могут принимать больше двух значений. Был проведен сравнительный анализ эффективности представленных моделей на псевдореальных данных. Был написан генератор входных данных, соответствующих портам Дальнего Востока России. Размерность примеров, сгенерированных в целях тестирования предлагаемых подходов, аналогична рассматриваемым в научной литературе для задач данного типа [5].

Работа частично поддержана грантом РФФИ, НТУ «Сириус», ОАО «РЖД» и Образовательного Фонда «Талант и успех» No. 20-38-51010.

[1] *Eskandari H., Amin M., Hassan E., Memarpour M., Ashkan S.* Evaluation of different berthing scenarios in Shahid Rajaee container terminal using discrete-event simulation // Simulation Conference (WSC), 2013.

[2] *Sammarra M., Monaco M.* The Berth Allocation Problem: A Strong Formulation Solved by a Lagrangean Approach // TTransportation Science, 2007. Vol. 41(2). Pp. 265–280.

[3] *РЖД.* Грузовые перевозки // URL: https://company.rzd.ru/ru/9377

[4] *Wittmann D., Krumsiek J., Saez-Rodriguez J. et al.* Transforming Boolean models to continuous models: methodology and application to T-cell receptor signaling // TTransportation Science. London: BMC Syst Biol, 2003.

[5] *Buhrkal K., Zuglian S., Ropke S., Larsen J., Lusby R.* Models for the discrete berth allocation problem: a computational comparison // Transp. Res. Part E: Log. and Transp. Rev: Amsterdam: Elsevier, 2011. Vol. 47. Pp. 461–473.

Comparison of boolean and integer mathematical models for the seaport-railway transhipment point problem

Morozov Niloay[1]⋆ morozov.nikolay@physics.msu.ru
Grishin Egor[1] grishin.em16@physics.msu.ru
Korovkin Dmitry[2] korovkin.dm19@physics.msu.ru
[1]Moscow, Institute of Control Science RAS
[2]Moscow, Moscow State University

The work deals with the complex problem arising in a seaport for transporting cargo on ships. First of all, it is necessary to designate for each ship a berth where it is to be landed. Not all berths are identical, and unloaders on them can unload only certain types of cargo. This problem is called the berth allocation problem (BAP) in scientific literature. The [1] presents various formulations of this problem. In addition, this problem is NP-hard [2].

After unloading cargo from ships arriving at different times, trains need to be formed. For each cargo, the destination, its volume (expressed in twenty-foot equivalent units, TEU) and importance are specified. The objective function is to minimise the weighted cost of delivering all arriving cargo to its destination. This task is extremely important for Russian Railways, as freight volumes are increasing annually [3].

Two mathematical models have been created: a Boolean model and an integer model [4]. Each has advantages and disadvantages. For example, in the Boolean model a large number of variables were created and additional restrictions were imposed on them. In the integer model, the number of variables (and constraints) is much smaller, but the variables themselves can take more than two values. A comparative analysis of the effectiveness of the presented models on pseudo-real data has been carried out. An input data generator corresponding to the ports of the Russian Far East was written. The dimensions of the examples generated to test the proposed approaches is similar to those considered in the scientific literature for problems of this type [5].

This work was partly supported by a grant from the RFBR, NTU Sirius, Russian Railways and the Talent and Success Educational Foundation No. 20-38-51010.

[1] *Eskandari H., Amin M., Hassan E., Memarpour M., Ashkan S.* Evaluation of different berthing scenarios in Shahid Rajaee container terminal using discrete-event simulation // Simulation Conference (WSC), 2013.

[2] *Sammarra M., Monaco M.* The Berth Allocation Problem: A Strong Formulation Solved by a Lagrangean Approach // TTransportation Science, 2007. Vol. 41(2). Pp. 265–280.

[3] *RZD* Freight transport // URL: https://company.rzd.ru/ru/9377

[4] *Wittmann D., Krumsiek J., Saez-Rodriguez J. et al.* Transforming Boolean models to continuous models: methodology and application to T-cell receptor signaling // TTransportation Science. London: BMC Syst Biol, 2003.

[5] *Buhrkal K., Zuglian S., Ropke S., Larsen J., Lusby R.* Models for the discrete berth allocation problem: a computational comparison // Transp. Res. Part E: Log. and Transp. Rev: Amsterdam: Elsevier, 2011. Vol. 47. Pp. 461–473.

О свойствах решения и целевой функции в теории расписаний

Галахов Семен Алексеевич[1] galakhov.sa16@physics.msu.ru
[1] Москва, ИПУ РАН

Одно из предположений классической теории расписаний гласит о том, что времена поступления всех требований к исполнительным машинам определены до начала процесса построения расписания и не могут быть изменены во время его реализации. Такое предположение необходимо для создания алгоритмов, однако делает их менее значимыми с практической точки зрения. Таким образом возникает задача анализа поведения целевой функции при измененииии начальных параметров [1].

В данной работе рассматривается двухприборная задача с одинаковыми временами обслуживания, известными временами поступления и отсутствующими дедлайнами. В каждый момент времени каждый прибор может обслуживать не более одного требования, каждое требование может быть обслужено не более чем на одном приборе . Прерывания в обслуживании запрещены. Целевая функция – суммарное время ожидания.

Для описанной задачи введен некий аналог непрерывности и проанализировано поведение целевой функции при изменении входных параметров задачи. Изучена проблема устойчивости полученного расписания.

Работа поддержана грантом РФФИ No. 20-58-S52006.

[1] *Сотсков Ю. Н., Сотскова Н. Ю.* Теория расписаний. Системы с неопределенными числовыми параметрами // Мн: ОИПИ НАН Беларуси, 2004. 290 с.

On solution properties and the objective function in scheduling theory

Galkhov Semen[1,2] galakhov.sa16@physics.msu.ru
[1]Moscow, ICS RAS

One of the assumptions of classical scheduling theory is that the arrival times of all the jobs on the machines are determined before the scheduling process begins and cannot be changed during its implementation. This assumption is necessary for creating algorithms, but makes them less meaningful for a practice. Thus the problem arises of analysing the behaviour of the objective function as the initial parameters change [1].

This paper considers a two-device problem with equal processing times, known arrival times and deadlines equal to infinity. At any given time, each machine can serve no more than one demand from the set of all demands, each demand can be served on no more than one machine from the set of all machines. Maintenance interruptions are prohibited. The objective function is the total waiting time.

For the described problem the behaviour, a certain analog of continuity, of the target function when the input parameters of the problem change is analysed. The problem of stability of the obtained schedule is studied.

This research is funded by RFBR, grant 20-58-S52006.

[1] *Sotskov Y., Sotskova N.* Schedule theory. Systems with uncertain numerical parameters // Minsk: OIPI of the National Academy of Sciences of Belarus, 2004. 290 p.

Аппроксимация весов задачи $1||\sum w_jC_j$

Барашов Егор Борисович[1]⋆ barashov.eb@gmail.com
Лемтюжникова Дарья Владимировна[1] darabbt@gmail.com
Тюняткин Андрей Александрович[1] andtun@yandex.ru
[1]Москва, ИПУ РАН

Исследуется задача $1||\sum w_jC_j$: имеются один прибор и множество $J = \{1, 2, ..., n\}$ из n требований, которые необходимо обслужить на приборе. Для каждого требования $j \in J$ заданы длительности обслуживания p_j. Отношения предшествования отсутствуют (не накладывается ограничений на очередность обслуживания требований), прерывания в обслуживании требований и искусственные простои прибора запрещены. Расписание π однозначно задаётся порядком обслуживания требований $(j_1, \ldots, j_n)$. В задаче $1||\sum w_jC_j$ необходимо найти расписание π^0, минимизирующее суммарное взвешенное время завершения обслуживания требований $\sum w_jC_j$, где C_j - момент окончания обслуживания требования j, а $w_j > 0$ весовой коэффициент соответствующего времени завершения обслуживания.

Выберем произвольным образом пару различных требований $i, j \in (1, ..., n), i \neq j$. Разобьем множество K на два подмножества $K_{i,j}$ и $K_{j,i}$. Идея алгоритма – в решении эффективной системы неравенств и интерполяции получаемых граничных значений.

