Российская академия наук
Федеральный исследовательский центр
«Информатика и управление» РАН
Вычислительный центр им. А. А. Дородницына РАН
Московский физико-технический институт
Harbour.Space University Barcelona

при поддержке

Российского фонда фундаментальных исследований,
компаний Форексис и ЦСПиР

Интеллектуализация обработки информации

ИОИ-2016

Тезисы докладов
11-й Международной конференции
(Москва, Россия – Барселона, Испания)

ТОРУС ПРЕСС МОСКВА 2016

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

И 73 **Интеллектуализация обработки информации:** Тезисы докладов 11-й Международной конференции (Москва, Россия – Барселона, Испания). — М.: ТОРУС ПРЕСС, 2016. — 238 с.
ISBN 978-5-94588-207-2

В сборнике представлены тезисы докладов 11-й Международной конференции «Интеллектуализация обработки информации», проводимой Вычислительным центром им. А. А. Дородницына ФИЦ ИУ РАН, Московским физико-техническим иститутом и Harbour.Space University Barcelona.

Конференция проводится с 1989 г., начиная с 2000 г. — регулярно один раз в два года, и является представительным научным форумом в области интеллектуального анализа данных, машинного обучения, распознавания образов, анализа изображений, обработки сигналов, дискретного анализа.

Организационный комитет ИОИ-2016 выражает особую благодарность РФФИ и компаниям Форексис и ЦСПиР, оказавшим неоценимую помощь при подготовке и проведении конференции.

Сайт конференции `http://mmro.ru`.

ББК 22.1:32.973.26-018.2

ISBN 978-5-94588-207-2

© Авторы докладов, 2016
© ФИЦ ИУ РАН, 2016

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

Intelligent Data Processing: Theory and Applications: Book of abstracts of the 11th International Conference (Moscow, Russia – Barcelona, Spain). 2016. — Moscow: TORUS PRESS. 238 p.
ISBN 978-5-94588-207-2

The volume contains the abstracts of the 11th International Conference "Intelligent Data Processing: Theory and Applications." The conference is organized by Dorodnicyn Computing Centre FRC CSC RAS, Moscow Institute of Physics and Technology, and Harbour.Space University Barcelona. 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 Organizing Committee of IDP-2016 is grateful to the Russian Foundation for Basic Research, Forecsys Co. and CFRS Co. for providing assistance in the conference preparation and execution.

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

BBK 22.1:32.973.26-018.2

ISBN 978-5-94588-207-2

© Authors of the abstracts, 2016
© FRC CSC RAS, 2016

Международный оргкомитет

Сопредседатели: Журавлев Юрий Иванович,
академик РАН, Россия
Великанова Светлана Сергеевна, *Испания*

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

Борисова Татьяна Игоревна
Громов Андрей Николаевич
Ивахненко Андрей Александрович, *к.ф.-м.н.*
Инякин Андрей Сергеевич, *к.ф.-м.н.*
Ишкина Шаура Хабировна
Помазкова Евгения Владимировна
Ромашкова Леонтина Леонтьевна, *к.ф.-м.н.*
Татарчук Александр Игоревич, *к.ф.-м.н.*
Чехович Юлия Викторовна

Программный комитет

Сопредседатели: Рудаков Константин Владимирович,
чл.-корр. РАН, Россия
Зорин Денис Николаевич, *проф., США*

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

Воронцов Константин Вячеславович, *д.ф.-м.н.*
Гимади Эдуард Хайрутдинович, *д.ф.-м.н.*
Горнов Александр Юрьевич, *д.т.н.*
Громова Ольга Алексеевна, *д.м.н.*
Гупал Анатолий Михайлович, *д.м.н.*
Двоенко Сергей Данилович, *д.ф.-м.н.*
Дорофеюк Александр Александрович, *д.т.н.*
Кельманов Александр Васильевич, *д.ф.-м.н.*
Краснопрошин Виктор Владимирович, *д.т.н.*
Мерцалов Константин Сергеевич, *PhD*
Местецкий Леонид Моисеевич, *д.т.н.*
Моттль Вадим Вячеславович, *д.т.н.*
Осипов Геннадий Семенович, *д.ф.-м.н.*
Пытьев Юрий Петрович, *д.ф.-м.н.*
Рейер Иван Александрович, *к.т.н.*
Рязанов Владимир Васильевич, *д.ф.-м.н.*
Сойфер Виктор Александрович, *чл.-корр. РАН*
Тузиков Александр Васильевич, *д.ф.-м.н.*
Устинин Михаил Николаевич, *д.ф.-м.н.*
Чуличков Алексей Иванович, *д.ф.-м.н.*
Хачай Михаил Юрьевич, *д.ф.-м.н.*
Шананин Александр Алексеевич, *д.ф.-м.н.*

International organizing committee

Co-Chairs: Yury Zhuravlev, *acad. of RAS, Moscow*
Svetlana Velikanova, *Barcelona*

Secretary: Yury Chehovich, *C.Sc.*

Tatyana Borisova
Andrey Gromov
Andrey Inyakin, *C.Sc.*
Shaura Ishkina
Andrey Ivakhnenko, *C.Sc.*
Evgeniya Pomazkova
Leontina Romashkova
Aleksandr Tatarchuk, *C.Sc.*
Yulia Chehovich

Program Committee

Chair: Konstantin Rudakov, *corr. member of RAS, Moscow*
Denis Zorin, *Prof., New York*

Secretary: Vadim Strijov, *D.Sc.*

Aleksey Chulichkov, *D.Sc.*
Vladimir Donskoy, *D.Sc.*
Alexander Dorofeyuk, *D.Sc.*
Sergey Dvoenko, *D.Sc.*
Edward Gimadi, *D.Sc.*
Alexander Gornov, *D.Sc.*
Olga Gromova, *D.Sc.*
Anatoliy Gupal, *D.Sc.*
Alexander Kel'manov, *D.Sc.*
Michael Khachay, *D.Sc.*
Viktor Krasnoproshin, *D.Sc.*
Leonid Mestetskiy, *D.Sc.*
Konstantin Mertsalov, *PhD*
Vadim Mottl, *D.Sc.*
Gennadiy Osipov, *D.Sc.*
Yury Pytiev, *D.Sc.*
Ivan Reyer, *C.Sc.*
Vladimir Ryazanov, *D.Sc.*
Aleksandr Shananin, *D.Sc.*
Viktor Soyfer, *D.Sc.*
Alexander Tuzikov, *D.Sc.*
Mikhail Ustinin, *D.Sc.*
Konstantin Vorontsov, *D.Sc.*

Российская секция
11-й Международной конференции «Интеллектуализация обработки информации»

Оргкомитет

Председатель: Журавлев Юрий Иванович, *академик РАН*
Ученый секретарь: Чехович Юрий Викторович, *к.ф.-м.н.*

Борисова Татьяна Игоревна
Громов Андрей Николаевич
Ивахненко Андрей Александрович, *к.ф.-м.н.*
Инякин Андрей Сергеевич, *к.ф.-м.н.*
Ишкина Шаура Хабировна
Помазкова Евгения Владимировна
Татарчук Александр Игоревич, *к.ф.-м.н.*
Чехович Юлия Викторовна

Программный комитет

Председатель: Рудаков Константин Владимирович, *чл.-корр. РАН*
Ученый секретарь: Стрижов Вадим Викторович, *д.ф.-м.н.*

Воронцов Константин Вячеславович, *д.ф.-м.н.*
Гимади Эдуард Хайрутдинович, *д.ф.-м.н.*
Горнов Александр Юрьевич, *д.т.н.*
Громова Ольга Алексеевна, *д.м.н.*
Гупал Анатолий Михайлович, *д.ф.-м.н.*
Двоенко Сергей Данилович, *д.ф.-м.н.*
Дорофеюк Александр Александрович, *д.т.н.*
Кельманов Александр Васильевич, *д.ф.-м.н.*
Краснопрошин Виктор Владимирович, *д.т.н.*
Местецкий Леонид Моисеевич, *д.т.н.*
Моттль Вадим Вячеславович, *д.т.н.*
Осипов Геннадий Семенович, *д.ф.-м.н.*
Пытьев Юрий Петрович, *д.ф.-м.н.*
Рейер Иван Александрович, *к.т.н.*
Рязанов Владимир Васильевич, *д.ф.-м.н.*
Сойфер Виктор Александрович, *чл.-корр. РАН*
Тузиков Александр Васильевич, *д.ф.-м.н.*
Устинин Михаил Николаевич, *д.ф.-м.н.*
Чуличков Алексей Иванович, *д.ф.-м.н.*
Хачай Михаил Юрьевич, *д.ф.-м.н.*
Шананин Александр Алексеевич, *д.ф.-м.н.*

Russian section
of the 11-th International Conference on Intelligent Data Processing: Theory and Applications

Organizing Committee

Chair: Yury Zhuravlev, *acad. of RAS*

Secretary: Yury Chehovich, *C.Sc.*

Tatyana Borisova
Andrey Gromov
Andrey Inyakin, *C.Sc.*
Shaura Ishkina
Andrey Ivakhnenko, *C.Sc.*
Evgeniya Pomazkova
Aleksandr Tatarchuk, *C.Sc.*
Yulia Chehovich

Program Committee

Chair: Konstantin Rudakov, *corr. member of RAS*
Secretary: Vadim Strijov, *D.Sc.*

Aleksey Chulichkov, *D.Sc.*
Vladimir Donskoy, *D.Sc.*
Alexander Dorofeyuk, *D.Sc.*
Sergey Dvoenko, *D.Sc.*
Edward Gimadi, *D.Sc.*
Alexander Gornov, *D.Sc.*
Olga Gromova, *D.Sc.*
Anatoliy Gupal, *D.Sc.*
Alexander Kel'manov, *D.Sc.*
Michael Khachay, *D.Sc.*
Viktor Krasnoproshin, *D.Sc.*
Leonid Mestetskiy, *D.Sc.*
Vadim Mottl, *D.Sc.*
Gennadiy Osipov, *D.Sc.*
Yury Pytiev, *D.Sc.*
Ivan Reyer, *C.Sc.*
Vladimir Ryazanov, *D.Sc.*
Aleksandr Shananin, *D.Sc.*
Viktor Soyfer, *D.Sc.*
Alexander Tuzikov, *D.Sc.*
Mikhail Ustinin, *D.Sc.*
Konstantin Vorontsov, *D.Sc.*

Рецензенты

Адуенко А. А.
Бахтеев О. Ю.
Борисова И. А.
Гасников А. В.
Гнеушев А. Н.
Гороховский К. Ю.
Дьяконов А. Г.
Животовский Н. К.
Игнатьев В. Ю.
Инякин А. С.
Ишкина Ш. Х.
Каркищенко А. Н.
Катруца А. М.
Красоткина О. В.
Крымова Е. А.
Кудинов М. С.
Кузнецов М. П.
Кузнецова М. В.
Ланге М. М.
Майсурадзе А. И.
Максимов Ю. В.
Матвеев И. А.
Местецкий Л. М.
Мотренко А. П.
Мурашов Д. М.
Неделько В. М.
Новик В. П.
Одиноких Г. А.
Панов А. И.
Панов М. Е.
Рейер И. А.
Сенько О. В.
Середин О. С.
Скипор К. С.
Стрижов В. В.
Сулимова В. В.
Талипов К. И.
Торшин И. Ю.
Трёкин А. Н.
Турдаков Д. Ю.
Фрей А. И.
Хачай М. Ю.
Черепанов Е. В.
Чуличков А. И.

Reviewers

Aduenko A.
Bakhteev O.
Borisova I.
Cherepanov E.
Chulichkov A.
D'yakonov A.
Frei O.
Gasnikov A.
Gneushev A.
Gorokhovskiy K.
Ignat'ev V.
Inyakin A.
Ishkina Sh.
Karkishchenko A.
Katrutsa A.
Khachay M.
Krasotkina O.
Krymova E.
Kudinov M.
Kuznetsov M.
Kuznetsova M.
Lange M.
Maksimov Yu.
Matveev I.
Maysuradze A.
Mestetskiy L.
Motrenko A.
Murashov D.
Nedelko V.
Novik V.
Odinokikh G.
Panov A.
Panov M.
Reyer I.
Senko O.
Seredin O.
Skipor K.
Strijov V.
Sulimova V.
Talipov K.
Torshin I.
Trekin A.
Turdakov D.
Zhivotovskiy N.

Краткое оглавление

Brief contents

Keynote Talks

Boris Polyak

Institute for Control Science, Moscow

Robust Principal Component Analysis

The main trend of modern data analysis is to reduce huge data bases to their low-dimensional approximations. Classical tool for this purpose is Principal Component Analysis (PCA). However it is sensitive to outliers and other deviations from standard assumptions. There are numerous approaches to robust PCA. We propose two novel models. One is based on minimization of Huber-like distances from low-dimensional subspaces. Simple method for this nonconvex matrix optimization problem is proposed. The second is robust version of maximum likelihood method for covariance and location estimation for contaminated multivariate Gaussian distribution; again we arrive to nonconvex vector-matrix optimization. Both methods are based on Reweighted Least Squares Approximations. They demonstrated fast convergence in simulations, however statistical validation as well as convergence behavior of both approaches remain open problems.

Victor Lempitsky

Skolkovo Institute of Science and Technology, Moscow

Image Synthesis with Deep Neural Networks

Using deep convolutional networks for pattern recognition in images has by now become a mature and well-known technology. More recently, there is a growing interest to using convolutional networks in a "reverse" mode, i.e. to synthesize images with certain properties rather than to recognize image content. In the talk, I will present several algorithmic results and application examples obtained for this very promising direction of research.

Анализ пространства параметров в задачах выбора мультимоделей

Адуенко Александр Александрович[1]⋆ aduenko1@gmail.com
Стрижов Вадим Викторович[1,2] strijov@ccas.ru
[1]Россия, Москва, МФТИ
[2]Россия, Москва, ФИЦ ИУ РАН

Рассматривается задача выбора мультимоделей двухклассовой классификации. Мультимодели являются интерпретируемым обобщением случая одной модели, учитывающим неоднородности в данных. Признаковые пространства моделей в мультимодели могут не совпадать, а мультимодель может содержать большое число близких моделей, что ведет к низкому качеству прогноза и отсутствию интерпретируемости. Для решения этой проблемы предлагается метод статистического сравнения моделей для прореживания мультимодели. Вводится понятие адекватной мультимодели, все модели в которой попарно статистически различимы. Для статистического сравнения моделей вводится функция близости между апостериорными распределениями параметров моделей. Функция должна быть определена для пары распределений с несовпадающими носителями, а также не различать два распределения, одно из которых является малоинформативным. Предлагается функция близости для пары распределений, которая удовлетворяет этим требованиям, и доказаны асимптотические свойства ее распределения в условиях истинности гипотезы о совпадении моделей. Получены верхняя и нижняя оценки на максимальное число попарно различимых моделей в мультимодели для выборки фиксированного размера. Диагональная оценка максимума обоснованности для ковариационной матрицы весов признаков используется для отбора признаков в мультимодели. Показана асимптотическая вырожденность недиагональной оценки ковариационной матрицы [1].

Работа поддержана грантом РФФИ № 16-07-01158.

[1] *Адуенко А. А., Стрижов В. В.* Совместный выбор объектов и признаков в задачах многоклассовой классификации коллекции документов // Инфокоммуникационные технологии, 2014. № 1. С. 47–53.

Features space analysis for multimodel selection

Aduenko Aleksandr[1]⋆ aduenko1@gmail.com
Strijov Vadim[1,2] strijov@ccas.ru
[1]Moscow, Russia, MIPT
[2]Moscow, Russia, FRC CSC RAS

The problem of multimodel selection for two-class classification task has been considered. Multimodels are interpretable generalization of the single model case which addresses data inhomogeneity. Feature spaces may differ across the models. Moreover, a multimodel may contain a big number of similar models, which leads to poor forecast quality and lack of interpretability. The method of statistical model comparison is suggested to address this problem. The notion of an adequate multimodel is introduced, for which all the constituting models are pairwise statistically distinguishable. The authors suggest to introduce a similarity function between posterior distribution of model parameters for model comparison. Such a similarity function must be defined for the pair of distribution with different supports. Moreover, it must not distinguish the distributions one of which is noninformative. The similarity function which satisfies the conditions is suggested, and the asymptotic properties of its distribution are proved in case the models' true parameters are identical. The upper and the lower bounds on the maximum number of pairwise distinguishable models in a multimodel for a sample of the fixed size are obtained. Diagonal maximum evidence estimate of features' weights' covariance matrix is used for feature selection for the multimodel. Asymptotic degeneracy of nondiagonal estimate of this matrix is proved [1].

This research is funded by the Russian Foundation for Basic Research, grant 14-07-31205.

[1] Aduenko, A. A., and V. V. Strijov. 2014. Joint feature and object selection for multiclass classification of documents' collection. *Infocommunications Technol.* 1:47–53.

Выбор модели глубокого обучения субоптимальной сложности с использованием вариационной оценки правдоподобия

Бахтеев Олег Юрьевич bakhteev@phystech.edu
Москва, Россия, МФТИ

Рассматривается задача построения моделей глубокого обучения субоптимальной сложности. Под моделью понимается суперпозиция функций, решающая задачу классификации или регрессии. В качестве критерия выбора модели используется субоптимальная сложность модели. Под сложностью модели понимается минимальная длина описания, т. е. минимальное количество информации, которое требуется для передачи информации о модели и о рассматриваемых данных в совокупности. Под субоптимальной сложностью понимается вариационная оценка правдоподобия модели, т. е. оценка, полученная с использованием аппроксимации неизвестного распределения другим заданным распределением. Предлагается получение ее приближенной оценки, основанной на связи минимальной длины описания и правдоподобия модели. Для получения оценки правдоподобия используются вариационные методы.

Предлагается метод получения вариационной нижней оценки интеграла правдоподобия модели с использованием модифицированного алгоритма стохастического градиентного спуска. Рассматривается ряд модификаций базового алгоритма, а также приводится сравнение с алгоритмом получения вариационной нижней оценки, основанном на аппроксимации нормальным распределением.

Работа представленного алгоритма проилюстрирована на выборке изображений рукописных цифр, а также на искусственных данных. Эксперименты подтверждают работоспособность представленного алгоритма [1].

Работа поддержана грантом РФФИ № 16-37-00488.

[1] *Бахтеев О.Ю., Попова М.С., Стрижов В.В.* Системы и средства глубокого обучения в задачах классификации // Системы и средства информатики, 2016. Т. 26. № 2. С. 4–22.

Deep learning model selection using variatonal inference method

Bakhteev Oleg bakhteev@phystech.edu

Moscow, Russia, MIPT

The paper describes a method of suboptimal complexity model selection for deep learning nets, i. e., multilevel classification or regression superpositions of models. The maximum marginal likelihood or evidence is used as a criterion for model selection. A minimum description length is considered as a complexity of a model. Minimum description length of a model is a minimal length of the message required to compress information about the model and data. Consider variational approximation of marginal likelihood as a suboptimal complexity. Due to high computational cost, marginal likelihood is used as a value close to the complexity. In order to estimate marginal likelihood, variational inference method, i. e., a method based on replacing intractable distribution with another, was used.

The paper describes a method of variational inference estimation using a modified algorithm of stochastic gradient descent. The method is compared with the algorithm based on approximation using Gaussian distribution.

The experiments were conducted on MNIST handwritten digits dataset and simulation data. The experiments show that the proposed method can approximate marginal likelihood of the model [1].

[1] Bakhteev, O. Yu., M. S. Popova, and V. V. Strijov. 2-16. Systems and means of deep learning for classification problems. *Sistemy i Sredstva Informatiki — Systems and Means of Informatics* 26(2):4–22. Available at: http://www.ipiran.ru/journal_system/article/08696527160201.html (accessed September 28, 2016).

Бэггинг нейронных сетей в многозадачной классификации биологической активности ядерных рецепторов

Владимирова Мария Руслановна[1]⋆ mrvladimirova@gmail.com
Стрижов Вадим Викторович[2] strijov@ccas.ru
[1]Россия, Долгопрудный, МФТИ
[2]Россия, Москва, ФИЦ ИУ РАН

Рассматривается проблема многозадачной классификации на данных, описывающих взаимодействие ядерных рецепторов. Ядерные рецепторы представляют собой класс находящихся в клетках белков. Рецепторы влияют на транскрипцию генов: регулируют развитие, гомеостаз и обмен веществ в организме. Регулирование происходит в основном тогда, когда рецептор и лиганд — молекула, воздействующая на поведение рецептора, — взаимодействуют. Требуется предсказать, будет ли объект относиться к определенному классу, т. е. будет ли взаимодействовать данный лиганд с определенным рецептором.

В качестве модели классификации используется двухслойная нейронная сеть. Рассматриваются задачи линейной и логистической регрессий с квадратичной и кросс-энтропийной функциями потерь. Проводится декомпозиция функции ошибки на смещение и дисперсию. Для повышения качества предсказаний за счет уменьшения дисперсии ошибки предлагается использовать композицию двухслойных нейронных сетей — бэггинг. Бэггинг генерирует из элементов обучающей выборки семейство подвыборок того же размера с помощью процедуры бутстрэп. На каждой подвыборке настраивается классификатор. Ответы классификаторов агрегируются путем простого голосования. Предложенный метод позволяет повысить качество классификации исследуемой выборки [1].

Работа поддержана грантом РФФИ № 16-07-01155.

[1] *Владимирова М. Р.* Бэггинг нейронных сетей // Машинное обучение и анализ данных, 2016 (в печати).

Bagging of neural networks in multitask classification of biological acivity for nuclear receptors

Vladimirova Maria[1]⋆ mrvladimirova@gmail.com
Strijov Vadim[2] strijov@ccas.ru
[1]Dolgoprudniy, Russia, MIPT
[2]Moscow, Russia, FRC CSC RAS

The paper is devoted to the multitask classification problem. The main purpose is building an adequate model to predict whether the object belongs to a particular class, precisely, whether the ligand binds to a specific nuclear receptor. Nuclear receptors are a class of proteins found within cells. These receptors work with other proteins to regulate the expression of specific genes, thereby controlling the development, homeostasis, and metabolism of the organism. The regulation of gene expression generally only happens when a ligand — a molecule that affects the receptor's behavior — binds to a nuclear receptor.

Two-layer neural network is used as a classification model. The paper considers the problems of linear and logistic regressions with squared and cross-entropy loss functions. To analyze the classification result, the authors propose to decompose the error into bias and variance terms. To improve the quality of classification by reducing the error variance, the authors suggest the composition of neural networks — bagging. Bagging generates a set of subsamples from the training sample using the bootstrap procedure. All subsamples have the same size as initial sample. Classifiers are trained on each subsample separately. Then their individual predictions are aggregated by voting. The proposed method improves the quality of investigated sample classification [1].

This research is funded by the Russian Foundation for Basic Research, grant 16-07-01155.

[1] Vladimirova, M. 2016 (in press). Bagging of neural networks. *Machine Learning Data Anal.*

О полных регрессионных решающих деревьях

Генрихов Игорь Евгеньевич[1] ingvar1485@rambler.ru
Дюкова Елена Всеволодовна[2] edjukova@mail.ru
Журавлёв Вадим Игоревич[3]⋆ vadim091294@gmail.com
[1]Россия, Химки, ООО «Мобайл парк ИТ»
[2]Россия, Москва, ФИЦ ИУ РАН
[3]Россия, Москва, МГУ

Рассматривается одна из актуальных задач машинного обучения — задача восстановления регрессии. По постановке данная задача близка к задаче классификации по прецедентам и отличается от последней видом целевой функции. Для решения обеих задач применяется, в частности, аппарат решающих деревьев (РД). Наиболее известные алгоритмы восстановления регрессии (например, алгоритмы CART и Random Forest) основаны на использовании бинарных РД. Реже используются k-арные РД.

Очевидным недостатком классической модели РД является то, что для построения очередной вершины дерева среди всех признаков, удовлетворяющих критерию ветвления, выбирается только один признак и выбирается этот признак фактически случайным образом. В работах И. Е. Генрихова, Е. В. Дюковой и Н. В. Пескова для задачи классификации разработана качественно новая модель РД, а именно: модель полного РД (ПРД), лишенная указанного недостатка. Конструкция ПРД позволяет более существенно использовать имеющуюся информацию, при этом описание распознаваемого объекта порождается не одной ветвью, как в классическом РД, а несколькими ветвями. Каждая такая ветвь участвует в процедуре голосования. В зарубежных и отечественных публикациях нет сведений о применении деревьев типа ПРД для задачи восстановления регрессии, поэтому представляет интерес проведенные в [1] исследования по разработке регрессионных ПРД.

Работа поддержана грантом РФФИ № 16-01-00445.

[1] *Генрихов И. Е., Дюкова Е. В., Журавлёв В. И.* О полных регрессионных решающих деревьях // Машинное обучение и анализ данных, 2016 (в печати).

About full regressive decision trees

Genrikhov Igor[1] ingvar1485@rambler.ru
Djukova Elena[2] edjukova@mail.ru
Zhuravlyov Vadim[3]⋆ vadim091294@gmail.com
[1]Khimki, Russia, LLC Mobail park IT
[2]Moscow, Russia, FRC CSC RAS
[3]Moscow, Russia, MSU

The regression restoration problem as one of the actual tasks of machine learning is considered. This problem is similar to the problem of classification by precedents and differs by the type of target function. The instrument of decision trees is used to solve both problems. The most well-known algorithms for regression restoration (e. g., CART and Random Forest algorithms) are based on the use of the basic decision trees (DT), namely, the binary DT. Rarely k-ary DT are used.

The evident disadvantage of the classical model of the DT is that to build another node among all features, which satisfy the criterion of branching, only one feature can be selected. Moreover, this feature is selected randomly. Qualitatively new DT model was developed for the classification problem in the works of Genrikhov I. E., Djukova E. V., and Peskov N. V. Namely, it is the model of the full DT (FDT), which is lacking this disadvantage. The FDT construction allows to use available information more significantly; herewith, the description of the recognizable object is generated not by a single branch, as in the classical DT, but by several branches. Each branch is involved in the voting procedure. There is no information on the use of the FDT-type trees for the regression restoration problems in the foreign and domestic publications. Therefore, it is of interest to investigate development of regressive FDT. The results of such investigations are presented in [1].

This research is funded by the Russian Foundation for Basic Research, grant 16-01-00445.

[1] Genrikhov, I. E., E. V. Djukova, and V. I. Zhuravlyov. 2016 (in press). About full regressive decision trees. *Machine Learning Data Anal.*

Группировка признаков на основе оптимальной последовательности миноров корреляционной матрицы

Двоенко Сергей Данилович dsd@tsu.tula.ru
Пшеничный Денис Олегович* denishshenichny@yandex.ru
Россия, Тула, ТулГУ

При решении задачи группировки возникает проблема содержательной интерпретации полученных факторов и групп признаков. Тем не менее, факторы групп является синтетическими признаками, интерпретация которых может быть затруднена. Поэтому после выделения групп признаков и построения соответствующих им факторов в каждой группе обычно определяется ее представитель как наиболее сильно коррелирующий с фактором группы признак. Тогда оказывается возможным содержательно интерпретировать результат группировки прямо в терминах исходных признаков.

Предложен новый подход для выбора подмножества признаков, способных адекватно представить скрытые факторы, без определения собственных или центроидных направлений в качестве промежуточных преобразований. Данный подход основан на построении оптимальной последовательности значений главных миноров корреляционной матрицы признаков. В начале такой оптимальной последовательности расположены наименее коррелированные друг с другом и с остальными признаки, а к ее концу выстраиваются все более коррелированные с остальными признаки, выбранные в последнюю очередь.

Показано, что предложенный подход позволяет формировать начальное решение для других алгоритмов группировки и также может применяться самостоятельно для оценки числа групп и построения содержательных группировок [1].

Работа частично поддержана грантами РФФИ №№ 15-07-02228, 15-07-08967, 14-07-00527 и 14-07-00964.

[1] *Двоенко С. Д., Пшеничный Д. О.* Группировка признаков на основе оптимальной последовательности миноров корреляционной матрицы // Машинное обучение и анализ данных, 2016 (в печати).

Feature grouping based on the optimal sequence of correlation matrix minors

Dvoenko Sergey dsd@tsu.tula.ru
Pshenichny Denis⋆ myhoangthanh@yahoo.com
Tula, Russia, TulSU

It is known that data analysis problems usually arise in early stages of investigations, when a model of a phenomenon in researching has not been developed yet. Hence, it is too early to introduce a problem of a model identification. It needs to collect and study a lot of miscellaneous information about most significant characteristics of a phenomenon under investigation in this case. Such a situation forces us to use inconsistent approach, since it is not known what characteristics are important and what knowledge needs to be collected. Therefore, data analysis methods must resolve the contradiction and focus on the correct description of the phenomenon.

The problem of informal interpretation of factors and groups arises in the grouping problem. Factors are synthetic features, and difficulties can arise in informal interpretation of them. Therefore, after groups and corresponding factors have been built, the representative usually is defined for each group as a feature the most correlated with the group factor. As a result, it is possible to name groups informally as such initial features. The new approach to specify a feature subset is proposed to represent correctly the hidden factors. In this approach, it does not need to define eigenvectors or centroid ones as intermediate transformations. It is based on the optimal sequence of correlation matrix minors, since the less correlated features are placed at the beginning of the sequence, and the more correlated ones are placed closer to the end of it.