Индекс l выбирается произвольным образом, для аппроксимации коэффициентов w_j необходимо:

1. построить множества $K_{i,j}, K_{j,i}$;
2. заполнить матрицы $X(i,j) = max_{k \in K_{j,i}}(\frac{p_j^k}{p_i^k}), Y(i,j) = max_{k \in K_{i,j}}(\frac{p_j^k}{p_i^k})$;
3. вычислить $w_j = \begin{cases} 1, \ j = l; \\ \sum_{l=0}^{n} \frac{\prod_{i \neq l}(\frac{n}{2} - i)}{\prod_{m \neq l}(\frac{n}{2} - j)} \xi(l,j), \ j \neq l, \end{cases}$, где индекс l любое число.

Формула в пункте 3 получена из интерполяционного полинома Лагранжа[1].

$$L_m(x) = \sum_{k=0}^{m} \frac{\prod_{i \neq k}(x - x_i)}{\prod_{j \neq k}(x - x_j)} f(x_k), \tag{1}$$

а как ξ обозначалась следующая величина: $\xi(l,j) = X(l,j)$ при $l <= n$ и $\xi(l,j) = Y(l-n,j)$ при $l > n$. Таким образом, по всем вычисленным значениям границ строится интерполяционный полином, а затем находится его значение при $n/2$ - то есть оценивается промежуточное значение между $X(n,j)$ и $Y(1,j)$.

Результат работы алгоритма – набор таких весовых коэффициентов w_j, $j = \overline{1,n}$, что для каждого из N заданных примеров оптимальное расписание, най-

денное для аппроксимированных значений весовых коэффициентов либо полностью совпадает с его заданным оптимальным расписанием, соответствующим неизвестному истинному набору весовых коэффициентов w_j^0, либо имеет с ним одинаковое значение целевой функции.

Работа выполнена при частичной финансовой поддержке Российского фонда фундаментальных исследований No. 20-58-S52006.

[1] *Лазарев А. А., Лемтюжникова Д. В., Тюняткин А. А.* Метрическая интерполяция для задачи минимизации максимального временно́го смещения для одного прибора / Автоматика и телемеханика, 2021. No. 10. С. 93–109.

Approximation of the weights of the problem $1||\sum w_jC_j$

Barashov Egor[1]⋆ barashov.eb@gmail.com
Lemtiuzhnikova Daria[1] darabbt@gmail.com
Tyunyatkin Andrey[1] andtun@yandex.ru
[1]Moscow, V. A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences

The problem $1||\sum w_jC_j$ is under research: there is one machine and a set of jobs $J = \{1, 2, \ldots, n\}$ that have to be processed on the machine. For each job $j \in J$ processing time p_j are known. There are no precedence relations, and interrupts in the processing of jobs are prohibited. The schedule π is the order in which jobs are processed $(j_1, ..., j_n)$. In the problem $1|r_j|\sum w_jC_j$ we need to construct a schedule π^0 that minimizes the total weighted completion time $\sum w_jC_j$, where C_j is the job j completion time, and $w_j > 0$ is the weight coefficient of the corresponding completion time.

A pair of different requirements is randomly chosen $i, j \in (1, ..., n), i \neq j$. We divide the set K into two subsets $K_{i,j}$ and $K_{j,i}$. The idea of the algorithm is to solve an effective inequality system[1] and to interpolate the bound values.

Index l is chosen randomly, then to approximate w_j:

1. compute the sets $K_{i,j}, K_{j,i}$;
2. fill the matrices $X(i,j) = max_{k \in K_{j,i}}(\frac{p_j^k}{p_i^k}), Y(i,j) = max_{k \in K_{i,j}}(\frac{p_j^k}{p_i^k})$;
3. calculate $w_j = \begin{cases} 1, \ j = l; \\ \sum_{l=0}^{n} \frac{\prod_{i \neq l}(\frac{n}{2} - i)}{\prod_{m \neq l}(\frac{n}{2} - j)} \xi(l,j), \ j \neq l, \end{cases}$ where the index l is any number.

The formula in step 3 is derived from the Lagrange interpolation polynomial.

$$L_m(x) = \sum_{k=0}^{m} \frac{\prod_{i \neq k}(x - x_i)}{\prod_{j \neq k}(x - x_j)} f(x_k), \tag{1}$$

and as ξ, the following value was denoted: $\xi(l,j) = X(l,j)$ for $l <= n$ and $\xi(l,j) = Y(l-n,j)$ for $l > n$. Thus, an interpolation polynomial is constructed for all calculated values of the boundaries, and then its value is found at $n/2$ - that is, an intermediate value between $X(n,j)$ and $Y(1,j)$ is estimated.

The result of the algorithm is a set of such weights w_j, $j = \overline{1,n}$, that for any of N considered examples, the optimal schedule found for the approximated values of the weighting coefficients completely coincides with its true optimal schedule, corresponding to an unknown true set of weighting coefficients w_j^0, or has the same value of the objective function obtained on it.

This research is partially funded by RFBR, grant 20-58-S52006.

[1] *Lazarev A., Lemtyuzhnikova D., Tyunyatkin A.* Metric Interpolation for the Problem of Minimizing the Maximum Lateness for a Single Machine // Automation and Remote Control, 2021. Vol. 82(10). Pp. 1706–1719.

Определение оптимального варианта оснащенности грузового фронта

Барашов Егор Борисович[1]⋆ barashov.eb@gmail.com
Лемтюжникова Дарья Владимировна[1] darabbt@gmail.com
Давыдов Денис Олегович[2] denius2000@gmail.com
[1]Москва, ИПУ РАН
[2]Москва, Московский Авиационный Институт

Задача выбора оптимального варианта оснащенности грузового фронта или склада является частью задачи управления движением материальных и информационных потоков и входит в круг проблем, которыми занимается логистика.

Грузовой фронт, оснащенный погрузочно-разгрузочными машинами, представляет собой систему массового обслуживания. Входящим потоком заявок являются поступающие на грузовой фронт транспортные средства: вагоны, автомобили, суда.

Оптимальное техническое оснащение грузового фронта определяют в такой последовательности: устанавливают условия работы проектируемого грузового фронта – объем погрузки и выгрузки, режим работы, характер поступления транспортных средств по времени и по количеству, порядок обслуживания грузовых фронтов маневровыми средствами, вид промышленного транспорта, связывающий прирельсовый склад с производственными цехами предприятия; выбирают рациональные варианты погрузочно-разгрузочных и складских работ. Затем строится математическая модель грузового фронта и составляется выражение для подсчета приведенных расходов, представляющее собой функцию количества погрузочно-разгрузочных машин, числа подач вагонов, времени работы грузового фронта, длины повышенных путей или разгрузочной эстакады.

Функционал годовых приведённых расходов для грузового фронта, оснащённого передвижными или стационарными погрузочно-разгрузочными машинами, когда вагоны поступают маршрутами или отдельными группами имеет вид:

$R(X,Y,T = \alpha_1 Y + \alpha_2 \frac{1}{XY} + \alpha_3 X + \alpha_4 YT + \alpha_5 T + \alpha_6[\frac{T}{X} + \frac{(24-T)^2}{24}] + \alpha_7 YT +$
$+ \alpha_8 \frac{X-1}{XY} + \alpha_9 \frac{1}{X} + \alpha_{10} \frac{1}{(X-1)XY} + \alpha 11 \frac{X}{T - X * \tau_m})$, руб.

X, Y, T – параметры оснащенности грузового фронта;

X – количество подач вагонов на грузовой фронт в течение суток;

Y – количество ПРМ данного типа;

T – количество часов работы грузового фронта в сутки.

Коэффициенты при переменных учитывают следующие виды расходов:

α_1 – сумма амортизационных и годовых приведенных расходов, составляющих долю от капитальных затрат на одну ПРМ рассматриваемого типа;

α_2 – стоимость простоев подвижного состава (например, вагонов) под погрузочно-разгрузочными операциями – технологических простоев;

α_3 – стоимость маневровых работ;

α_4 – расходы на оплату труда бригады механизаторов, обслуживающих одну погрузочно-разгрузочную машину;

α_5 – расходы на оплату труда складскому персоналу, обслуживающему данный грузовой фронт;

α_6 – расходы, связанные с неполносуточным режимом работы грузового фронта;

α_7 – расходы на оплату энергоносителей;

α_8 – расходы, связанные с вагоно-часами ожидания подачи (X – 1) групп вагонов на подъездной путь склада и накопления на станционных путях при погрузке-выгрузке маршрута частями;

α_9 – приведенные расходы на содержание эстакады повышенного пути (учитываются для вариантов разгрузки навалочных грузов на повышенном пути);

$\alpha_{10} = \alpha_2 K_{on}, K_{on}$ – коэффициент увеличения объема операций погрузки-выгрузки;

$\alpha_{11} = 2\alpha_2 \tau_m^2, \tau_m$ – суммарная средняя продолжительность маневровых операций на грузовом фронте по постановке-уборке одной подачи вагонов.