The proposed approach can produce initial partitioning for other grouping algorithms and additionally can be used to evaluate a number of groups and to get informal partitions.

This research is funded partially by the Russian Foundation for Basic Research, grants 15-07-02228, 15-07-08967, 14-07-00527, and 14-07-00964.

[1] Dvoenko, S. D., and D. O. Pshenichny. 2016 (in press). Feature grouping based on the optimal sequence of correlation matrix minors. *J. Machine Learning Data Anal.*

Методы построения хорошо интерпретируемых классификаций

Дорофеюк Александр Александрович* daa2@mail.ru
Дорофеюк Юлия Александровна dorofeyuk_julia@mail.ru
Покровская Ирина Вячеславовна ivp750@mail.ru
Чернявский Александр Леонидович achern@ipu.ru
Россия, Москва, ИПУ РАН

Описаны три алгоритма построения содержательно хорошо интерпретируемых классификаций — покоординатная, спрямляющая и содержательно-экспертная. Для алгоритмов покоординатной и спрямляющей классификации множество значений каждого показателя разбивается на небольшое число диапазонов, при этом классификация фактически задается набором границ этих диапазонов. Покоординатная классификация осуществляет для каждого показателя независимое от других разбиение на заданное число диапазонов, т. е. искомая классификация задается сочетанием одномерных классификаций по каждому из исходных показателей. Основное преимущество такой классификации — возможность использования алгоритмов глобально-оптимальной одномерной классификации. Спрямляющая классификация строится как своего рода аппроксимация многомерной классификации. Приводится сравнение покоординатной и спрямляющей классификаций. Содержательно-экспертная классификация используется тогда, когда каждый из имеющегося множества объектов, подлежащих классификации, задан только своим названием и содержательным описанием. Кроме того, экспертным путем задана так называемая «содержательная» классификация, которая определяется названиями и содержательным описанием классов. Требуется по этой информации о заданном множестве объектов провести их классификацию [1]. Работа выполнена при частичной финансовой поддержке РФФИ, проекты №№ 14-07-00463-а, 15-07-06713-а, 16-07-00895-а и 16-07-00896-а.

[1] *Чернявский А. Л., Дорофеюк А. А., Гольдовская М. Д.* Методы экспертизы в задаче построения хорошо интерпретируемых классификаций // Управление развитием крупномасштабных систем. — М.: ИПУ РАН, 2011. Т. 1. С. 331–337.

Well-interpreted classifications constructing methods

Dorofeyuk Alexander ⋆ daa2@mail.ru
Dorofeyuk Yuliya dorofeyuk_julia@mail.ru
Pokrovskaya Irina ivp750@mail.ru
Chernyavskiy Aleksandr achern@ipu.ru
Moscow, Russia, ICS RAS

This paper describes three algorithms for building a meaningful well-interpreted classifications — coordinates, rectifiable, and content-expert. For algorithms of the coordinates and rectifiable classification, the values set of each indicator is divided into a small number of ranges. The wisecoordinate classifications are actually defined by the set of boundaries of these ranges. The coordinatewise classification provides for each indicator independent spanning on the specified number of ranges, i. e., the required classification is set by the combination of one-dimensional (1D) classifications for each of the benchmarks. The main advantage of such classification is the ability to use algorithms for globally optimal 1D classification. The rectifiable classification is constructed as a kind of a multidimensional classification approximation. A comparison of the coordinates and rectifiable classifications has been done. Content-expert classification is used when each of the available objects set to be classified specify only its name and a meaningful description. Furthermore, the expert is specified by the so-called "meaningful" classification, which is determined by titles and content description of the classes. For this information, it is required to classify the objects set [1].

This research is executed at partial financial support of the Russian Foundation for Basic Research, grants Nos. 14-07-00463-a, 15-07-06713-a, 16-07-00895-a, and 16-07-00896-a.

[1] Chernyavskiy, A. L., A. A. Dorofeyuk, and M. D. Goldovskaya. 2013. Methods of examination in the objective of building a well-interpreted classifications. *Management of large-scale systems development*. Moscow: ICS RAS. 1:331–337.

Применение обучения с подкреплением для одновременного выбора модели алгоритма классификации и ее структурных параметров

Ефимова Валерия Александровна*⋆* efimova@rain.ifmo.ru
Фильченков Андрей Александрович afilchenkov@corp.ifmo.ru
Шалыто Анатолий Абрамович shalyto@mail.ifmo.ru
Россия, Санкт-Петербург, Университет ИТМО

Для решения задач классификации разработано множество алгоритмов. Для обработки конкретного набора данных производится выбор модели алгоритма и настройка ее структурных параметров. Решение задачи данной работы существует только для частных случаев или осуществляется полным перебором, реализованным в библиотеке Auto-WEKA. Цель данной работы — предложить метод одновременного выбора модели алгоритма классификации и ее структурных параметров.

Предложенный метод основан на сведении к задаче о многоруком бандите. Ручкам соответствуют модели алгоритмов. Выбору ручки — запуск процесса настройки структурных параметров в течение некоторого отрезка времени. Предложены две функции награды.

Эксперименты проводились на десяти наборах реальных данных, в них использовались шесть известных моделей алгоритмов классификации, для оптимизации структурных параметров запускался алгоритм SMAC. Использовались следующие алгоритмы решения задачи о многоруком бандите: ε-жадный, UCB1 и Softmax. Результаты показывают, что предложенный метод позволяет найти решение, не уступающее решению, найденному библиотекой Auto-WEKA, а в большинстве случаев и превосходящее его [1].

Работа выполнена при поддержке Правительства Российской Федерации, грант 074-U01 и РФФИ, проект № 16-37-60115.

[1] *Ефимова В. А., Фильченков А. А., Шалыто А. А.* Применение обучения с подкреплением для одновременного выбора модели алгоритма классификации и ее структурных параметров // Машинное обучение и анализ данных, 2016 (в печати).

Reinforcement-based simultaneous classification model and its hyperparameters selection

Efimova Valeria* efimova@rain.ifmo.ru
Filchenkov Andrey afilchenkov@corp.ifmo.ru
Shalyto Anatoly shalyto@mail.ifmo.ru
St. Petersburg, Russia, ITMO University

Many algorithms for data analysis exist, especially for classification problem. To solve a data analysis problem, a proper algorithm should be chosen and also, its hyperparameters should be selected. These two problems, algorithm selection and hyperparameter optimization, are commonly solved independently. The full-model selection process requires unacceptable time budgets. Thus, this is one of the factors preventing the spread of automated model selection methods.

The goal of this work is to suggest a method for simultaneous algorithm and its parameters selection to reduce full-model selection time. In order to do so, this problem was produced to a multiarmed bandit problem. An algorithm was considered as an arm, and algorithm for hyperparameters search during a fixed time was considered as the corresponding arm play. Also, several reward functions were described.

The experiments were held on ten popular labeled datasets from the UCI repository. To compare the proposed method, several well-known classification algorithms were used from WEKA library and SMAC algorithm for hyperparameter optimization was used from Auto-WEKA library. The proposed method was compared with the brute force search implemented in WEKA library and a random time budget assignment policy. The results show significant time reduction of selecting proper algorithm and its hyperparameters for processing the given dataset. The proposed method often produces classification results much better than Auto-WEKA state-of-the-art automatic algorithm selection and hyperparameter optimization tool [1].

The research was supported by the Government of the Russian Federation (grant 074-U01) and the Russian Foundation for Basic Research (project no. 16-37-60115).

[1] Efimova, V., A. Filchenkov, and A. Shalyto. 2016 (in press). Reinforcement-based simultaneous classification model and its hyperparameters selection. *Machine Learning Data Anal.*

Классификация демографических последовательностей на основе узорных структур

Игнатов Дмитрий Игоревич[1]⋆ dignatov@hse.ru
Гиздатуллин Данил Кутдусович[1] gizdatullindanil@gmail.com
Митрофанова Екатерина Сергеевна[1] emitrofanova@hse.ru
Муратова Анна Александровна[1] anyamuratova@yandex.ru
Башерье Жауме[2] jbaixer@lsi.upc.edu
[1]Россия, Москва, НИУ ВШЭ
[2]Испания, Барселона, Политехнический университет Каталонии

Представлены результаты первых экспериментов применения узорных структур на последовательностях к анализу демографических данных в России. Использованы данные об 11 поколениях с 1930 по 1984 гг. для панели из трех волн, имевших место в 2004, 2007 и 2011 гг. Основная задача состояла в поиске таких закономерностей, которые являются (замкнутыми) частыми префиксами без «разрывов». Эти ограничения — естественное требование демографов, необходимое для изучения первых событий на этапе взросления. Для решения этой задачи использованы узорные структуры неразрывных последовательностей и модифицированные FP-деревья (Frequent Pattern Trees). Наилучшие результаты в терминах TPR-FPR (True Positive Rate – False Positive Rate) были получены при больших значениях параметра роста (с некоторым числом отказов от классификации) [1].

Статья подготовлена в ходе проведения исследования № 16-05-0011 «Разработка и апробация методик анализа демографических последовательностей» в рамках Программы «Научный фонд Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ)» в 2016 г. и с использованием средств субсидии на государственную поддержку ведущих университетов Российской Федерации в целях повышения их конкурентоспособности среди ведущих мировых научно-образовательных центров, выделенной НИУ ВШЭ.

[1] *Ignatov D., Mitrofanova E., Muratova A., Gizdatullin D.* Pattern mining and machine learning for demographic sequences // Knowledge Engineering and Semantic Web: 6th Conference (International) Proceedings. — Cham: Springer, 2015. P. 225–239. http://dx.doi.org/10.1007/978-3-319-24543-0_17.

Pattern-based classification of demographic sequences

Ignatov Dmitry[1]⋆ dignatov@hse.ru
Gizdatullin Danil[1] gizdatullindanil@gmail.com
Mitrofanova Ekaterina[1] emitrofanova@hse.ru
Muratova Anna[1] anyamuratova@yandex.ru
Baixeries Jaume[2] jbaixer@lsi.upc.edu
[1]Moscow, Russia, HSE
[2]Barcelona, Spain, Universitat Politècnica de Catalunya

This paper presents the first results of studies in application of sequence-based pattern structures and emerging patterns to analysis of demographic sequences in Russia. This study is performed on data of 11 generations from 1930 till 1984 for the panel of three waves of the Russian part of Generation and Gender Survey, which took place in 2004, 2007, and 2011. The main goal is to develop methods of extracting emerging patterns (EP) with the following restrictions: the obtained patterns need to be (closed) frequent gapless prefixes of the input sequences. These constraints were required by demographers since it is necessary for proper interpretation and understanding of early life course events that lead to adulthood. To solve this problem, pattern structures of gapless prefixes and modified FP-trees have been used. After extraction of EP, CAEP classifier was used to predict gender of respondents with their demographic sequences of the first life course events. The best results in terms of TPR-FPR have been obtained for large values of minimum growth-rate parameter (with some objects left without classification) [1]. The paper was prepared within the framework of the Academic Fund Program at HSE in 2016 (grant 16-05-0011 "Development and testing of demographic sequence analysis and mining techniques") and supported within the framework of a subsidy granted to the HSE by the Government of the Russian Federation for the implementation of the Global Competitiveness Program.

[1] Ignatov, D., E. Mitrofanova, A. Muratova, and D. Gizdatullin. 2015. Pattern mining and machine learning for demographic sequences. *Knowledge Engineering and Semantic Web: 6th Conference (International) Proceedings*. Cham: Springer. 225–239. Available at: http://dx.doi.org/10.1007/978-3-319-24543-0_17 (accessed September 29, 2016).

Аппроксимация комбинаторных оценок переобучения пороговых классификаторов

Ишкина Шаура Хабировна shaura-ishkina@yandex.ru
Москва, ФИЦ ИУ РАН

Повышение точности оценок обобщающей способности, зависящих от данных, остается одной из открытых задач теории статистического обучения. Комбинаторная теория переобучения позволяет получать неасимптотические и неулучшаемые оценки для некоторых искусственных частных случаев семейств классификаторов. Проблема заключается как в расширении класса реальных задач классификации, для которых применимы комбинаторные оценки, так и в сокращении их вычислительной сложности. Вычисление оценок непосредственно по определению имеет экспоненциальную сложность по объему выборки L.

В данной работе рассматривается семейство пороговых классификаторов над одномерными признаками. Оценки обобщающей способности вычисляются путем подсчета числа траекторий случайного блуждания по трехмерной сетке с ограничениями. Алгоритм их комбинаторного вычисления имеет полиномиальную сложность $O(L^5)$, что затрудняет его практическое применение. Для решения этой проблемы применяется суррогатное моделирование: точные оценки вычисляются по случайно порожденным выборкам данных, затем строится аппроксимирующая модель. Признаками в этой модели являются различные геометрические свойства семейства классификаторов как набора бинарных векторов ошибок. Полученные аппроксимации обобщающей способности используются в качестве критерия отбора признаков в линейном наивном байесовском классификаторе.

Эксперименты в задаче медицинской диагностики показывают, что применение приближенных оценок обобщающей способности приводит к сокращению числа признаков и повышению качества классификации [1].

Работа поддержана грантом РФФИ № 14-07-00908.

[1] *Ишкина Ш. Х.* Аппроксимация комбинаторных оценок переобучения пороговых классификаторов // Машинное обучение и анализ данных, 2016 (в печати).

Approximation of combinatorial generalization bounds for threshold classifiers

Ishkina Shaura shaura-ishkina@yandex.ru

Moscow, Russia, FRC CSC RAS

Tightening data dependent generalization bounds remains one of the open problems in statistical learning theory. Combinatorial theory of overfitting gives unimprovable nonasymptotic bounds for some special families of classifiers. The problem is to extend the combinatorial techniques to the more realistic classification tasks as well as to reduce the complexity of bounds computation. The complexity of combinatorial generalization bounds if computed directly by definition is commonly exponential in the size L of the dataset.

In this work, a family of one-dimensional threshold classifiers is considered. In this case, exact generalization bounds can be calculated by counting the number of random walks on a three-dimensional grid with restrictions. The algorithm of bound computation has polynomial complexity $O(L^5)$ which hinders its practical application. The surrogate modeling is used to tackle this problem: first, the exact bounds are calculated on the randomly generated datasets and then, one learns an approximation function via parametric nonlinear regression. A set of regressors that describe geometric properties of the family of threshold classifiers considered as a sequence of binary error vectors has been developed. Finally, the approximated generalization bound is used as a feature selection criterion in linear Naïve Bayes classifier.