Данную модель мы можем использовать для различных сценариев, таких как например сценарий со следующей технологической схемой: вагон-электропогрузчик-склад. Формирование груза в вагоне выполняется вручную на поддон размером 1200 х 800 мм. На каждый поддон укладывается 12 ящиков. Общая масса подъема груза на поддоне составляет 300 кг (25 кг х 12 ящ.).

Исследование выполнено при финансовой поддержке РФФИ, НТУ «Сириус», ОАО «РЖД» и Образовательного Фонда «Талант и успех» в рамках научного проекта №20-38-51010.

[1] *Барашов Е. Б., Лемтюжникова Д. В., Гришин Е. М.* Оптимизация загрузки грузовых фронтов // Материалы XVII Всероссийской школы-конференции молодых ученых Управление Большими Системами, 2021.

Determination of the optimal option for equipping the cargo front

Barashov Egor[1]⋆ barashov.eb@gmail.com
Lemtiuzhnikova Daria[1] darabbt@gmail.com
Davydov Denis[2] denius2000@gmail.com
[1]Moscow, V. A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences
[2]Moscow, Moscow Aviation Institute

The problem of selecting the optimum equipment option for a loading complex or warehouse is part of the task of managing the flow of material and information flows and is part of the range of problems dealt with by logistics.

A loading complex equipped with loading and unloading machines is a queueing system. The entering flux of applications is the vehicles arriving at the loading complex: wagons, cars, ships.

The optimal technical equipment of the loading complex is determined in the following sequence: set working conditions of the projected loading complex — the volume of loading and unloading, operating mode, the nature of the receipt of vehicles by time and by quantity, the order of service loading complex by shunting means, type of industrial transport connecting a riverside warehouse with production shops of the enterprise; choose rational options of loading and unloading and warehouse work. Then a mathematical model of the loading complex is built and the formula for calculating reduced costs is written as a function of the number of loading and unloading machines, number of car transfers, operating time of the loading complex, length of the elevated tracks or unloading pier.

The function of the annual present value cost for a loading complex with mobile or stationary loading and unloading machines, when wagons arrive in routes or in separate groups, is as follows:

$R(X,Y,T) = \alpha_1 Y + \alpha_2 \frac{1}{XY} + \alpha_3 X + \alpha_4 YT + \alpha_5 T + \alpha_6 [\frac{T}{X} + \frac{(24-T)^2}{24}] + \alpha_7 YT + + \alpha_8 \frac{X-1}{XY} + \alpha_9 \frac{1}{X} + \alpha_{10} \frac{1}{(X-1)XY} + \alpha 11 \frac{X}{T - X * \tau_m})$, rub.

X, Y, T parameters of the equipment of the loading complex;

X — number of wagon feeds to the loading complex during the day;

Y — number of carriages of the given type;

T — number of hours of operation of the loading complex per day.

The coefficients at variables take into account the following types of costs:

α_1 is the sum of depreciation and annual present value costs as a proportion of the capital cost per PWM of the type in question;

α_2 is the cost of downtime of rolling stock (e.g. wagons) for loading and unloading operations — technological downtime;

α_3 is the cost of shunting operations;

α_4 is the labour costs of a team of mechanics, operating one loading and unloading machine;

α_5 is the labour costs of warehouse personnel operating a given loading complex;

α_6 is the costs associated with part-time daily operation of the loading complex;

α_7 is the energy costs;

α_8 is the costs associated with wagon-hours waiting for delivery (X-1) groups of cars on the sidings of the warehouse and accumulation on the station tracks during loading and unloading of the route in parts;

α_9 is the present cost of maintaining the elevated track trestle (considered for bulk unloading options on the elevated track);

$\alpha_{10} = \alpha_2 K_{on}, K_{on}$ is the increase in loading and unloading volume;

$\alpha_{11} = 2\alpha_2 \tau_m^2, \tau_m$ is the total average duration of shunting operations on the loading complex for staging-discharging one train feed.

We can use this model for different scenarios, such as the following workflow scenario: wagon-forklift-warehouse. The load in the wagon is formed manually on a pallet of 1200 x 800 mm. Each pallet is loaded with 12 boxes. The total weight of the goods lifted on the pallet is 300 kg (25 kg x 12 crates).

The research was supported by RFBR, NTU "Sirius", JSC "Russian Railways" and Educational Foundation "Talent and Success" within the framework of the scientific project No 20-38-51010.

[1] *Barashov E., Lemtyuzhnikova D., Grishin E.* Optimization of loading of cargo fronts // Materials of XVII All-Russian School-Conference of Young Scientists Management Of Large Systems.

Содержание

Интеллектуальный анализ данных 10

Двоенко С. Д., Пшеничный Д. О.
Метрическая коррекция парных сравнений на основе прямого изменения собственных значений 10

Двоенко С. Д., Курбаков М. Ю.
Проверка согласованности метрик качества изображений 14

Немирко А. П.
Оценка близости выпуклых оболочек для задач машинного обучения . 18

Бериков В. Б., Литвиненко А.
Слабо-контролируемое обучение на основе матрицы нечетких отношений 22

Генрихов И. Е., Дюкова Е. В.
Поиск неприводимых пороговых ассоциативных правил в частично упорядоченных данных 26

Викентьев А. А.,, Бериков В. Б.
Машинное обучение на логических высказываниях: меры сходства, нетривиальности и кластеризация логических формул 32

Тырсин А. Н.
Энтропийное моделирование сетевых структур 36

Драгунов Н. А., Дюкова Е. В.
Асимптотически оптимальная расшифровка монотонной логической функции 38

Фадеев Е. П., Ященко М. А., Зубюк А. В.
Комбинирование рейтингов, полученных из разных источников 43

Ланге М. М., Ланге А. М.
Теоретико-информационный подход к построению нижних границ вероятности ошибки в задачах кодирования дискретного источника и классификации данных 49

Сенько О. В., Салманов М. Ю.
Алгоритм распознавания, основанный на иерархической кластеризации с метрикой специального вида 55

Дюкова Е. В., Масляков Г. О.
Корректная классификация над произведением частичных порядков . 59

Неделько В. М.
О корреляции риска с оценкой скользящего экзамена 64

Карандашев Я. М., Шамин А. Ю.
О нейросетевом подходе к решению классов дифференциальных уравнений 67

Королев Н. С., Сенько О. В.
Метод повышения эффективности обучения градиентного бустинга, основанный на модифицированных функциях потерь 73

Сурков Е. Э., Середин О. С., Копылов А. В., Двоенко С. Д.
Визуализация многомерных данных на основе построения кратчайшего незамкнутого пути 78

Кирилюк И. Л., Сенько О. В.
Применение методов Монте-Карло в задачах анализа временных рядов с мультиколлинеарностью 84

Шибзухов З. М.
Об одной робастной схеме градиентного бустинга 88

Визильтер Ю. В.
Морфологическая теория простоты 94

Визильтер Ю. В.
Теория простоты и мозаичные покрытия 99

Смирнов В. Ю., Кузнецова А. В.
Моделирование циклических процессов решениями кусочно-линейных разностных уравнений с постоянными коэффициентами по экспериментальным данным в виде временных рядов 105

Майсурадзе А. И., Колосов А. М.
Нейросетевой метод решения задачи обобщённого неметрического многомерного шкалирования 111

Нейронные сети и глубокое обучение 117

Яковлев К. Д., Гребенькова О. С., Бахтеев О. Ю., Стрижов В. В.
Дифференцируемый алгоритм поиска архитектуры с контролем сложности 117

Горпинич М., Бахтеев О. Ю., Стрижов В. В.
Градиентные методы оптимизации метапараметров в задаче дистилляции знаний 119

Григорьев А. Д., Гнеушев А. Н.
Регуляризация параметров нейронной сети на основе неравенства Рисса 121

Гребенькова О. С., Бахтеев О. Ю., Стрижов В. В.
Порождение моделей заданной сложности с использованием байесовских гиперсетей 123