The experiments on a medical diagnostics problem show that the approximated bounds help to reduce the number of features and to improve the classification quality [1].

This research is funded by the Russian Foundation for Basic Research, grant 14-07-00908.

[1] Ishkina, Sh. 2016 (in press). Approximation of combinatorial generalization bounds for threshold classifiers. *Machine Learning Data Anal.*

Информационный критерий для сравнения классификаторов на ансамбле источников

Ланге Михаил Михайлович lange_mm@ccas.ru

Россия, Москва, ФИЦ ИУ РАН

Исследуются MV и GM классификаторы на основе голосования решений по отдельным источникам (Majority Voting) и на основе обобщенной меры для ансамбля источников (General Measure). Пусть $\Omega = \{\omega_1, \ldots, \omega_c\}$, $c \geqslant 2$, — множество классов с априорными вероятностями $p(\omega_i) > 0$, где $\mathbf{X}^M = (X_1, \ldots, X_M)$ — ансамбль источников; $X_m = \{\mathbf{x}_m = (x_{m1}, \ldots, x_{mV_m})\}$, $m = 1, \ldots, M$, — множество N_m-мерных векторов (объектов) m-го источника; $\mathbf{x}^M = (\mathbf{x}_1, \ldots, \mathbf{x}_M) \in \mathbf{X}^M$ — объект ансамбля. Классификатор объектов из множества X_m выполняет преобразование $F_m : X_m \to \Omega$; классификатор объектов из ансамбля $\mathbf{X}^M$ — преобразование $F^M : \mathbf{X}^M \to \Omega$. На множествах X_m заданы меры различия $d(\mathbf{x}_m, \hat{\mathbf{x}}_m) = \sum_{n=1}^{N_m} (x_{mn} - \hat{x}_{mn})^2/\sigma_{mn}^2$ с параметрами $0 < \sigma_{mn}^2 < \infty$, $n = 1, \ldots, N_m$; на ансамбле $\mathbf{X}^M$ — обобщенная мера различия $D(\mathbf{x}^M, \hat{\mathbf{x}}^M) = \sum_{m=1}^{M} w_m d(\mathbf{x}_m, \hat{\mathbf{x}}_m)$ с весами $W = \{w_m > 0\}$. Обобщенная мера является обобщением многокомпонентной меры, используемой в [1]. Используя априорное распределение классов и указанные меры, введены функционалы средней взаимной информации $I(X_m; \Omega)$ и $I_W(\mathbf{X}^M; \Omega)$, соответствующие преобразованиям F_m и F^M. Предлагаемый критерий сформулирован в виде неравенства

$$\max_W \sum_{m=1}^{c} I(X_m; \Omega) \frac{w_m}{\sum_{m=1}^{c} w_m} \leqslant \frac{1}{M} I_{W^*}(\mathbf{X}^M; \Omega),$$

где левая часть равна наибольшей средней взаимной информации MV классификатора, достигаемой при весах W^*, а правая — средней взаимной информации GM классификатора по обобщенной мере с весами W^*.

Работа поддержана РФФИ, проекты 15-01-04671 и 15-07-09324.

[1] *Ланге М. М., Ганебных С.Н., Ланге А.М.* Многоклассовое распознавание образов в пространстве представлений с многоуровневым разрешением // Машинное обучение и анализ данных, 2016. Т. 2. №1. С. 70–88. jmlda.org/papers/doc/2016/no1/Lange2016Recognition.pdf.

Information criterion for comparison of metric classifiers in ensemble of sources

Lange Mikhail lange_mm@ccas.ru

Moscow, Russia, FRC CSC RAS

Given ensemble of sources, MV classifier based on majority voting decisions about individual sources and GM classifier based on a general dissimilarity measure in the ensemble are investigated. Let $\Omega = \{\omega_1, \ldots, \omega_c\}$, $c \geqslant 2$, be a set of classes with *a priori* probabilities $p(\omega_i) > 0$, and $\mathbf{X}^M = (X_1, \ldots, X_M)$ be an ensemble of sources, where $X_m = \{\mathbf{x}_m = (x_{m1}, \ldots, x_{mV_m})\}$, $m = 1, \ldots, M$, is a set of N_m-dimensional vectors (objects) of the mth source, and $\mathbf{x}^M = (\mathbf{x}_1, \ldots, \mathbf{x}_M) \in \mathbf{X}^M$ is a composite object of the ensemble. A classifier of the objects from the set X_m and a classifier of the composite objects from the ensemble $\mathbf{X}^M$ produce the corresponding transformations $F_m : X_m \to \Omega$ and $F^M : \mathbf{X}^M \to \Omega$. A dissimilarity measure $d(\mathbf{x}_m, \hat{\mathbf{x}}_m) = \sum_{n=1}^{N_m} (x_{mn} - \hat{x}_{mn})^2 / \sigma_{mn}^2$ with the parameters $0 < \sigma_{mn}^2 < \infty$, $n = 1, \ldots, N_m$, is given in each set X_m, and a general dissimilarity measure $D(\mathbf{x}^M, \hat{\mathbf{x}}^M) = \sum_{m=1}^{M} w_m d(\mathbf{x}_m, \hat{\mathbf{x}}_m)$ with the weights $W = \{w_m > 0\}$ is produced in the ensemble $\mathbf{X}^M$. The general measure is a generalization of a multicomponent measure that is used in [1]. Using the a priori probabilities and the above measures, the functions of average mutual information $I(X_m; \Omega)$ and $I_W(\mathbf{X}^M; \Omega)$ for F_m and F^M transformations are defined. The suggested criterion is formed by the inequality

$$\max_W \sum_{m=1}^{c} I(X_m; \Omega) \frac{w_m}{\sum_{m=1}^{c} w_m} \leqslant \frac{1}{M} I_{W^*}(\mathbf{X}^M; \Omega)$$

where the maximum of the average mutual information of the MV classifier (left part) is provided by the weights W^* and the average mutual information per one source of the GM classifier (right part) is taken at the same W^*.

Research is funded by the Russian Foundation for Basic Research, grants 15-01-04671 and 15-07-09324.

[1] Lange, M. M., S. N. Ganebnykh, and A. M. Lange. 2016. Multiclass pattern recognition in a space of multiresolution representations. *J. Mach. Learn. Data Anal.* 2(1):70–88. Available at: jmlda.org/papers/doc/2016/no1/Lange2016Recognition.pdf (accessed August 28, 2016).

Оценка объема выборки в задачах классификации

Мотренко Анастасия Петровна

anastasya.motrenko@phystech.edu

Россия, Долгопрудный, МФТИ

Задача оценки объема выборки, возникающая при планировании эксперимента, состоит в оценке минимального количества измерений некоторого параметра или набора параметров, требуемых для выполнения некоторых заранее сформулированных условий. Выбор метода оценки объема выборки определяется формулировкой этих условий. Предлагается метод оценки объема выборки, объединяющий частотный и байесовский подходы к оценке объема выборки. Предполагается, что на этапе сбора данных зафиксировано несколько возможных моделей. Предлагается метод оценки объема выборки, гарантирующий, что оптимальная модель будет выбрана с достаточно высокой вероятностью. Предлагаемый способ оценки объема выборки основан на сравнении оценок функции распределения данных, полученных на различных подвыборках исследуемой выборки. На основе получаемых оценок достаточного объема выборки предлагается выбирать оптимальный подход к построению модели классификации. Данный подход является оригинальным и предлагается впервые. Приводятся рекомендации по использованию предлагаемого метода для оценки достаточного объема выборки при наличии избыточного объема данных и по корректировке получаемых оценок при поступлении измерений [1].

Работа поддержана грантом РФФИ № 16-37-00111.

[1] *Motrenko A., Strijov V., Weber G.-W.* Bayesian sample size estimation for logistic regression // J. Comput. Appl. Math., 2014. Vol. 255. P. 743–752.

Sample size estimation in classification problems

Motrenko Anastasia anastasiya.motrenko@phystech.edu
Dolgopudny, Russia, MIPT

The problem of sample size determination, which is an important step in the design of experiments, involves estimation of the minimal number of observations, required to satisfy some predefined criterion. The choice of criterion defines the method of sample size determination. A new method of sample size determination, which combines frequentist and bayesian points of view, has been suggested. It was assumed that several models were fixed before the data were collected. The proposed method guarantees that the optimal will be selected with sufficient probability. To verify this condition, the probability distribution functions of model parameters, evaluated at different subsamples, have been compared. The estimates of sample size for each model were then used to select the optimal model. It is discussed how the proposed method can be used is cases where (*i*) no data are available before the study; (*ii*) an excessively large sample is available; and (*iii*) a medium-size (insufficient) sample is available [1].

This research is funded by the Russian Foundation for Basic Research, grant 16-37-00111.

[1] Motrenko, A., V. Strijov, and G.-W. Weber. 2014. Bayesian sample size estimation for logistic regression. *J. Comput. Appl. Math.* 255:743–752.

Сокращение размерности признакового пространства на основе критерия разделимости классов

Немирко Анатолий Павлович apn-bs@yandex.ru
Россия, Санкт-Петербург, СПбГЭТУ «ЛЭТИ»

Рассматривается линейное преобразование данных в многомерном признаковом пространстве на основе критерия Фишера. Дискриминантный анализ Фишера позволяет сократить размерность исходного пространства и применить различные методы распознавания в полученном пространстве меньшей размерности. Введение добавочных весовых векторов позволяет более точно представить взаимное положение классов в результирующем пространстве. Предложена процедура для рекуррентного вычисления весовых векторов, которые образуют новые признаки. Для линейной классификации в сокращенном пространстве признаков рассмотрена мера близости классов на основе оценки минимального расстояния между выпуклыми оболочками классов. Предложен алгоритм вычисления такого расстояния для двумерного пространства. На примерах показано, что за счет более точного представления данных с помощью найденных признаков они позволяют обнаружить линейную разделимость классов, которую методами главных компонент и классического линейного дискриминанта Фишера обнаружить не удается [1]. Результаты использованы в задачах автоматического анализа биомедицинских данных.

Работа поддержана грантами РФФИ 15-07-01790 и 16-01-00159 и медицинским проектом CardioQVARK — кардиограмма с помощью телефона (www.cardioqvark.ru).

[1] *Nemirko A. P.* Transformation of feature space based on Fisher's linear discriminant // Pattern Recogn. Image Anal., 2016. Vol. 26. No. 2. P. 257–261.

Reduction of feature space dimension based on separability criterion

Nemirko Anatolii apn-bs@yandex.ru
Saint Petersburg, Russia, ETU

Linear transformation of data in multidimensional feature space based on Fisher's criterion is considered. Fischer's discriminant analysis enables to reduce the dimension of initial space and to apply various methods of recognition in the derived space of smaller dimension. Introduction of additional weight vectors allows to represent the mutual location of classes in the resulting space more precisely. Procedure for recurrent calculation of weight vectors which form new features is suggested. For linear classification in the reduced space of features, the measure of proximity of classes is considered. It is based on the assessment of the minimum distance between convex covers of classes. The algorithm for calculation of such distance for two-dimensional space is proposed. The considered examples show that the newly found features which represent the data more accurately make it possible to achieve linear separability of classes which remains impossible using the technique of principal components and the classic Fisher's linear discriminant [1]. The obtained results are used in the tasks of the automatic analysis of biomedical data.

This research is funded by the Russian Foundation for Basic Research, grants Nos. 15-07-01790 and 16-01-00159, and by medical project CardioQVARK — Cardiogram by Phone (www.cardioqvark.ru).

[1] Nemirko, A. 2016. Transformation of feature space based on Fisher's linear discriminant. *Pattern Recogn. Image Anal.* 26(2):257–261.

Решающие правила для ансамбля из цепей вероятностных классификаторов при решении задач классификации с пересекающимися классами

Остапец Андрей Александрович aostapec@mail.ru
Россия, Москва, МГУ

Рассматривается задача классификации с пересекающимися классами. Исследовано применение ансамбля из цепей вероятностных классификаторов с использованием основных типов решающих правил для формирования итоговых предсказаний.

Схема решения рассматривается с точки зрения алгебраического подхода. Алгебраический подход заключается в представлении алгоритма решения задачи в виде суперпозиции двух алгоритмов. На первом этапе строится первый алгоритм (распознающий оператор), который в качестве ответа выдает вектор оценок принадлежности к каждому из классов. В качестве распознающих операторов рассматриваются следующие семейства алгоритмов: линейные классификаторы (базовые классификаторы), цепь вероятностных классификаторов из линейных классификаторов и ансамбль из цепей вероятностных классификаторов. На следующем этапе второй алгоритм (решающее правило) трансформирует этот вектор оценок в финальный ответ. Приведен обзор основных типов решающих правил, и исследовано их применение для различных распознающих операторов.

Экспериментально показана возможность эффективного использования решающих правил, построенных над результатами прогнозов базовых классификаторов [1].

[1] *Остапец А. А.* Решающие правила для ансамбля из цепей вероятностных классификаторов при решении задач классификации с пересекающимися классами // Машинное обучение и анализ данных, 2016 (в печати).

Decision rules for ensembled probabilistic classifier chain for multilabel classification

Ostapets Andrey aostapec@mail.ru

Moscow, Russia, MSU

This work considers using of the main types of decision rules for the multilabel classification task.

The algorithm is presented as a superposition of two algorithms: a recognition operator and a decision rule. The recognition operator converts feature vectors of objects to be recognized into scores for each class. This work considers several families of algorithms to be the recognition operator: linear models (base classifiers), probabilistic classifier chain of linear models, and ensembled probabilistic classifier chain. The decision rule converts the scores into final answers. In this survey, main types of decision rules are described and their performance for several recognition operators is also shown.

It is experimentally demonstrated that the quality of the forecast of the proposed composition exceeds the quality of the base classifiers [1].

[1] Ostapets, A. 2016 (in press). Decision rules for ensembled probabilistic classifier chain for multilabel classification. *J. Machine Learning Data Anal.*

О взаимосвязи мер кластеров и распределений расстояний в компактных метрических пространствах

Пушняков Алексей Сергеевич pushnyakovalex@mail.ru
Долгопрудный, Россия, МФТИ

Рассматривается компактное метрическое пространство (X, ρ) с ограниченной борелевской мерой μ. Под *r-кластером* понимается любое измеримое множество диаметра не более r. Набор, состоящий из k $2r$-кластеров, назовем *r-кластерной структурой порядка k*, если любые два кластера набора отделены на расстояние не менее r. Под мерой кластерной структуры понимается суммарная мера кластеров, входящих в нее. Используя теорему Бляшке, можно показать, что среди всех r-кластерных структур порядка k найдется структура максимальной меры $\mathcal{X}^*$. Исследуется зависимость величины $\mu(\mathcal{X}^*)$ от распределения расстояний. Основной вопрос состоит в следующем: какие ограничения нужно наложить на распределение расстояний, чтобы $\mu(\mathcal{X}^*)$ была *близка* к $\mu(X)$? Предлагается следующая дискретизация распределения расстояний: пару точек $(x, y) \in X^2$ назовем *коротким ребром*, если $\rho(x, y) \leqslant r$; *длинным ребром*, если $\rho(x, y) > 3r$, и *средним ребром* иначе. Набор точек $(x_1, \ldots, x_k)$ назовем *антикликой порядка k*, если $\rho(x_i, x_j) > r$ при всех $1 \leqslant i < j \leqslant k$. Потребуем, чтобы в метрическом пространстве (X, ρ) мера средних ребер была *мала* в следующем смысле:

$$\mu\{(x, y) \in X^2 \colon r < \rho(x, y) \leqslant 3r\} \leqslant \alpha\mu(X)^2, \tag{1}$$

а мера антиклик порядка $k + 1$ была *мала* в следующем смысле:

$$\mu\{(x_1, \ldots x_{k+1}) \in X^{k+1} \colon \rho(x_i, x_j) > r, \quad 1 \leqslant i < j \leqslant k + 1\} \leqslant \beta\mu(X)^{k+1}, \tag{2}$$

где $\alpha, \beta > 0$ — параметры. Основным результатом работы является нижняя оценка величины $\mu(\mathcal{X}^*)$ при ограничениях (1) и (2):

$$\frac{\mu(\mathcal{X}^*)}{\mu(X)} \geqslant 1 - \sqrt{\alpha}(2k + 1) - (k(e + 1) + 1)\beta^{1/(k+1)}.$$

Вначале данная оценка получается для конечного полуметрического пространства, а затем с помощью теоремы Бляшке обобщается на случай компактного пространства.

Interdependence of clusters measures and distance distribution in compact metric spaces

Pushnyakov Alexey pushnyakovalex@mail.ru
Russia, Dolgoprudny, MIPT

A compact metric space (X, ρ) is given. Let μ be a Borel measure on X. By r-cluster, let mean a measurable subset of X with diameter at most r. A family of k $2r$-clusters is called an *r-cluster structure of order k* if any two clusters from the family are separated by a distance at least r. By measure of a cluster structure, let mean a sum of clusters measures from the cluster structure. Using the Blaschke selection theorem, one can prove that there exists a cluster structure $\mathcal{X}^*$ of maximum measure. Let us study dependence $\mu(\mathcal{X}^*)$ on distance distribution. The main issue is to find restrictions for distance distribution which guarantee that $\mu(\mathcal{X}^*)$ is *close* to $\mu(X)$. Let us propose the following discretization of distance distribution. Let say that a pair $(x, y) \in X^2$ is a *short edge* if $\rho(x, y) \leqslant r$, (x, y) is a *long edge* if $\rho(x, y) > 3r$ and (x, y) is a *medium edge* otherwise. A sequence $(x_1, \ldots, x_k)$ is called an *anticlique of order k* if $\rho(x_i, x_j) > r$ for all $1 \leqslant i < j \leqslant k$. Suppose, one has the following conditions for measure of medium edges

$$\mu\{(x, y) \in X^2 \colon r < \rho(x, y) \leqslant 3r\} \leqslant \alpha\mu(X)^2 \tag{1}$$

and anticliques of order k:

$$\mu\{(x_1, \ldots x_{k+1}) \in X^{k+1} \colon \rho(x_i, x_j) > r, \\ 1 \leqslant i < j \leqslant k+1\} \leqslant \beta\mu(X)^{k+1}. \tag{2}$$

Under conditions (1) and (2), the following lower bound for $\mu(\mathcal{X}^*)$ has been proved:

$$\frac{\mu(\mathcal{X}^*)}{\mu(X)} \geqslant 1 - \sqrt{\alpha}(2k+1) - (k(e+1)+1)\beta^{1/(k+1)}.$$

First, this bound was obtained in the case of a finite metric space and then, it was generalized for a compact metric space.

О метрических пространствах, возникающих при формализации задач распознавания и классификации: свойства компактности

Торшин Иван Юрьевич[1] tiy1357@yandex.ru
Рудаков Константин Владимирович[1,2]⋆ rudakov@ccas.ru
[1]Россия, Долгопрудный, МФТИ
[2]Россия, Москва, ФИЦ ИУ РАН

В контексте алгебраического подхода к распознаванию научной школы Ю. И. Журавлёва метрический анализ признаковых описаний необходим для получения адекватных постановок для плохо формализованных задач распознавания/классификации. Проблематика формализации задач распознавания лежит на стыке между «контролируемым» (*англ.* supervised) и «неконтролируемым обучением» (*англ.* unsupervised machine learning). Проведен анализ компактности метрических пространств, возникающих в ходе формализации задач распознавания. Исследованы необходимые и достаточные условия компактности метрических пространств над решетками множеств признаковых описаний. Сформулированы подходы к пополнению изучаемых дискретных метрических пространств (пополнение расширением решетки или вариацией оценки). Показано, что при анализе компактности метрических пространств могут быть получены используемые в кластерном анализе эвристические критерии кластера. В ходе анализа свойств компактности возникает центральное понятие ро-сети как подмножества точек, позволяющего оценить произвольное расстояние в произвольной метрической конфигурации. Анализ свойств компактности и вводимый понятийный аппарат (ро-сети и их функционалы качества, условие метрического диапазона, i- и ро-спектры, эпсилон-окрестности в метрическом конусе, эпсилон-изоморфизм полных взвешенных графов и др.) позволяют применять методы функционального анализа, теория вероятностей, метрической геометрии и теории графов для анализа плохо формализованных задач распознавания и классификации [1]. Работа выполнена при поддержке грантов РФФИ №№ 12-07-00485, 12-07-00457, 13-07-12053 и 14-07-00852.

[1] *Torshin I. Yu. Rudakov K. V.* Metric spaces in formalization of problems of recognition and classification // Pattern Recogn. Image Anal., 2016. No. 4. P. 145–155.

On metric spaces arising during formalization of problems of recognition and classification: Properties of compactness

Torshin Ivan[1] tiy1357@yandex.ru
Rudakov Konstantin[1,2]⋆ rudakov@ccas.ru
[1]Dolgoprudniy, Russia, MIPT
[2]Moscow, Russia, FRC CSC RAS

In the context of the Zhuravlev's algebraic approach, metric analysis of the feature descriptions is required to obtain adequate formulations for poorly formalized problems of recognition/classification. Problematics of formalizing the recognition problems lies at the junction between supervised machine learning and unsupervised machine learning. This work presents results of analysis of compact metric spaces arising during the process of formalization. Necessary and sufficient conditions of compactness of the metric space over lattices of sets of features have been investigated and approaches to study to completeness of the studied discrete metric spaces (completion by lattice expansion or completion by variation of the lattice isotonic estimate) have been obtained. In particular, it is shown that the analysis of compactness of metric spaces leads to some of the heuristic criteria of clusters commonly used in the cluster analysis. During the analysis of the properties of compactness, an important concept of a "ro-network" as a subset of points of an arbitrary metric configuration arises that allows to estimate any distance in the configuration. The conceptual apparatus introduced (ro-networks, functionals of their quality, the condition of the metric range, i-spectra and ro-spectrum, epsilon-neighborhood in a metric cone, epsilon-isomorphism of complete weighted graphs, etc.) makes it possible to use various methods of functional analysis, probability theory, metric geometry, and graph theory to analyze poorly formalized problems of recognition and classification [1].

This research is funded by RFBR, grants Nos. 12-07-00485, 12-07-00457, 13-07-12053, and 14-07-00852.

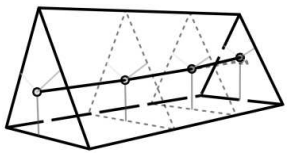
[1] Torshin, I. Yu., and K. V. Rudakov. 2016. Metric spaces in formalization of problems of recognition and classification. *Pattern Recogn. Image Anal.* 4:145–155.

Прогнозирование результатов обучения студентов с использованием смешанных диагностических тестов и 2-симплекс призмы

Янковская Анна Ефимовна[1,2,3,4]⋆ ayyankov@gmail.com
Дементьев Юрий Николаевич[4] dementev@tpu.ru
Ямшанов Артем Вячеславович[3] yav@keva.tusur.ru
Ляпунов Данил Юрьевич[3,4] lyapdy@gmail.com

[1]Россия, Томск, ТГАСУ
[2]Россия, Томск, НИ ТГУ
[3]Россия, Томск, ТУСУР
[4]Россия, Томск, ТПУ

Прогнозирование результатов обучения студентов является весьма актуальной задачей в образовательном процессе. Обсуждается текущее состояние исследований в области электронного обучения. Приводятся математические основы оценки обучения студентов на основе смешанных диагностических тестов (СДТ). Описываются два когнитивных средства: 2-симплекс и 2-симплекс призма; математические основы и специфика визуализации. Предлагается новый подход к прогнозированию результатов обучения, основанный на СДТ и 2-симплекс призме (см. рисунок). Приводятся результаты обработки тестов для студентов и примеры их визуализации с использованием 2-симплекс призмы для курса «Selected Chapters of Electronics». Обсуждается предложенный подход прогнозирования и когнитивная визуализация. Описывается специфика кросс-платформенной программной реализации когнитивных средств, инвариантных к проблемным областям. Обсуждаются результаты и дальнейшие исследования [1].



Прогнозирование с применением 2-симплекс призмы

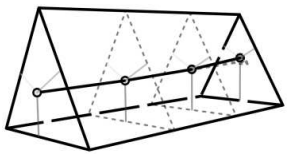
Работа поддержана грантом РФФИ № 16-07-00859 и частично грантом РФФИ № 14-07-00673а.

[1] *Yankovskaya A., Dementyev Yu., Yamshanov A., Lyapunov D.* Prediction of students' learning results with usage of 2-simplex prism and mixed diagnostic tests // Машинное обучение и анализ данных, 2016 (в печати).

Prediction of students' learning results with usage of mixed diagnostic tests and 2-simplex prism

Yankovskaya Anna[1,2,3,4]⋆ ayyankov@gmail.com
Dementyev Yury[4] dementev@tpu.ru
Yamshanov Artem[3] yav@keva.tusur.ru
Lyapunov Danil[3,4] lyapdy@gmail.com
[1]Tomsk, Russia, TSUAB
[2]Tomsk, Russia, TSU
[3]Tomsk, Russia, TUSUR
[4]Tomsk, Russia, TPU

Prediction of students' learning results is one of the "hottest" problems in modern learning process. Current state of research in e-learning area is discussed. Mathematical framework of students' learning results assessment on the base of mixed diagnostic tests (MDT) is given. Cognitive tools 2-simplex and 2-simplex prism, the mathematical basis, and visualization essentials are described. A new approach to the prediction of students' learning results, based on MDT and 2-simplex prism (see the figure), is proposed. Students' test results for course "Selected Chapters of Electronics" and examples of their cognitive visualization are given. The proposed approach to prediction and cognitive visualization are discussed. Specificity of cross-platform software implementation of cognitive graphics tools invariant to problem areas is described. Results and future research are discussed [1].



Prediction with 2-simplex prism application

This research is funded by the Russian Foundation for Basic Research No. 16-07-00859 and partially No. 14-07-00673a.

[1] Yankovskaya, A., Yu. Dementyev, A. Yamshanov, and D. Lyapunov. 2016 (in press). Prediction of students' learning results with usage of 2-simplex prism and mixed diagnostic tests. *Machine Learning Data Anal.*

Восстановление произвольных нестационарных зависимостей в линейном пространстве наблюдений

Красоткина Ольга Вячеславовна[1] o.v.krasotkina@yandex.ru
Моттль Вадим Вячеславович[2] vmottl@yandex.ru
Турков Павел Анатольевич[3]⋆ pavel-turkov@yandex.ru
[1]Россия, Москва, МГУ
[2]Россия, Москва, ФИЦ ИУ РАН
[3]Россия, Тула, ТулГУ

Предположение о том, что всякий объект реального мира представляется в компьютере как точка линейного пространства, позволяет построить единую методологию восстановления зависимостей по множеству прецедентов, допускающую измерение целевой характеристики объекта в произвольной шкале. От этой шкалы зависит лишь выбор так называемой параметрической функции потерь, содержащей три аргумента: специфическую целевую переменную и две точки линейного пространства, одна из которых отображает наблюдаемые свойства объекта, а вторая, называемая направляющей точкой, полностью определяет модель зависимости в данный момент времени. Другим важнейшим элементом методологии является регуляризующая функция от направляющей точки линейного пространства, или, в вероятностных терминах, ее априорное распределение, выбор которого не связан с типом искомой зависимости. В данной работе это распределение рассматривается как марковский случайный процесс, развивающийся во времени. Обучение в реальном времени сводится к оцениванию двух его параметров — размерности базиса в исходном линейном пространстве и волатильности нестационарной зависимости по времени. Хотя вид текущего апостериорного распределения направляющей точки зависит от типа целевой переменной, его нормальная аппроксимация обеспечивает универсальность алгоритма обучения на основе фильтра Калмана [1].

Работа поддержана грантом РФФИ № 14-07-00964.

[1] *Turkov P., Krasotkina O., Mottl V., Sychugov A.* Feature selection for handling concept drift in the data stream classification // Machine learning and data mining in pattern recognition / Ed. P. Perner. — Lecture notes in computer science ser. — Springer, 2016. Vol. 9729. P. 614–629.

Estimation of arbitrary nonstationary dependences in a linear observation space

Krasotkina Olga[1] o.v.krasotkina@yandex.ru
Mottl Vadim[2] vmottl@yandex.ru
Turkov Pavel[3]⋆ pavel-turkov@yandex.ru
[1]Moscow, Russia, MSU
[2]Moscow, Russia, FRC CSC RAS
[3]Tula, Russia, TulSU

The assumption that any real-world object is represented in the computer as a point of a linear space allows for constructing a unified technique of dependence estimation from a set of precedents, which admits for measuring the object's goal characteristics in an arbitrary scale. In this work, a nonstationary formulation of the generalized dependence-estimation problem has been considered. The scale of the goal variable affects only the choice of the so-called parametric loss function that contains three arguments: the goal variable and two points of the linear space, one of which is the image of the observable properties of the object, whereas the second one, called the direction point, completely determines the model of the dependence at the current moment of time. Another crucial element of the technique is the regularization function of the direction point or, in probabilistic terms, its *a priori* distribution, the choice of which is not tied to the kind of the sought-for dependence. The latter distribution is treated here as a Markov random process on the time axis. The real-time training consists in estimating two parameters of the hidden process: the dimensionality of the basis in the initial linear space and the time-volatility of the nonstationary dependence. Despite the fact that the class of the current *a posteriori* distribution of the direction point depends on the kind of the goal variable, a normal approximation of this distribution provides the universality of the training algorithm which is based on the Kalman filtration principle [1].