Сороковиков П. С., Горнов А. Ю.
Вычислительные технологии поиска низкопотенциальных состояний кластеров Морса размерностей от 460 до 690 атомов 127

Анциперов В. Е.
Генеративная модель автокодировщиков, обучающихся на изображениях представленых выборкмаи отсчетов 131

Гаджиев И. М., Доленко С. А.
Сверточный иерархический нейросетевой классификатор 137

Грабовой А. В., Стрижов В. В.
Априорное распределение параметров в задачах выбора моделей глубокого обучения 142

Методы оптимизации для интеллектуального анализа данных 145

Горнов А. Ю., Зароднюк Т. С.
Методика оценки степени несепарабельности функции 145

Ручкин К., Ручкин А.
Детектирование периодических решений с помощью алгоритма BFOA на неполных картах Пуанкаре 147

Горнов А. Ю.
Q-поиск: удачный метод для задачи безусловной минимизации 149

Аникин А. С.
Модификация метода LBFGS с экономичным одномерным поиском . . 151

Вычислительная сложность и приближенные методы 155

Кутненко О. А., Плясунов А. В.
Вычислительная сложность задачи цензурирования данных 155

Ерохин В. И., Кадочников А. П., Сотников С. В.
Достаточные условия полиномиальной сложности решения интервальных систем линейных алгебраических уравнений в задачах построения линейных зависимостей с интервальной неопределенностью данных . . 161

Ваганов С. Е., Хашин С. И.
Порождение целочисленных алгоритмов генетическими методами . . . 167

Коптелов Д. А., Местецкий Л. М.
Построение диаграммы Вороного для сайтов-многоугольников на основе алгоритма заметания 171

Ломов Н. А.
Скелет многоугольной фигуры с выпуклым многоугольным структурирующим элементом: формализация и эффективный алгоритм построения 176

Карацуба Е. А.
Сложность вычисления Гамма-функции Эйлера 182

Обработка и анализ изображений и сигналов, компютерное зрение 187

Гришин В. А.
Постановка задачи формирования оптимального покрытия области неопределенности эталонами для систем оптической навигации 187

Бобков А. В., Дай И.
Методы 3D-реконструкции поверхности в задаче автономной навигации робота-марсохода 193

Захаров А. А., Жизняков А. Л.
Методы компьютерного зрения на основе спектральной теории графов 198

Свитов Д. В., Алямкин С. А.
Уменьшение числа ложных срабатываний детектора для домофонов с био-идентификацией 200

Свитов Д. В., Алямкин С. А.
Дистилляция моделей для распознавания лиц, обученных с применением softmax с отступами 205

Бобков А. В., Хтет А.
Идентификация человека в реальном времени с помощью конструктора приложений MATLAB на основе YOLO v2 и VGG 16 211

Бахтеев О. Ю., Горленко Т. А., Каприелова М. С., Кильдяков А. С., Огальцов А. В., Финогеев Е. Л., Чехович Ю. В.
Поиск заимствованных изображений в больших коллекциях научных документов 215

Обухов Д. С.
Клонирование и конверсия произвольного голоса с использованием генеративных потоков 220

Алешновский В. С., Безрукова А. В., Зюзина Н. А., Газарян В. А., Курбатова Ю. А., Чуличков А. И., Шапкина Н. Е.
Модели восстановления пропущенных данных во временных рядах концентрации углекислого газа 223

Минаев Е. Ю., Жердева Л. А., Фурсов В. А.
Визуальная одометрия по изображениям опорной поверхности с малыми межкадро-выми поворотами 229

Бериков В. Б., Козинец Р. М.
Интерпретируемое распознавание изображений с помощью логических решающих функций 235

Чучупал В. Я.
Экономная модель трансформера для акустического моделирования речи 239

Филин А. И., Копылов А. В., Середин О. С., Грачева И. А., Сурков Е. Э., Спицын Д. Р., Давыдкин Д. Р., Костинский А. Н.
NIGHT-HAZE: набор данных для оценки алгоритмов удаления тумана с изображений, полученных в темное время суток 245

Князев Д. В., Мурашов Д. М.
Сравнительный анализ алгоритмов в задаче сегментации срезов красочного слоя картин 251

Карандашев Я. М., Марков А. С., Котляров Е. Ю.
Использование нейронных сетей для выявления аномалий на снимках полученных на сканерах персонального досмотра 257

Ефимов Ю. С., Матвеев И. А.
Детектирование подделок в мобильных системах распознавания по лицу при помощи стереокамеры 264

Иванов Д. А., Ольховников С. Ю.
Детекция устаревших регионов в карте для лидарной локализации при помощи синтетичских модификаций лидарных облаков 266

Рихтер А. А., Мурынин А. Б., Гвоздев О. Г., Козуб В. А., Пуховский Д. Ю.
Параметрическая оценка наблюдаемых объектов по перспективным изображениям на базе методов перспективной геометрии, типизированных элементов и свёрточных нейронных сетей 271

Мурашов Д. М.
Комбинирование сегментированных изображений на основе минимизации информационной избыточности 277

Добролюбова О. А.
Особенности моделей прогнозирования на основе компонент временного ряда в эпидемиологии и экономике 281

Качура А. С., Липкина А. Л., Литовских Е. В., Рейер И. А.
Поиск ключевых слов на изображениях рукописей средневековых исландских нарративных памятников 283

Информационный поиск и анализ текстов 288

Инякин А. С., Кормаков Г. В., Каширин Д. О., Мусин Ш. Н., Сотнезов Р. М., Разин Н. А., Саутенков И. С.
Классификация сообщений новостного потока, предположительно «участвующих» в реализации PUMP-стратегий на фондовом рынке 288

Михайлов Д. В., Емельянов Г. М.
Согласование смысловых эталонов и взаимная релевантность документов тематического корпуса 294

Сафин К. Ф., Чехович Ю. В.
Определение факта заимствования в текстовых документах без указания источника 300

Галеев Д. Т., Панищев В. С.
Применение нейронных сетей в вопросно-ответных системах на русском языке 304

Скачков Н. А., Воронцов К. В.
Улучшение качества машинного перевода с использованием обратной модели 308

Алексеев В. А., Воронцов К. В.
Банк тем: сбор интерпретируемых тем с помощью множественного обучения тематических моделей и их дальнейшее использование для оценки качества тематических моделей 313

Сулейманова Е. А., Трофимов И. В.
Дата или не дата? Опыт обучения нейросети разрешению неоднозначности темпоральных выражений 319

Максимов Н. В.
Когнитивно-подобный документальный информационный поиск: концепция и технологии 324

Шевченко О. В., Гращенков К. В., Чащин А. В., Грабовой А. В.
Многозадачное обучение в задаче рубрикации научных документов . . 329

Крыжановская С. Ю., Власов А. В., Еремеев М. А., Воронцов К. В.
Полуавтоматическая суммаризация тематических подборок научных публикаций: задачи и подходы 333

Хрыльченко К. Я., Воронцов К. В.
Оптимизация весов модальностей в тематических моделях транзакционных данных 339

Герасименко Н. А., Чернявский А. С., Никифорова М. А., Никитин М. Д., Воронцов К. В.
Инкрементные тематические модели с аддитивной регуляризацией для выделения трендовых научных тем 344

Рамазанова А., Янина А. О., Воронцов К. В.
Нейронные тематические модели для рекомендации статей 350

Сердюк Ю. А., Воронцов К. В.
Реализация EM-алгоритма для аддитивно регуляризованных тематических моделей на GPU 356

Воронцов К. В.
Задачи и методы понимания естественного языка для мониторинга медиа-пространства 362

Анализ данных веба и социальных сетей 368

Дюличева Ю. Ю.
Подходы к выявлению тревожных расстройств на основе автоматического анализа текстов комментариев в социальных сетях 368

Индустриальные приложения науки о данных 374

Инякин А. С., Мотренко А. П., Руденко И. Н., Кормаков Г. В., Каширин Д. О., Чипак Е. О.
Классификация видов активности домашних животных по данным, получаемым с сенсоров носимых устройств 374

Мортин К. В.
Разработка системы детекции аномалий с целью автоматизации визуального контроля поверхности листового металлопроката 380