This research is funded by the Russian Foundation for Basic Research, grant 14-07-00661.

[1] Turkov, P., O. Krasotkina, V. Mottl, and A. Sychugov. 2016. Feature selection for handling concept drift in the data stream classification. *Machine learning and data mining in pattern recognition*. Ed. P. Perner. Lecture notes in computer science ser. Vol. 9729. P. 614–629.

Верификация волатильности модели в задачах оценивания нестационарных зависимостей

Красоткина Ольга Вячеславовна[1]⋆ o.v.krasotkina@yandex.ru
Моттль Вадим Вячеславович[2] vmottl@yandex.ru
Черноусова Елена Олеговна[2] lena-ezhova@rambler.ru
[1]Россия, Москва, МГУ
[2]Россия, Москва, ФИЦ ИУ РАН
[3]Россия, Москва, МФТИ

Рассматривается задача обучения по прецедентам, в которой объекты упорядочены вдоль некоторой координаты. Дополнительно предполагается, что искомая зависимость изменяется во времени, что в терминах линейной регрессии либо распознавания двух классов объектов означает свой набор элементов направляющего вектора в каждый момент времени. Такая постановка является некорректной без принятия дополнительных предположений о характере изменения направляющего вектора. В качестве априорной модели его динамики рассматривается марковский случайный процесс, важнейшим структурным параметром которого является его волатильность, изменение которой от нуля до бесконечности определяет последовательность почти вложенных классов моделей динамики скрытого процесса. На сегодняшний день для решения оценивания структурного параметра волатильности в нестационарных моделях в основном используются кроссвалидационные методы, которые склонны к выбору слишком стационарных моделей, что в ряде практических задач является неприемлемым. Впервые предложено использовать для подбора структурного параметра критерий обоснованности иерархической вероятностной модели нестационарной линейной регрессии. Проведено экспериментальное сравнение предложенного способа оценивания структурного параметра волатильности с кроссвалидационными способами подбора параметра и обобщением информационного критериея Акаике [1].

Работа поддержана грантом РФФИ № 14-07-00964.

[1] *Красоткина О. В., Моттль В. В.* Использование критерия обоснованности модели для подбора структурных параметров в задаче восстановления нестационарной линейной регрессии // Известия Тульского гос. ун-та, 2016 (в печати).

Model volatility verification in problems of nonstationary dependence estimation

Krasotkina Olga[1]⋆ o.v.krasotkina@yandex.ru
Mottl Vadim[2] vmottl@ccas.ru
Chernousova Elena[3] lena-ezhova@rambler.ru
[1]Moscow, Russia, Moscow State University
[2]Moscow, Russia, FRC CSC RAS
[3]Moscow, Russia, MIPT

The problem of supervised learning, in which the objects are arranged along some axis, usually time, has been considered. It is assumed, in addition, that the sought-for regression dependence is changing over time. However, such a learning problem statement is incorrect without additional assumptions about the nature of changes in regression coefficients. As the *a priori* model of the coefficient dynamics, a Markov random process that contains a structural parameter having the sense of the volatility of hidden process has been considered. The volatility parameter ranges from the full invariability of regression coefficients in time to their absolute independence of each other. The leave-one-out cross-validation in nested sets of data models is traditionally considered in Machine Learning as the basic instrument of finding the most appropriate values of structural parameters. But in the case of nonstationary models, they lead to excessively static estimates of the hidden process that is unacceptable in a number of practical situations. In this paper, a new criterion for time-volatility adjustment, which is based on the evidence estimation for hierarchical probabilistic model of time-dependent data, has been proposed. An integral part of the work is a detailed comparison of the proposed method with the cross-validation procedure and generalized Akakike criterion [1]. This research is funded by the Russian Foundation for Basic Research, grant 14-07-00964.

[1] Krasotkina, O. V., and V. V. Mottl. 2016 (in press). Ispol'zovanie kriteriya obosnovannosti modeli dlya podbora strukturnykh parametrov v zadache vosstanovleniya nestatsionarnoy lineynoy regressii [Use of the model of the validity criterion for the selection of the structural parameters in the problem of recovery of nonstationary linear regression]. *Izvestiya Tul'skogo gos. un-ta* [News of the Tula State. Univ.]

Численные методы проверки обоснованности обобщенных линейных моделей зависимостей

Левдик Павел Владимирович[1]⋆ cold62@mail.ru
Моттль Вадим Вячеславович[2] vmottl@yandex.ru
Красоткина Ольга Вячеславовна[3] o.v.krasotkina@yandex.ru
Татарчук Александр Игоревич[2] aitech@yandex.ru
[1]Россия, Москва, МФТИ
[2]Россия, Москва, ФИЦ ИУ РАН
[3]Россия, Москва, МГУ

Алгоритмическая реализация принципа проверки обоснованности модели в задачах восстановления зависимостей требует вычисления смеси условных распределений целевой переменной модели (marginal likelihood) относительно ее параметра, априорное распределение которого играет роль регуляризующего смешивающего распределения. Обобщенный линейный подход к восстановлению зависимостей основан на понимании этого параметра как точки в том же линейном пространстве, в котором представлены объекты реального мира. Это позволяет без потери общности ограничиться классом смешивающих априорных распределений, построенных на основе нормального распределения. Однако параметрическое семейство смешиваемых распределений не может быть унифицировано для всех видов целевой переменной, и вычисление показателя обоснованности остается задачей, традиционно квалифицируемой как алгоритмически трудная. Обсуждаемый универсальный метод итерационной максимизации показателя обоснованности использует классический EM (expectation–maximization) принцип максимизации правдоподобия смеси распределений. В частности, в задачах линейной регрессии, распознавания образов и оценивания продолжительности жизни метод позволяет находить наиболее обоснованную размерность базиса в линейном пространстве наблюдений без перебора всех значений размерности, как в популярном методе Regularization Path [1].

Работа поддержана грантом РФФИ № 14-07-00661.

[1] *Chernousova E., Levdik P., Tatarchuk A., Mottl V., Windridge D.* Non-enumerative cross validation for the determination of structural parameters in feature-selective SVMs // ICPR Proceedings, 2014. P. 3654–3659.

Numerical evidence evaluation for generalized linear dependence models

Levdik Pavel[1]⋆ cold62@mail.ru
Mottl Vadim[2] vmottl@yandex.ru
Krasotkina Olga[3] o.v.krasotkina@yandex.ru
Tatarchuk Alexander[2] aitech@yandex.ru
[1]Moscow, Russia, MIPT
[2]Moscow, Russia, FRC CSC RAS
[3]Moscow, Russia, MSU

Algorithmic implementation of model evidence evaluation in the problems of dependence estimation is bound to the necessity to compute the mixtures of conditional distributions of the target variable (marginal likelihood) with respect to the random model parameter, whose *a priori* distribution plays the role of the mixing regularization distribution. The generalized linear approach to dependence estimation is based on treating this parameter as a point in the same linear space in which real-world objects are meant to be represented. This allows to restrict the consideration, without loss of generality, to the class of mixing *a priori* distributions being generalization of the normal distribution. But the parametric family of the distributions to be mixed cannot be unified for all the kinds of the target variable, and the evaluation of the evidence remains to be the problem which is traditionally qualified as an algorithmically difficult one. The present authors consider a universal method of iterative maximization of the model evidence, which exploits the classical EM (expectation–maximization) principle of mixture likelihood maximization and does not require direct computation of the evidence. In particular, for linear regression, pattern recognition, and survival analysis, the method provides finding the most evident dimension of the basis in the linear object-representation space without enumeration of all the values of the dimension, how it is presupposed by the popular regularization path method [1].

This research is funded by RFBR, grant 14-07-00661.

[1] Chernousova, E., P. Levdik, A. Tatarchuk, V. Mottl, and D. Windridge. 2014. Non-enumerative cross validation for the determination of structural parameters in feature-selective SVMs. *ICPR Proceedings*. 3654–3659.

Быстрые последовательные методы обучения обобщенных линейных моделей зависимостей

Маленичев Антон Александрович[1]⋆ malenichev@mail.ru
Красоткина Ольга Вячеславовна[2] o.v.krasotkina@yandex.ru
Моттль Вадим Вячеславович[3] vmottl@yandex.ru
[1]Россия, Тула, ТулГУ
[2]Россия, Москва, МГУ
[3]Россия, Москва, ФИЦ ИУ РАН

Рассматривается новый принцип построения последовательных процедур восстановления произвольных зависимостей (регрессионных моделей, решающих правил распознавания образов, моделей выживаемости, порядковых моделей) по очень большим массивам прецедентов на основе обобщенного линейного подхода к проблеме обучения. Процесс обучения построен как последовательный пересчет функций Беллмана в некоторой специальной процедуре динамического программирования. Строго оптимальный результат обучения произвольной (не обязательно линейной) модели зависимости на каждом шаге относительно уже полученной обучающей совокупности сводится к минимизации очередной функции Беллмана. Показано, что для весьма широкого класса моделей допустима квадратичная аппроксимация каждой функции Беллмана в окрестности точки ее минимума, приводящая к классическому фильтру Калмана. Предлагается компенсировать неизбежную потерю оптимальности обучения дополнительной коррекцией каждого шага по классическому принципу стохастической аппроксимации, который гарантирует сходимость обучения «почти наверное». Хотя такое «дотягивание» обучения является очень медленным, что характерно для стохастической аппроксимации, оценка на каждом шаге остается в небольшой окрестности строгого минимума «глобального» критерия для обработанной части массива обучающих данных [1].

Работа поддержана грантом РФФИ № 14-07-00964.

[1] *Turkov P., Krasotkina O., Mottl V., Sychugov A.* Feature selection for handling concept drift in the data stream classification // Machine learning and data mining in pattern recognition. — Lecture notes in computer science ser. — Springer, 2016. Vol. 9729. P. 614–629.

Quick methods of online learning for generalized linear models of arbitrary dependences

Malenichev Anton[1]⋆ malenichev@mail.ru
Krasotkina Olga[2] o.v.krasotkina@yandex.ru
Mottl Vadim[3] vmottl@ccas.ru
[1]Tula, Russia, TulSU
[2]Moscow, Russia, MSU
[3]Moscow, Russia, FRC CSC RAS

A new method of constructing online procedures for estimation of arbitrary dependences (regression models, pattern-recognition rules, survival models, ranking models) from very large data sets on the basis of the generalized linear approach to the problem of training is considered. The training process is built as iterative recomputing of Bellman functions in a special dynamic programming procedure. The strictly optimal result of training the model of the respective dependence, not obligatory a linear one, with respect to the part of the training set, which is already registered by the moment of current observation, is nothing else than the minimum point of the last Bellman function. It is shown that a quite broad class of models allow for quadratic approximation of each Bellman function in a vicinity of its minimum point that results in the classical Kalman filter. The present authors propose to compensate the inevitable loss of optimality of training by an additional correction of the respective step in accordance with the classical stochastic approximation principle, which guaranties the convergence of the training process "almost for sure." Despite the fact that such a fine-tuning is extremely slow, as it is characteristic for stochastic approximation, the current estimate remains in a small vicinity of the strict minimum of the "global" criterion for the already processed part of the entire training data set [1].

This research is funded by the Russian Foundation for Basic Research, grant 14-07-00964.

[1] Turkov, P., O. Krasotkina, V. Mottl, and A. Sychugov. 2016. Feature selection for handling concept drift in the data stream classification. *Machine learning and data mining in pattern recognition*. Lecture notes in computer science ser. Springer. 9729:614–629.

Проверка обоснованности обучаемых моделей зависимостей: обобщенный линейный подход

Моттль Вадим Вячеславович[1]⋆ vmottl@yandex.ru
Левдик Павел Владимирович[2] cold62@mail.ru
Красоткина Ольга Вячеславовна[3] o.v.krasotkina@yandex.ru
[1]Россия, Москва, ФИЦ ИУ РАН
[2]Россия, Москва, МФТИ
[3]Россия, Москва, МГУ

В отличие от кросс-валидации, требующей многократного повторения процесса обучения при разных разбиениях совокупности прецедентов, метод проверки обоснованности модели (marginal likelihood) предполагает однократное вычисление критерия, зависящего только от исходных данных и варьируемых параметров регуляризации. Однако вычисление показателя обоснованности является простой задачей только для модели линейной регрессии. В остальных случаях, в частности в задачах распознавания образов, анализа выживаемости, оценивания порядковой регрессии, вычисление показателя обоснованности становится трудной задачей. В данной работе используется специфика обобщенного линейного подхода к восстановлению зависимостей, основанного на необременительном предположении, что каждый объект реального мира воспринимается компьютером как точка в некотором линейном пространстве. Это обстоятельство позволяет практически без потери общности ограничиться классом априорных распределений параметра модели, построенных на основе нормального распределения. Однако в случае целевой переменной, измеряемой в произвольной шкале, апостериорное распределение параметра модели все равно не будет нормальным. Его нормальная аппроксимация позволила построить единый алгоритм выбора размерности линейного пространства представления объектов путем максимизации показателя обоснованности, применимый к задачам восстановления произвольных зависимостей [1].

Работа поддержана грантом РФФИ № 14-07-00661.

[1] *Chernousova E., Levdik P., Tatarchuk A., Mottl V., Windridge D.* Non-enumerative cross validation for the determination of structural parameters in feature-selective SVMs // ICPR Proceedings, 2014. P. 3654–3659.

Evidence evaluation for trainable models of arbitrary dependences: A generalized linear approach

Mottl Vadim[1]⋆ vmottl@yandex.ru
Levdik Pavel[2] cold62@mail.ru
Krasotkina Olga[3] o.v.krasotkina@yandex.ru
[1]Russia, Moscow, FRC CSC RAS
[2]Russia, Moscow, MIPT
[3]Russia, Moscow, MSU

As distinct from cross-validation, which requires multiple repetition of training with different partitions of the set of precedents, the evidence evaluation method (marginal likelihood) implies computation of a single criterion that depends only on the original data set and regularization parameters to be varied. However, computation of the evidence criterion is an easy procedure only for the linear regression model. In other cases, in particular, in the problems of pattern recognition, survival analysis, and rank regression estimation, evidence evaluation becomes a difficult problem. The present authors exploit the specificity of the generalized linear approach to dependence estimation, which is based on the lenient assumption that any real-world object is perceived by the computer as a point in some linear space. This circumstance allows one to confine the consideration, practically without the loss of generality, to the class of *a priori* distributions of the model parameter, which are the extensions of the normal distribution. Nevertheless, in the case of the target variable measured in an arbitrary scale, the *a posteriori* distribution will not be normal anyway. Its normal approximation has led to a unified technique of choosing the dimensionality of the object representation linear space by the way of maximizing the evidence value. The technique is valid for estimation of arbtrary depebdences [1].