Астафьев А. В., Жизняков А. Л., Демидов А. А., Кондрушин И. Е.
Исследование применимости использования информации о состоянии канала передачи данных для организации позиционирования внутри помещений 385

Старожилец В. О., Чехович Ю. В.
Использование мезоскопической модели для моделирования транспортных потоков на МКАД и управления въездами 389

Грызлова Т. П.
Информативные образы нестационарных цифровых сигналов в диагностических системах 393

Анализ биомедицинских данных, биоинформатика 399

Сушкова О. С., Морозов А. А., Габова А. В., Кершнер И. А., Чигалейчик Л. А., Карабанов А. В.
Анализ фаз огибающих ЭМГ мышц-антагонистов у пациентов с нейродегенеративными заболеваниями 399

Торшин И. Ю., Рудаков К. В.
Топологическая теория анализа хемографов как перспективный подход к имитационному моделированию квантово-механических свойств молекул 403

Манило Л. А., Немирко А. П., Алексеев Б. Э.
Распознавание опасных аритмий для выявления рисков осложнений сердечно-сосудистых заболеваний 405

Обухов Ю. В., Кершнер И. А., Синкин М. В.
Применение хребтов вейвлет спектров в выделении диагностических признаков ЭЭГ: Длительный мониторинг эпилепсии 410

Махортых С. А., Москаленко А. В.
Перспективы использования обобщённого спектрального анализа в решении задач распознавания паттернов вариабельности ритма сердца . 412

Сенько О. В., Кузнецова А. В., Добролюбова О. А., Воронин Е. М., Акимкин В. Г., Плоскирева А. А.
Использование методов кластерного анализа в исследовании эпидемических процессов COVID-19 в странах мира. 415

Обухов Ю. В., Толмачева Р. А., Жаворонкова Л. А.
Оценка восстановления межканальных фазовых связей электроэнцефалограмм при когнитивных тестах у пациентов с черепно-мозговой травмой средней тяжести до и после реабилитации 420

Рыкунов С. Д., Устинин М. Н., Бойко А. И.
Оценка корреляций между компартментами мозга при синдроме дефицита внимания и гиперактивности методом виртуальных электродов . 422

Бойко А. И., Рыкунов С. Д., Устинин М. Н.
Программный комплекс для компьютерного моделирования магнитной и электрической активности головного мозга человека и интеллектуального анализа модельных данных 428

Устинин М. Н., Бойко А. И., Рыкунов С. Д.
Изучение особенностей пространственного распределения откликов на различные речевые стимулы по данным электроэнцефалографии . . . 432

Ямаев А. В., Чуличков А. И.
Алгоритм точного обучения частотного фильтра для задачи малоракурсной компьютерной томографии 434

Методы математического моделирования в интеллектуальном анализе данных . 438

Фаломкина О. В., Пытьев Ю. П., Чуличков А. И., Пятков Ю. В., Жучко В. Е., Каманин Д. В., Горяйнова З. И.
Новый метод определения скорости тяжелого иона, основанный на математическом формализме субъективного моделирования 438

Интеллектуальный анализ геопространственных данных 442

Исаев И. В., Оборнев И. Е., Оборнев Е. А., Родионов Е. А., Шимелевич М. И., Доленко С. А.
Использование классификации
при решении регрессионной обратной задачи
разведочной геофизики
как способ повышения устойчивости решения
к шумам в данных 442

Интеллектуальная оптимизация и эффективный менеджмент 446

Белозуб В. А., Козлова М. Г., Лукъяненко В. А.
Восстановление решений уравнений типа Урысона 446

Германчук М. С., Козлова М. Г., Руденко Л. И.
Интеллектуализация обработки информации социальных сетей 452

Гришин Е. М.
Радиус устойчивости и робастное расписание на примере задачи железнодорожного планирования 457

Лукъяненко В. А., Германчук М. С., Макаров О. О.
Специфика задач маршрутизации в условиях локальных преобразований сети 460

Букуева Е. С., Кудинов И. Д., Лемтюжникова Д. В.
Метрический подход для задач на быстродействие 466

Лазарев А. А., Лемтюжникова Д. В., Сомов М. Л.
Планирование операций с участием анестезиолога в операционных комнатах 468

Морозов Н. Ю., Гришин Е. М., Коровкин Д. М.
Сравнение булевой и целочисленной моделей для задачи пункта перевалки морской порт – железная дорога 470

Галахов С. А.
О свойствах решения и целевой функции в теории расписаний 474

Барашов Е. Б., Лемтюжникова Д. В., Тюняткин А. А.
Аппроксимация весов задачи $1||\sum w_j C_j$ 476

Барашов Е. Б., Лемтюжникова Д. В., Давыдов Д. Д.
Определение оптимального варианта оснащенности грузового фронта 480

Содержание 484

Авторский указатель 501

Contents

Data mining 10

Dvoenko S., Pshenichny D.
Metric correction of paired comparisons based on direct changing of eigenvalues 12

Dvoenko S., Kurbakov M.
Checking the consistency of image quality metrics 16

Nemirko A.
Estimating the proximity of convex hulls for machine learning problems . 20

Berikov V., Litvinenko A.
Weakly supervised learning based on a fuzzy relationship matrix 24

Genrikhov I., Djukova E.
Finding irredundant threshold association rules in partially ordered data . 29

Vikentiev A., Berikov V.
Machine learning on logical statements: measures of similarity, nontrivialities and clustering of logical formulas 34

Tyrsin A.
Entropy modeling of network structures 37

Dragunov N., Djukova E.
Asymptotically optimal decoding of a monotone logical function 41

Fadeev E., Yashchenko M., Zubyuk A.
Aggregation of ratings from various sources 46

Lange M., Lange A.
Information-theoretical approach to construct lower bounds for error probability in tasks of discrete source coding and data classification 52

Senko O., Salmanov M.
Recognition algorithm based on hierarchical clustering with a metric of a special kind 57

Djukova E., Masliakov G.
Correct classification over a product of partial orders 62

Nedel'ko V.
On the correlation of risk with the cross-validation estimate 66

Karandashev I., Shamin A.
On a neural network approach to solving classes of differential equations . 70

Korolev N., Senko O.
Method for improving generalization performance of gradient boosting . . 75

Surkov E., Seredin O., Kopylov A., Dvoenko S.
Multidimensional data visualization based on the shortest unclosed path search 81

Kirilyuk I., Senko O.
Application of Monte Carlo methods in the tasks of analysis of time series with multicollinearity 86

Shibzukhov Z.
One Robust Scheme of Gradient Boosting 91

Vizilter Yu.
The Morphological Simplicity Theory 96

Vizilter Yu.
The Simplicity Theory and Mosaic Coverings 102

Smirnov V., Kuznetsova A.
Modeling of Cyclic Processes (of time series type experimental date) by Solving of Piecewise Linear Differential Equations with Constant Coefficients 108

Maysuradze A., Kolosov A.
Neural network method for solving the problem of generalized non-metric multidimensional scaling 114

Neural networks and deep learning 117

Yakovlev K., Grebenkova O., Bakhteev O., Strijov V.
Differentiable architecture search with model complexity control 118

Gropinich M., Bakhteev O., Strijov V.
Gradient-based metaparameter optimization in knowledge distillation task 120

Grigorev A., Gneushev A.
Neural network parameters regularization based on Riesz inequality 122

Grebenkova O., Bakhteev O., Strijov V.
Model selection using Bayesian hypernetworks 125

Sorokovikov P., Gornov A.
Computational technologies for the search for low-potential states of Morse clusters with dimensions from 460 to 690 atoms 129

Antsiperov V.
Generative model of autoencoders learning images by count sample representations 134

Gadzhiev I., Dolenko S.
Convolutional hierarchical neural network classifier 140

Grabovoy A., Strijov V.
Prior distribution of parameters for the deep learning model selection problem 144

Data mining optimization techniques 145

Gornov A., Zarodnyukr T.
Technique for estimating the inseparability degree of a function 146

Ruchkin C., Ruchkin A.
Detecting periodic solutions using the BFOA algorithm on the incomplete cards Poincare 148

Gornov A.
Q-Search: a successful method for the unconstrained minimization problem 150

Anikin A.
Modification of the LBFGS method with economical line-search 153

Algorithmic complexity and approximate methods 155

Kutnenko O., Plyasunov A.
Computational complexity of data cleaning problem 158

Erokhin V., Kadochnikov A., Sotnikov S.
Sufficient conditions for the polynomial complexity of solving interval systems of linear algebraic equations in problems of constructing linear dependencies with interval uncertainty of data 164

Khashin S., Vaganov S.
Generation of integer algorithms by genetic method 169

Koptelov D., Mestetskiy L.
Sweepline algorithm for Voronoi diagram of polygonal sites 174

Lomov N.
Skeleton of a Polygonal Figure with a Convex Polygonal Structuring Element: Formalization and an Efficient Algorithm of Construction 179

Karatsuba E.
The Complexity of computation of the Euler Gamma function 185

Image and signal processing, computer vision 187

Grishin V.
Formulation of the problem of generation the optimal coverage by reference images of the uncertainty region of optical navigation systems 190

Bobkov A., Dai Y.
3D surface reconstruction in the task of autonomous navigation of a Martian rover 196

Zakharov A., Zhiznyakov A.
Computer vision methods based on spectral graph theory 199

Svitov D., Alyamkin S.
Reducing the number of false positive detections for intercoms with bio-identification 203
Svitov D., Alyamkin S.
Distillation for face recognition neural networks with margin-based softmax 208
Bobkov A., Htet A.
Identification of person in real-time using the MATLAB application designer based on YOLOv2 and VGG 16 213
Bakhteev O., Gorlenko T., Kaprielova M., Kildyakov A., Ogaltsov A., Finogeeev E., Chekhovich Yu.
Image reuse detection in large-scale document scientific collection 218
Obukhov D.
Voice cloning and any-to-any voice conversion using generative flows . . . 222
Aleshnovskiy V., Bezrukova A., Zyuzina N., Gazaryan V., Kurbatova J., Chulichkov A., Shapkina N.
Recovery models of missing data in time series of carbon dioxide concentration226
Minaev E., Zherdeva L., Fursov V.
Visual odometry on surface image sequences with small inter-frame rotations232
Berikov V., Kozinets R.
Interpretable image recognition using logical decision functions 237
Chuchupal V.
A computationally-effective transformer model for acoustic speech modeling 242
Filin A., Kopylov A., Seredin O., Gracheva I., Surkov E., Spitsyn D., Davydkin D., Kostinskii A.
NIGHT-HAZE: A dehazing benchmark with real hazy and haze-free low-light indoor images 248
Knyazev D., Murashov D.
Experimental research of algorithms for segmenting paint layer cross-sections of paintings 254
Karandashev I., Markov A., Kotlyarov E.
Anomaly Detection on full body scanner images with neural networks . . . 261
Efimov I., Matveev I.
Stereo Face Liveness Detection for Mobile Biometric Applications 265
Ivanov D., Olkhovnikov S.
Detection of outdated regions in map for lidar localization using synthetic lidar clouds modifications 269
Rihter A., Murynin A., Gvozdev O., Kozub V., Puhovskij D.
Parametric estimation of observed objects from perspective images based on methods of perspective geometry, typed elements and convolutional neural networks 274

Murashov D.
Combining segmented images based on information redundancy minimization 279
Dobrolyubova O.
Features of forecasting models based on the time series components in epidemiology and economics 282
Kachura A., Lipkina A., Litovskikh E., Reyer I.
Keyword search in images of manuscripts of medieval Icelandic narrative sources 286

Information retrieval and text analysis 288
Inyakin A., Kormakov G., Kashirin D., Musin S., Sotnezov R., Razin N., Sautenkov I.
Classification of news stream messages supposedly "involved" in the implementation of PUMP strategies on the stock market 291
Mikhaylov D., Emelyanov G.
Coherence of semantic patterns and mutual relevancy of topical corpus documents 297
Safin K., Chekhovich Yu.
Intrinsic methods for plagiarised texts detection 302
Galeev D., Panishchev V.
Application of neural networks in question answering systems in Russian . 306
Skachkov N., Vorontsov K.
Machine translation quality improvement using reverse translation model . 311
Alekseev V., Vorontsov K.
TopicBank: Collection of coherent topics using multiple model training with their further use for topic model validation 316
Suleymanova E., Trofimov I.
A date or not a date? Word sense disambiguation for temporal expression recognition 322
Maksimov N.
Cognitive-like documentary information search: concept and technologies . 327
Shevchenko O., Grashchenkov K., Chashchin A., Grabovoy A.
Multi-task Learning in the Problem of Rubrication of Scientific Documents 331
Kryzhanovskaya S., Vlasov A., Eremeev M., Vorontsov K.
Machine Aided Human Summarization of scientific articles: tasks and approaches 336
Khrylchenko K., Vorontsov K.
Optimizing modality weights for topic models of transaction data 342

Gerasimenko N., Chernyavskiy A., Nikiforova M., Nikitin M., Vorontsov K.
Incremental ARTM for Scientific Trend Topics Detection 347

Ramazanova A., Ianina A., Vorontsov K.
Neural Topic Models for Article Recommendation 353

Serdyuk J., Vorontsov K.
GPU implementation of parallel EM-algorithm for additive regularization of topic models 359

Vorontsov K.
Problems and approaches of natural language understanding for media monitoring 365

Analysis of web and social network data 368

Dyulicheva Yu.
The Approaches to the Anxiety Disorders Detection based on the Text Mining of Social Media Comments 371

Industrial data science applications 374

Inyakin A., Motrenko A., Rudenko I., Kormakov G., Kashirin D., Chipak E.
Pets activities classification based on analysis of data obtained from sensors of wearable devices 377

Mortin K.
Development of an anomaly detection system for the purpose of automating visual inspection of the surface of sheet metal 383

Astafiev A., Zhiznyakov A., Demidov A., Kondrushin I.
Investigation the applicability of using channel state information for indoor positioning 387

Starozhilets V., Chekhovich Yu.
About use of the mesoscopic model for traffic flows modeling on the Moscow Ring Road and enters control 391

Gryzlova T.
The informative Images of the non-stationary digital signals in the systems of diagnostics 396

Analysis of biomedical data, bioinformatics 399

Sushkova O., Morozov A., Gabova A., Kershner I., Chigaleichik L., Karabanov A.
Phase analysis of EMG envelopes of antagonist muscles in patients with neurodegenerative diseases 401

Torshin I., Rudakov K.
Topological theory of chemograph analysis as a promising approach to simulation modeling of quantum mechanical properties of molecules 404

Manilo L., Nemirko A., Alekseev B.
Recognition of dangerous arrhythmias to identify the risks of complications of cardiovascular diseases 408

Obukhov Yu., Kershner I., Sinkin M.
Wavelet Ridges in EEG Diagnostic Features Extraction: Epilepsy Long-Time Monitoring 411

Makhortykh S., Moskalenko A.
Perspectives for using of generalized spectral analysis in solving problems of recognizing patterns of heart rate variability 414

Senko O., Kuznetsova A., Dobrolyubova O., Voronin E., Akimkin V., Ploskireva A.
Clustering methods for COVID-19 epidemic analytics in countries across the world 418

Obukhov Yu., Tolmacheva R., Zhavoronkova L.
Estimation of recovery of interchannel phase connections of electroencephalograms during cognitive tests in patients with moderate traumatic brain injury before and after rehabilitation 421

Rykunov S., Ustinin M., Boyko A.
Evaluation of correlations between brain compartments in attention deficit hyperactivity disorder using the method of virtual electrodes 425

Boyko A., Rykunov S., Ustinin M.
Software for computer modeling of magnetic and electrical activity of the human brain and intelligent analysis of model data 430

Ustinin M., Boyko A., Rykunov S.
Study of the features of spatial distribution of responses to various speech stimuli based on the electroencephalography 433

Yamaev A., Chulichkov A.
Algorithm of exact frequency filter finding for few view tomography problem436

Methods of mathematical modeling in data mining 438

Falomkina O., Pyt'ev Yu., Chulichkov A., Pyatkov Yu., Zhuchko V., Kamanin D., Goryaynova Z.
A new method for determining the velocity of a heavy ion based on the mathematical formalism of subjective modeling 440

Geospatial data mining 442

Isaev I., Obornev I., Obornev E., Rodionov E., Shimelevich M., Dolenko S.
Using classification
in solving regression inverse problem
of exploration geophysics
as a way to improve the resilience of the solution
to noise in data 444

Intelligent optimization and effective management 446

Belozub V., Kozlova M., Lukianenko V.
Recovery of solutions of Urysohn-type equations 449