This research is funded by the Russian Foundation for Basic Research, grant 14-07-00661.

[1] Chernousova, E., P. Levdik, A. Tatarchuk, V. Mottl, and D. Windridge. 2014. Non-enumerative cross validation for the determination of structural parameters in feature-selective SVMs. *ICPR Proceedings*. 3654–3659.

Обобщенный линейный подход к восстановлению зависимостей по эмпирическим данным

Моттль Вадим Вячеславович[1,2]⋆ vmottl@yandex.ru
Середин Олег Сергеевич[2] oseredin@yandex.ru
[1]Россия, Москва, ФИЦ ИУ РАН
[2]Россия, Тула, ТулГУ

Обычно методы обучения по прецедентам рассматривают отдельно для каждого вида зависимости — оценивания числовой регрессии, обучения распознаванию образов, прогнозирования продолжительности жизни. В любом случае требуется построить модель связи между объектом реального мира, воспринимаемым компьютером через измеренное значение его наблюдаемого свойства, и некоторой скрытой характеристикой объекта. Рассматривается единая методология восстановления произвольной зависимости по множеству прецедентов. Предполагается лишь, что наблюдаемая характеристика объекта представлена как элемент нормированного линейного пространства. Обучение сводится к поиску направляющей точки линейного пространства, являющейся моделью искомой зависимости. Наряду со способом погружения объектов в нормированное линейное пространство наблюдатель должен выбрать неотрицательную параметрическую функцию потерь, имеющую три аргумента: (1) точку отображения объекта; (2) его целевую характеристику, измеряемую в произвольной шкале; (3) направляющий элемент. Наблюдатель выбирает также неотрицательную регуляризующую функцию, выражающую априорное представление о положении направляющей точки. Выбор этих двух функций полностью охватывает все известные способы обучения. Здесь нет различия между линейными и нелинейными моделями зависимостей: модели, линейные относительно одной нормы, будут нелинейными в терминах другой. Естественная байесовская интерпретация обеспечивает проверку обоснованности выбора модели [1].

Работа поддержана грантом РФФИ № 14-07-00661.

[1] *Середин О. С., Моттль В. В.* Метод опорных объектов для обучения распознаванию образов в произвольных метрических пространствах // Известия ТулГУ, Естественные науки, 2015. Вып. 4. С. 49–66.

A generalized linear approach to estimation of dependences from empirical data

Mottl Vadim[1,2]⋆ vmottl@yandex.ru
Seredin Oleg[2] oseredin@yandex.ru
[1]Moscow, Russia, FRC CSC RAS
[2]Tula, Russia, TulSU

The methods of dependence estimation from empirical data are usually considered separately for each particular problem — regression estimation, pattern recognition, survival prediction. It is required, in each case, to develop a model of dependence between a real-world entity, which is assumed to be perceived by the computer via the measured value of one of its observable properties, and its hidden goal variable. In this work, a unified methodology is considered that is meant to be able to estimate any arbitrary dependence from a set of precedents. It is only assumed that the observable properties of an entity are represented as points in a normed linear space. The training boils down to the search, in this linear space, for an appropriate direction point, which will serve as the model of the sought-for dependence. Along with the way of embedding any entity into a normed linear space, it is up to the observer to choose a nonnegative parametric loss function that includes three arguments: (*i*) the representation point of the entity; (*ii*) its goal characteristic measured in an arbitrary scale; and (*iii*) the direction point of the model. In addition, the observer has to assign a nonnegative regularization function meant to express his/her *a priori* suggestion on the position of the direction point. The choice of these two functions completely covers all the known methods of training. There is no distinction here between linear and nonlinear models of dependences — if a model is linear with respect to some norm, it will be nonlinear in terms of another one. A natural Bayesian interpretation provides the evidence control of the model choice [1].

This research is funded by the Russian Foundation for Basic Research, grant 14-07-00661.

[1] Seredin, O., and V. Mottl. 2015. The support vector method for pattern recognition in arbitrary metric spaces. *News of the Tula State University, Natural Sciences* 4:49–66. (In Russian.)

Исследование эффективности некоторых линейных методов классификации

Неделько Виктор Михайлович nedelko@math.nsc.ru
Россия, Новосибирск, ИМ СО РАН

Рассматривается проблема построения вероятностных моделей, позволяющих выявлять свойства методов построения решающих функций и проводить исследование этих методов. В частности, ставилась задача построения моделей, на которых заданный метод наиболее эффективен среди сравниваемых методов.

Для метода логистической регрессии были построены модели, на которых этот метод эквивалентен методу максимального правдоподобия. Для метода SVM (support vector machine) построена модель, на которой этот метод приближенно эквивалентен методу максимального правдоподобия. Для дискриминанта Фишера подобной модели построить не удалось.

Проведенное исследование демонстрирует принципиальную возможность построения набора «эталонных» вероятностных моделей для исследования и сравнения методов построения решающих функций.

Также в работе выявлены некоторые неочевидные свойства метода SVM и особенности его поведения, учет которых позволяет более эффективно применять данный метод.

Работа поддержана грантом РФФИ № 14-01-00590.

[1] *Неделько В. М.* Исследование эффективности некоторых линейных методов классификации на модельных распределениях // Машинное обучение и анализ данных, 2016 (в печати).

Investigation of effectiveness of several linear classifiers

Nedel'ko Victor nedelko@math.nsc.ru

Novosibirsk, Russia, IM SB RAS

The most common way to compare the effectiveness of data analysis methods is testing on tasks from UCI repository. However, this approach has several disadvantages, in particular, the incompleteness of the set of tasks and limited sample sizes.

In this paper, the possibility of building a repository of probabilistic distributions is considered. The distributions are constructed purposefully in such a way as to reveal properties of the studied methods. Such distributions will be called probabilistic models.

Let us choose some linear classification methods for research: logistic regression, Fisher discriminant, and SVM (support vector machine).

Several probabilistic models have been constructed to investigate the properties of these methods, in particular, for each method, a model, on which this method outperforms the other methods, was built. In addition, these models allow one to explain why a particular method is the best.

This research is funded by the Russian Foundation for Basic Research, grant 14-01-00590.

[1] Nedel'ko, V. M. 2016 (in press). Investigation of effectiveness of several linear classifiers by using synthetic distributions. *Machine Learning Data Anal.*

Методы погружения произвольных объектов реального мира в нормированное линейное пространство для реализации обобщенного линейного подхода к восстановлению зависимостей

Середин Олег Сергеевич[1]⋆ oseredin@yandex.ru
Моттль Вадим Вячеславович[1,2] vmottl@yandex.ru
[1]Россия, Тула, ТулГУ
[2]Россия, Москва, РАН

Обобщенный линейный подход к восстановлению зависимостей по эмпирическим данным предполагает, что рассматриваемое множество объектов реального мира погружено в некоторое нормированное линейное пространство. Показано, что такое погружение обеспечивает всякая симметричная двухместная числовая функция, определенная наблюдателем на множестве объектов как функция их парного сравнения. В получающемся бесконечномерном линейном пространстве эта функция определяет скалярное произведение, в общем случае индефинитное. Искомая зависимость формируется выбором направляющей точки, которую удобно задавать как линейную комбинацию образов объектов обучающей совокупности. Предлагается рассматривать скалярное произведение образа объекта и направляющей точки как обобщенный числовой признак объекта, играющий роль его единственной характеристики. Выбор метода обучения сводится к выбору двух числовых функций — трехместной обобщенной функции потерь, связывающей точку образа объекта, его целевую характеристику, измеряемую в произвольной шкале, и направляющую точку, а также одноместной регуляризующей функции от направляющей точки, выражающей априорные предпочтения относительно модели. Если принятая функция парного сравнения объектов обладает свойствами предевклидовой метрики, то из такой модели вытекают все кернельные методы обучения, в том числе конечномерные методы [1]. Работа поддержана грантами РФФИ № 14-07-00527 и № 16-57-52042.

[1] *Середин О. С., Моттль В. В.* Метод опорных объектов для обучения распознаванию образов в произвольных метрических пространствах // Известия ТулГУ, Естеств. науки, 2015. Вып. 4. С. 49–66.

Methods of embedding real-world entities into a normed linear space for implementing the generalized linear approach to dependence estimation

Seredin Oleg[1]⋆ oseredin@yandex.ru
Mottl Vadim[1,2] vmottl@yandex.ru
[1]Tula, Russia, TulSU
[2]Moscow, Russia, FRC CSC RAS

The generalized linear approach to dependence estimation suggests that the set of real-world entities of interest is embedded into a nirmed linear space. It is shown that such an embedding is provided by any symmetric numerical two-argument function defined by the observer over the set of ebtities as function of their pairwise comparison. In the resulting infinite-dimensional linear space, the preset pairwise comparison function defines a linear product which may be indefinite in the general case. The sought-for dependence is formed by the choice of a direction point, which is convenient to determine as a linear combination of the training-set entities. The present authors propose to consider the inner product of the entity's image and the direction point as a generalized numerical feature of the entity that will further play the role of its only characteristics. The choice of the training method boils down to the choice of two numerical functions — the three-argument generalized loss function meant to tie the image point of the entity, its goal characteristic that may be measured in an arbitrary scale, and the direction point as well as the one-argument regularization function of the direction point which should express the observer's *a priori* preferences regarding the model. If the accepted pairwise comparison function possesses the properties of a proto-Euclidean metric, then such a scheme yields all the kernel-based training methods including the finite-dimensional ones [1].

This research is funded by the Russian Foundation for Basic Research, grants 14-07-00527 and 16-57-52042.

[1] Seredin, O., and V. Mottl. 2015. The support vector method for pattern recognition in arbitrary metric spaces. *News of the Tula State University. Natural Sciences* 4:49–66. (In Russian.)

Точный псевдополиномиальный алгоритм для задачи поиска семейства непересекающихся подмножеств

Галашов Александр Евгеньевич[1]⋆ galashov.alexandr@gmail.com
Кельманов Александр Васильевич[1,2] kelm@math.nsc.ru
[1]Россия, Новосибирск, НГУ
[2]Россия, Новосибирск, ИМ СО РАН

Рассматривается NP-трудная в сильном смысле [1]

Задача. *Дано*: множество $\mathcal{Y} = \{y_1, \ldots, y_N\}$ точек из $\mathbb{R}^q$ и натуральные числа $M_1, \ldots, M_J$. *Найти:* семейство $\{\mathcal{C}_1, \ldots, \mathcal{C}_J\}$ непересекающихся подмножеств множества $\mathcal{Y}$ такое, что

$$\sum_{j=1}^{J} \sum_{y \in \mathcal{C}_j} \|y - \overline{y}(\mathcal{C}_j)\|^2 \to \min,$$

где $\overline{y}(\mathcal{C}_j) = (1/|\mathcal{C}_j|) \sum_{y \in \mathcal{C}_j} y$ — центроид (геометрический центр) подмножества $\mathcal{C}_j$, при ограничениях $|\mathcal{C}_j| = M_j$, $j = 1, \ldots, J$, на мощности искомых подмножеств.

Задача актуальна, в частности, для помехоустойчивого анализа данных.

Для случая задачи с целочисленными входами в работе предложен алгоритм, гарантирующий отыскание точного решения за время $\mathcal{O}(\mathcal{N}(\mathcal{N}^2 + q)(2MB + 1)^{qJ} + (J - 1) \lg N)$, где B — максимальное абсолютное значение координат входных точек, а M — наименьшее общее кратное чисел $M_1, \ldots, M_J$. В случае, когда размерность q пространства и число J искомых подмножеств не являются частью входа, предложенный алгоритм псевдополиномиален, а время его работы оценивается величиной $\mathcal{O}(\mathcal{N}^3(MB)^{qJ})$.

Работа поддержана грантами РФФИ № 15-01-00462 и № 16-07-00168.

[1] *Galashov A., Kel'manov A.* An exact pseudopolynomial algorithm for a problem of finding a family of disjoint subsets // 9th Conference (International) on Discrete Optimization and Operations Research Proceedings. Vladivostok, Russky Island, Russia, 2016 (in press).

An exact pseudopolynomial-time algorithm for a problem of finding a family of disjoint subsets

Galashov Alexandr[1]⋆ galashov.alexandr@gmail.com
Kel'manov Alexander[1,2] kelm@math.nsc.ru
[1]Novosibirsk, Russia, NSU
[2]Novosibirsk, Russia, IM SB RAS

The authors consider the following strongly NP-hard [1]

Problem. *Given* a set $\mathcal{Y} = \{y_1, \ldots, y_N\}$ of points from $\mathbb{R}^q$ and some positive integers $M_1, \ldots, M_J$. *Find* a family $\{\mathcal{C}_1, \ldots, \mathcal{C}_J\}$ of disjoint subsets of $\mathcal{Y}$ such that

$$\sum_{j=1}^{J} \sum_{y \in \mathcal{C}_j} \|y - \overline{y}(\mathcal{C}_j)\|^2 \to \min,$$

where $\overline{y}(\mathcal{C}_j) = (1/|\mathcal{C}_j|) \sum_{y \in \mathcal{C}_j} y$ is the centroid (geometrical center) of the subset $\mathcal{C}_j$, under constraints $|\mathcal{C}_j| = M_j, j = 1, \ldots J$, on the cardinalities of the required subsets.

The problem is relevant, in particular, for noise-proof data analysis.

In the current work, for the variation of the problem with an additional restriction that the coordinates of the input points are integer, an algorithm which finds an exact solution in $\mathcal{O}(\mathcal{N}(\mathcal{N}^2 + q)(2MB + 1)^{qJ} + (J - 1) \lg N)$ time is constructed, where B is the maximum absolute value of the coordinates of the input points and M is the least common multiple for the numbers $M_1, \ldots, M_J$. In the case of the fixed dimension q of the space and of the fixed number J of required subsets, the proposed algorithm is pseudopolynomial and its time complexity is bounded by $\mathcal{O}(N^3(MB)^{qJ})$.

This research is funded by the Russian Foundation for Basic Research, grants 15-01-00462 and 16-07-00168.

[1] Galashov, A., and A. Kel'manov. 2016 (in press). An exact pseudopolynomial algorithm for a problem of finding a family of disjoint subsets. *9th Conference (International) on Discrete Optimization and Operations Research Proceedings*. Vladivostok, Russky Island, Russia.