Germanchuk M., Kozlova M., Rudenko L.
Intellectualization of social network information processing 455

Grishin E.
Stability radius and robust scheduling in a rail planning problem as an example 459

Lukianenko V., Germanchuk M., Makarov O.
The specifics of routing tasks in the context of local network transformations 463

Bukueva E., Kudinov I., Lemtyuzhnikova D.
Metric approach for minimization makespan problem 467

Lazarev A., Lemtyuzhnikova D., Somov M.
Scheduling of Anaesthesia Operations in Operating Rooms 469

Morozov N., Grishin E., Korovkin D.
Comparison of boolean and integer mathematical models for the seaport-railway transhipment point problem 472

Galakhov N.
On solution properties and the objective function in scheduling theory . . 475

Barashov E., Lemtiuzhnikova D., Tyunyakin A.
Approximation of the weights of the problem $1||\sum w_j C_j$ 478

Barashov E., Lemtiuzhnikova D., Davydov D.
Determination of the optimal option for equipping the cargo front 482

Contents 484

Author index 504

Авторский указатель

А

Акимкин В. Г., ... 415
Алексеев Б. Э., ... 405
Алексеев В. А., ... 313
Алешновский В. С., ... 223
Алямкин С. А., ... 200, 205
Аникин А. С., ... 151
Анциперов В. Е., ... 131
Астафьев А. В., ... 385

Б

Барашов Е. Б., ... 476, 480
Бахтеев О. Ю., ... 117, 119, 123, 215
Безрукова А. В., ... 223
Белозуб В. А., ... 446
Бериков В. Б., ... 22, 32, 235
Бобков А. В., ... 193, 211
Бойко А. И., ... 422, 428, 432
Букуева Е. С., ... 466

В

Ваганов С. Е., ... 167
Визильтер Ю. В., ... 94, 99
Викентьев А. А.„ 32
Власов А. В., ... 333
Воронин Е. М., ... 415
Воронцов К. В., . 308, 313, 333, 339, 344, 350, 356, 362

Г

Габова А. В., ... 399
Гаджиев И. М., ... 137
Газарян В. А., ... 223
Галахов С. А., ... 474
Галеев Д. Т., ... 304
Гвоздев О. Г., ... 271
Генрихов И. Е., ... 26
Герасименко Н. А., ... 344
Германчук М. С., ... 452, 460
Гнеушев А. Н., ... 121
Горленко Т. А., ... 215
Горнов А. Ю., ... 127, 145, 149
Горпинич М., ... 119
Горяйнова З. И., ... 438
Грабовой А. В., ... 142, 329
Грачева И. А., ... 245
Гращенков К. В., ... 329
Гребенькова О. С., ... 117, 123
Григорьев А. Д., ... 121
Гришин В. А., ... 187
Гришин Е. М., ... 457, 470
Грызлова Т. П., ... 393

Д

Давыдкин Д. Р., ... 245
Давыдов Д. Д., ... 480
Дай И., ... 193
Двоенко С. Д., ... 10, 14, 78
Демидов А. А., ... 385
Добролюбова О. А., ... 281, 415
Доленко С. А., ... 137, 442
Драгунов Н. А., ... 38
Дюкова Е. В., ... 26, 38, 59
Дюличева Ю. Ю., ... 368

Е

Емельянов Г. М., ... 294
Еремеев М. А., ... 333
Ерохин В. И., ... 161
Ефимов Ю. С., ... 264

Ж

Жаворонкова Л. А., ... 420
Жердева Л. А., ... 229
Жизняков А. Л., ... 198, 385
Жучко В. Е., ... 438

З

Зароднюк Т. С., ... 145
Захаров А. А., ... 198

Зубюк А. В., ... 43
Зюзина Н. А., ... 223

И

Иванов Д. А., ... 266
Инякин А. С., ... 288, 374
Исаев И. В., ... 442

К

Кадочников А. П., ... 161
Каманин Д. В., ... 438
Каприелова М. С., ... 215
Карабанов А. В., ... 399
Карандашев Я. М., ... 67, 257
Карацуба Е. А., ... 182
Качура А. С., ... 283
Каширин Д. О., ... 288, 374
Кершнер И. А., ... 399, 410
Кильдяков А. С., ... 215
Кирилюк И. Л., ... 84
Князев Д. В., ... 251
Козинец Р. М., ... 235
Козлова М. Г., ... 446, 452
Козуб В. А., ... 271
Колосов А. М., ... 111
Кондрушин И. Е., ... 385
Коптелов Д. А., ... 171
Копылов А. В., ... 78, 245
Кормаков Г. В., ... 288, 374
Коровкин Д. М., ... 470
Королев Н. С., ... 73
Костинский А. Н., ... 245
Котляров Е. Ю., ... 257
Крыжановская С. Ю., ... 333
Кудинов И. Д., ... 466
Кузнецова А. В., ... 105, 415
Курбаков М. Ю., ... 14
Курбатова Ю. А., ... 223
Кутненко О. А., ... 155

Л

Лазарев А. А., ... 468
Ланге А. М., ... 49
Ланге М. М., ... 49
Лемтюжникова Д. В., 466, 468, 476, 480
Липкина А. Л., ... 283
Литвиненко А., ... 22
Литовских Е. В., ... 283
Ломов Н. А., ... 176
Лукьяненко В. А., ... 446, 460

М

Майсурадзе А. И., ... 111
Макаров О. О., ... 460
Максимов Н. В., ... 324
Манило Л. А., ... 405
Марков А. С., ... 257
Масляков Г. О., ... 59
Матвеев И. А., ... 264
Махортых С. А., ... 412
Местецкий Л. М., ... 171
Минаев Е. Ю., ... 229
Михайлов Д. В., ... 294
Морозов А. А., ... 399
Морозов Н. Ю., ... 470
Мортин К. В., ... 380
Москаленко А. В., ... 412
Мотренко А. П., ... 374
Мурашов Д. М., ... 251, 277
Мурынин А. Б., ... 271
Мусин Ш. Н., ... 288

Н

Неделько В. М., ... 64
Немирко А. П., ... 18, 405
Никитин М. Д., ... 344
Никифорова М. А., ... 344

О

Оборнев Е. А., ... 442
Оборнев И. Е., ... 442
Обухов Д. С., ... 220
Обухов Ю. В., ... 410, 420
Огальцов А. В., ... 215
Ольховников С. Ю., ... 266

П

Панищев В. С., ... 304
Плоскирева А. А., ... 415
Плясунов А. В., ... 155
Пуховский Д. Ю., ... 271
Пшеничный Д. О., ... 10
Пытьев Ю. П., ... 438
Пятков Ю. В., ... 438

Р

Разин Н. А., ... 288
Рамазанова А., ... 350
Рейер И. А., ... 283
Рихтер А. А., ... 271
Родионов Е. А., ... 442
Рудаков К. В., ... 403
Руденко И. Н., ... 374
Руденко Л. И., ... 452
Ручкин А., ... 147
Ручкин К., ... 147
Рыкунов С. Д., ... 422, 428, 432

С

Салманов М. Ю., ... 55
Саутенков И. С., ... 288
Сафин К. Ф., ... 300
Свитов Д. В., ... 200, 205
Сенько О. В., ... 55, 73, 84, 415
Сердюк Ю. А., ... 356
Середин О. С., ... 78, 245
Синкин М. В., ... 410
Скачков Н. А., ... 308
Смирнов В. Ю., ... 105
Сомов М. Л., ... 468
Сороковиков П. С., ... 127
Сотнезов Р. М., ... 288
Сотников С. В., ... 161
Спицын Д. Р., ... 245
Старожилец В. О., ... 389
Стрижов В. В., ... 117, 119, 123, 142
Сулейманова Е. А., ... 319
Сурков Е. Э., ... 78, 245
Сушкова О. С., ... 399

Т

Толмачева Р. А., ... 420
Торшин И. Ю., ... 403
Трофимов И. В., ... 319
Тырсин А. Н., ... 36
Тюняткин А. А., ... 476

У

Устинин М. Н., ... 422, 428, 432

Ф

Фадеев Е. П., ... 43
Фаломкина О. В., ... 438
Филин А. И., ... 245
Финогеев Е. Л., ... 215
Фурсов В. А., ... 229

Х

Хашин С. И., ... 167
Хрыльченко К. Я., ... 339
Хтет А., ... 211

Ч

Чащин А. В., ... 329
Чернявский А. С., ... 344
Чехович Ю. В., ... 215, 300, 389
Чигалейчик Л. А., ... 399
Чипак Е. О., ... 374
Чуличков А. И., ... 223, 434, 438
Чучупал В. Я., ... 239

Ш

Шамин А. Ю., ... 67
Шапкина Н. Е., ... 223
Шевченко О. В., ... 329
Шибзухов З. М., ... 88
Шимелевич М. И., ... 442

Я

Яковлев К. Д., ... 117
Ямаев А. В., ... 434
Янина А. О., ... 350
Ященко М. А., ... 43

Author index

A

Akimkin V., ... 418
Alekseev B., ... 408
Alekseev V., ... 316
Aleshnovskiy V., ... 226
Alyamkin S., ... 203, 208
Anikin A., ... 153
Antsiperov V., ... 134
Astafiev A., ... 387

B

Bakhteev O., ... 118, 120, 125, 218
Barashov E., ... 478, 482
Belozub V., ... 449
Berikov V., ... 24, 34, 237
Bezrukova A., ... 226
Bobkov A., ... 196, 213
Boyko A., ... 425, 430, 433
Bukueva E., ... 467

C

Chashchin A., ... 331
Chekhovich Yu., ... 218, 302, 391
Chernyavskiy A., ... 347
Chigaleichik L., ... 401
Chipak E., ... 377
Chuchupal V., ... 242
Chulichkov A., ... 226, 436, 440

D

Dai Y., ... 196
Davydkin D., ... 248
Davydov D., ... 482
Demidov A., ... 387
Djukova E., ... 29, 41, 62
Dobrolyubova O., ... 282, 418
Dolenko S., ... 140, 444
Dragunov N., ... 41
Dvoenko S., ... 12, 16, 81
Dyulicheva Yu., ... 371

E

Efimov I., ... 265
Emelyanov G., ... 297
Eremeev M., ... 336
Erokhin V., ... 164

F

Fadeev E., ... 46
Falomkina O., ... 440
Filin A., ... 248
Finogeeev E., ... 218
Fursov V., ... 232

G

Gabova A., ... 401
Gadzhiev I., ... 140
Galakhov N., ... 475
Galeev D., ... 306
Gazaryan V., ... 226
Genrikhov I., ... 29
Gerasimenko N., ... 347
Germanchuk M., ... 455, 463
Gneushev A., ... 122
Gorlenko T., ... 218
Gornov A., ... 129, 146, 150
Goryaynova Z., ... 440
Grabovoy A., ... 144, 331
Gracheva I., ... 248
Grashchenkov K., ... 331
Grebenkova O., ... 118, 125
Grigorev A., ... 122
Grishin E., ... 459, 472
Grishin V., ... 190
Gropinich M., ... 120
Gryzlova T., ... 396
Gvozdev O., ... 274

H

Htet A., ... 213

I

Ianina A., ... 353
Inyakin A., ... 291, 377
Isaev I., ... 444
Ivanov D., ... 269

K

Kachura A., ... 286
Kadochnikov A., ... 164
Kamanin D., ... 440
Kaprielova M., ... 218
Karabanov A., ... 401
Karandashev I., ... 70, 261
Karatsuba E., ... 185
Kashirin D., ... 291, 377
Kershner I., ... 401, 411
Khashin S., ... 169
Khrylchenko K., ... 342
Kildyakov A., ... 218
Kirilyuk I., ... 86
Knyazev D., ... 254
Kolosov A., ... 114
Kondrushin I., ... 387
Koptelov D., ... 174
Kopylov A., ... 81, 248
Kormakov G., ... 291, 377
Korolev N., ... 75
Korovkin D., ... 472
Kostinskii A., ... 248
Kotlyarov E., ... 261
Kozinets R., ... 237
Kozlova M., ... 449, 455
Kozub V., ... 274
Kryzhanovskaya S., ... 336
Kudinov I., ... 467
Kurbakov M., ... 16
Kurbatova J., ... 226
Kutnenko O., ... 158
Kuznetsova A., ... 108, 418

L

Lange A., ... 52
Lange M., ... 52
Lazarev A., ... 469
Lemtiuzhnikova D., ... 478, 482
Lemtyuzhnikova D., ... 467, 469
Lipkina A., ... 286
Litovskikh E., ... 286
Litvinenko A., ... 24
Lomov N., ... 179
Lukianenko V., ... 449, 463

M

Makarov O., ... 463
Makhortykh S., ... 414
Maksimov N., ... 327
Manilo L., ... 408
Markov A., ... 261
Masliakov G., ... 62
Matveev I., ... 265
Maysuradze A., ... 114
Mestetskiy L., ... 174
Mikhaylov D., ... 297
Minaev E., ... 232
Morozov A., ... 401
Morozov N., ... 472
Mortin K., ... 383
Moskalenko A., ... 414
Motrenko A., ... 377
Murashov D., ... 254, 279
Murynin A., ... 274
Musin S., ... 291

N

Nedel'ko V., ... 66
Nemirko A., ... 20, 408
Nikiforova M., ... 347
Nikitin M., ... 347

O

Obornev E., ... 444
Obornev I., ... 444
Obukhov D., ... 222
Obukhov Yu., ... 411, 421
Ogaltsov A., ... 218

Olkhovnikov S., ... 269

P

Panishchev V., ... 306
Ploskireva A., ... 418
Plyasunov A., ... 158
Pshenichny D., ... 12
Puhovskij D., ... 274
Pyatkov Yu., ... 440
Pyt'ev Yu., ... 440

R

Ramazanova A., ... 353
Razin N., ... 291
Reyer I., ... 286
Rihter A., ... 274
Rodionov E., ... 444
Ruchkin A., ... 148
Ruchkin C., ... 148
Rudakov K., ... 404
Rudenko I., ... 377
Rudenko L., ... 455
Rykunov S., ... 425, 430, 433

S

Safin K., ... 302
Salmanov M., ... 57
Sautenkov I., ... 291
Senko O., ... 57, 75, 86, 418
Serdyuk J., ... 359
Seredin O., ... 81, 248
Shamin A., ... 70
Shapkina N., ... 226
Shevchenko O., ... 331
Shibzukhov Z., ... 91
Shimelevich M., ... 444
Sinkin M., ... 411
Skachkov N., ... 311
Smirnov V., ... 108
Somov M., ... 469
Sorokovikov P., ... 129
Sotnezov R., ... 291
Sotnikov S., ... 164
Spitsyn D., ... 248
Starozhilets V., ... 391
Strijov V., ... 118, 120, 125, 144
Suleymanova E., ... 322
Surkov E., ... 81, 248
Sushkova O., ... 401
Svitov D., ... 203, 208

T

Tolmacheva R., ... 421
Torshin I., ... 404
Trofimov I., ... 322
Tyrsin A., ... 37
Tyunyakin A., ... 478

U

Ustinin M., ... 425, 430, 433

V

Vaganov S., ... 169
Vikentiev A., ... 34
Vizilter Yu., ... 96, 102
Vlasov A., ... 336
Voronin E., ... 418
Vorontsov K., ... 311, 316, 336, 342, 347, 353, 359, 365

Y

Yakovlev K., ... 118
Yamaev A., ... 436
Yashchenko M., ... 46

Z

Zakharov A., ... 199
Zarodnyukr T., ... 146
Zhavoronkova L., ... 421
Zherdeva L., ... 232
Zhiznyakov A., ... 199, 387
Zhuchko V., ... 440
Zubyuk A., ... 46
Zyuzina N., ... 226

MachineLearning.ru

http://www.machinelearning.ru/

Профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных. Цели ресурса — сконцентрировать информацию о достижения ведущих научных школ; способствовать обмену опытом, накоплению и распространению научных знаний; предоставить площадку для виртуальных научных семинаров и обсуждений.

Научное издание

Математические методы распознавания образов

Тезисы докладов
20-й Всероссийской конференции
с международным участием

Подписано в печать 27.12.2021
Формат 60×84 1/8
Усл.-печ. л. 22,2. Уч.-изд. л. 23,8
Тираж 50 экз

Издатель — Российская Академия Наук

Печать — УНИД РАН

Отпечатано в экспериментальной цифровой типографии РАН

Издается по распоряжению президиума РАН
и распространяется бесплатно