Реализация асимптотически точного подхода к построению полиномиальных алгоритмов решения некоторых трудных задач маршрутизации, назначения, покрытия и кластеризации

Гимади Эдуард Хайрутдинович[1,2] gimadi@math.nsc.ru
[1]Россия, Новосибирск, ИМ СО РАН
[2]Россия, Новосибирск, НГУ

За последние почти полвека (с 1969 г.) автором, совместно с его коллегами и учениками, был получен ряд эффективных реализаций асимптотически точного подхода к решению нижеперечисленных подклассов трудных задач дискретной оптимизации.

1. Задача коммивояжера (TSP — travelling salesman problem).
2. Задача упаковки в контейнеры и в полосу.
3. Задача построения связного подграфа с ограниченными степенями вершин.
4. Задача маршрутизации транспортных средств.
5. Трехиндексная задача о назначениях.
6. Задача размещения предприятий.
7. Задача отыскания m реберно непересекающихся маршрутов коммивояжера с экстремальным суммарным весом ребер.
8. Задача отыскания подмножества векторов с экстремальным суммарным весом;
9. Отыскание минимального остовного дерева с ограниченным снизу диаметром.
10. Покрытие графа заданным числом несмежных циклов.

Цель работы — представить обзор результатов по асимптотически точной разрешимости задач 7—10, полученных в самое последнее время. Один из таких результатов представлен в [1].

Исследования по задачам 7–8 поддержаны РНФ (грант 16-11-10041). Исследования по задачам 9–10 поддержаны РФФИ (гранты 15-01-00976 и 16-07-00168).

[1] *Gimadi E., Istomin A., Tsidulko O.* On the m-peripatetic salesman problem on random inputs // 7th Conference (International) “Optimization and Applications” Proceedings. Petrovac, Montenegro, 2016.

Implementation of the asymptotically optimal approach to polynomial time solving some hard discrete optimization problems of rooting, assigning, covering, and clustering

Gimadi Edward[1,2] gimad@math.nsc.ru
[1]Novosibirsk, Russia, IM SB RAS
[2]Novosibirsk, Russia, NSU

Over the past half-century (since 1969), the author, together with his colleagues and students, has built fast asymptotically optimal algorithms for a number of hard discrete optimization problems.

1. Travelling salesman problem.
2. Bin and strip packing problem.
3. Degree constrained connected subgraph problem.
4. Vehicle rooting problem.
5. Three-index assignment problem.
6. Capacitated facility location problem.
7. m-Peripatetic salesman problem.
8. Total weight subset problem.
9. Spanning tree problem with a bounded below diameter.
10. k Cycles covering problem.

In the report, some recent results of asymptotically optimal solvability for the problems 7–10 are presented. One such result is presented in [1].

The study of problems 7–8 was supported by the Russian Science Foundation (grant 16-11-10041). The study of problems 9–10 was supported by the Russian Foundation for Basic Research (grants 15-01-00976 and 16-07-00168).

[1] Gimadi, E., A. Istomin, and O. Tsidulko. 2016. On the m-peripatetic salesman problem on random inputs. *7th Conference (International) "Optimization and Applications" Proceedings*. Petrovac, Montenegro.

Алгоритмы и вычислительные технологии поиска экстремума в задачах оптимального управления

Горнов Александр Юрьевич* gornov@icc.ru
Зароднюк Татьяна Сергеевна tz@icc.ru
Аникин Антон Сергеевич htower@icc.ru
Финкельштейн Евгения Александровна finkel@icc.ru
Иркутск, Россия, ИДСТУ СО РАН

Задачи оптимального управления являются специальным классом экстремальных задач, связанных с поиском минимума функционала, определенного на траекториях динамической системы. Традиционно для описания динамики систем используются либо дифференциальные уравнения, либо рекуррентно-разностные уравнения. Принято выделять фазовые переменные, соотносимые с моделью явления, и переменные-управления. Имеет место множество обобщений этого формализма, включающих элементы неопределенности, возмущения, отклонения аргумента в уравнениях, интегральные компоненты и др.

С точки зрения теории экстремальных задач проблема поиска оптимального управления имеет ярко выраженную специфику, определяемую условиями типа равенства, задаваемыми системой дифференциальных или разностных уравнений. Усилиями множества известных специалистов за последние десятилетия найден набор подходов и методов, позволяющих эффективно решать задачи оптимального управления широкого класса на доступной вычислительной технике.

В работе рассматривается классификация проблем, возникающих при численном решении и обсуждаются различные подходы к их преодолению. Приводятся результаты вычислительных экспериментов и обсуждается опыт применения реализованных программных средств и вычислительных технологий при решении большого ряда прикладных задач из областей механики, динамики полета, космонавигации, электроэнергетики, робототехники, медицины, экологии, экономики, нанофизики и др. [1].

Работа частично поддержана грантом РФФИ № 15-07-03827.

[1] *Горнов А. Ю.* Вычислительные технологии решения задач оптимального управления. — Новосибирск: Наука, 2009. 278 с.

Algorithms and computational technology for extremum search in optimal control problems

Gornov Alexander⋆ gornov@icc.ru
Zarodnyuk Tatiana tz@icc.ru
Anikin Anton htower@icc.ru
Finkelstein Evgeniya finkel@icc.ru
Russia, Irkutsk, ISDCT of SB RAS

Optimal control problems are a special class of extremal problems concerned with the search of a functional minimum, which is defined on the trajectories of the dynamical system under additional constraints. Traditionally, for describing the dynamics of systems, one can use either differential equations or recurrence equations. It is customary to differ the phase variables, which are correlated with the model phenomena, and control variables in order to influence on the studied processes. There are a lot of generalizations of this formalism allowing to include uncertainties, disturbances, deviations of the argument in equations, integral components, etc.

From the point of view of the extremal problems theory, the problem of finding the optimal control has a distinct specificity determined by the equality constraint, it is given by a system of differential or recurrence equations. Many well-known experts made significant efforts over the past decade and found a set of approaches and methods for effectively solving the wide class optimal control on available computing technology.

The paper considers the classification of numerical problems and discusses various approaches to overcome these difficulties. The authors demonstrate the results of computational experiments and present the experience of using the implemented software and computing technologies for solving a wide range of tasks such as applied problems from the area of mechanics, flight dynamics, cosmonavigation, electrical power engineering, robotics, medicine, ecology, economics, nanophysics, etc. [1].

This research is partly supported by Grant No. 15-07-03827 of the Russian Foundation for Basic Research.

[1] Gornov, A. Yu. 2009. *Computational technologies for solving optimal control problems*. Novosibirsk: Nauka. 278 p.

О поиске подмножества векторов с минимальным нормированным квадратом длины суммы

Еремеев Антон Валентинович[1,2] eremeev@ofim.oscsbras.ru
Кельманов Александр Васильевич[2,3] kelm@math.nsc.ru
Пяткин Артем Валерьевич[2,3]⋆ artem@math.nsc.ru
[1]Россия, Омск, ОмГУ
[2]Россия, Новосибирск, ИМ СО РАН
[3]Россия, Новосибирск, НГУ

Анализируется статус вычислительной сложности известной дискретной экстремальной задачи с измененным направлением оптимизации, а именно: с max на min. Рассматривается евклидова задача поиска подмножества в конечном множестве точек (векторов). Целевая функция равна квадрату нормы суммы элементов подмножества, деленному на мощность искомого подмножества.

Такая задача имеет приложения в физике и технике, в частности при поиске подмножества сбалансированных сил. Она имеет истоки в проблеме data mining и возникает, например, в ситуации, когда требуется проверить гипотезу: содержат ли входные данные такое подмножество векторов (точек), что сумма всевозможных скалярных произведений элементов подмножества равна нулю.

Анализируются варианты постановки задачи, в которых размерность пространства является частью входа и не является частью входа. Установлено, что в первом случае задача NP-трудна в сильном смысле, а во втором — в обычном смысле даже для размерности 1. Кроме того, показано, что для этой задачи не существует полиномиального алгоритма с гарантированной оценкой точности, если P≠NP. Однако при фиксированной размерности пространства и целочисленных координатах задача разрешима с помощью псевдополиномиального алгоритма [1].

Работа поддержана грантами РНФ 16-11-10041 (результаты по сложности) и 15-11-10009 (псевдополиномиальный алгоритм).

[1] *Еремеев А. В., Кельманов А. В., Пяткин А. В.* О сложности некоторых задач оптимального суммирования // Докл. РАН, 2016. Т. 468. № 4. С. 372–375.

On searching for a vectors subset with the minimum normalized squared sum length

Eremeev Anton[1,2] eremeev@ofim.oscsbras.ru
Kel'manov Alexander[2,3] kelm@math.nsc.ru
Pyatkin Artem[2,3]⋆ artem@math.nsc.ru
[1]Omsk, Russia, OmSU n.a. F. M. Dostoevskiy
[2]Novosibirsk, Russia, IM SB RAS
[3]Novosibirsk, Russia, NSU

The complexity status of one discrete optimization problem of a subset search in a finite space of Euclidean points (vectors) is analyzed. This problem is obtained from the known one by changing the optimization direction from max to min. The objective function is equal to the squared sum length of the elements of the subset divived by the cardinality of the set.

This problem has application in physics and technics, in particular, in searching for a subset of balanced forces. It has also origins in data mining problem and arises, for instance, in verifying a hypothesis that the input data contain a subset of vectors (points) such that the sum of all scalar products of the elements of the subset is zero.

It is proved that if the dimension of the space is the part of input then the analyzed problem is strongly NP-hard and if the dimension is fixed then the problem is NP-hard even for dimension 1. It is shown that no approximation algorithm with a guaranteed performance ration exists for this problem. However, if the dimension of the space is fixed and the coordinates of the input points are integer, then it can be solved in a pseudopolynomial time [1].

This research is funded by the Russian Science Foundation, grants 16-11-10041 (complexity) and 15-11-10009 (pseudopolynomial algorithm).

[1] Eremeev, A., A. Kel'manov, and A. Pyatkin. 2016. On the complexity of some optimal summing problems. *Dokl. Math.* 93(3):286–288.

О некоторых задачах кластеризации: сложность и эффективные алгоритмы с оценками точности

Кельманов Александр Васильевич[1,2] kelm@math.nsc.ru
[1]Россия, Новосибирск, ИМ СО РАН
[2]Россия, Новосибирск, НГУ

Рассматривается несколько квадратичных евклидовых задач кластеризации на минимум, которые имеют приложения в анализе данных, машинном обучении, статистике, вычислительной геометрии. Задачи типичны для многих естественно-научных и технических приложений. Цель работы — обзор новых результатов о вычислительной сложности этих задач и об эффективных алгоритмах с оценками точности для их решения [1].

Ниже приведен список рассматриваемых задач и соответствующих этим задачам результатов.

Задача 1. *Поиск семейства непересекающихся подмножеств.*

Представлен точный псевдополиномиальный алгоритм для одного случая задачи.

Задача 2. *Сбалансированная* 2*-кластеризация.*

Задача 3. *Поиск подпоследовательности.*

Представлены полностью полиномиальные аппроксимационные схемы для специальных случаев задач 2 и 3.

Задача 4. *Разбиение последовательности на кластеры.*

Задача 5. *Разбиение последовательности на кластеры с ограничениями на их мощность.*

Представлены 2-приближенные полиномиальные алгоритмы для специальных случаев задач 4 и 5.

Задача 6. *Разбиение множества на броуновские кластеры.*

Представлены некоторые результаты о вычислительной сложности задачи.

Исследование проблем 1–4 поддержано РФФИ (гранты 15-01-00462 и 16-07-00168). Исследование проблем 5 и 6 поддержано РНФ (грант 16-11-10041).

[1] *Kel'manov A.* On some clustering problems: complexity and efficient algorithms with performance garantees for their solutions // 7th Conference (International) on Optimization and Applications Proceedings. Petrovac, Montenegro, 2016.

On some clustering problems: Complexity and efficient algorithms with performance garantees

Kel'manov Alexander[1,2] kelm@math.nsc.ru

[1]Novosibirsk, Russia, IM SB RAS

[2]Novosibirsk, Russia, NSU

Some quadratic Euclidean clustering problems most closely related to data mining, machine learning, statistics, and computational geometry have been considered. Minimum is the optimization direction for all considered problems. All of these problems are typical for many natural science and technical applications.

The purpose of the paper is the review of some new results on the computational complexity of these problems and on efficient algorithms with performance guarantees for their solutions [1].

Below is the list of the considered problems and the corresponding results.

Problem 1. *Finding a family of disjoint subsets.*

An exact pseudopolynomial-time algorithm for the case of the problem is proposed.

Problem 2. *Balanced* 2*-clustering with one given center.*

Problem 3. *Finding a subsequence in a sequence.*

Fully polynomial-time approximation schemes (FPTAS) for the cases of problems 2 and 3 are proposed.

Problem 4. *Partitioning a sequence into clusters.*

Problem 5. *Partitioning a sequence into clusters with restrictions on their cardinalities.*

A 2-approximation polynomial-time algorithms for the cases of problems 4 and 5 are proposed.

Problem 6. *Partitioning a set into Brownian clusters.*

Some results on complexity of the problem are proposed.

The study of problems 1–4 was supported by the Russian Foundation for Basic Research (grants 15-01-00462 and 16-07-00168). The study of problems 5 and 6 was supported by the Russian Science Foundation (grant 16-11-10041).

[1] Kel'manov, A. 2016. On some clustering problems: Complexity and efficient algorithms with performance garantees for their solutions. *7th Conference (International) on Optimization and Applications Proceedings*. Petrovac, Montenegro.

Приближенный алгоритм для задачи разбиения последовательности на кластеры при ограничениях на их мощность

Кельманов Александр Васильевич[1,2] kelm@math.nsc.ru
Михайлова Людмила Викторовна[1] mikh@math.nsc.ru
Хамидуллин Сергей Асгадуллович[1] kham@math.nsc.ru
Хандеев Владимир Ильич[1,2]⋆ khandeev@math.nsc.ru
[1]Россия, Новосибирск, ИМ СО РАН
[2]Россия, Новосибирск, НГУ

Рассматривается NP-трудная в сильном смысле [1]

Задача. *Дано*: последовательность $\mathcal{Y} = (y_1, \ldots, y_N)$ точек из $\mathbb{R}^q$, натуральные числа $T_{\min}$, $T_{\max}$, L и M. *Найти*: непустые непересекающиеся подмножества $\mathcal{M}_1, \ldots, \mathcal{M}_L$ множества $\mathcal{N} = \{1, \ldots, N\}$ номеров элементов последовательности $\mathcal{Y}$ такие, что

$$\sum_{l=1}^{L} \sum_{j \in \mathcal{M}_l} \|y_j - \overline{y}(\mathcal{M}_l)\|^2 + \sum_{i \in \mathcal{N} \setminus \mathcal{M}} \|y_i\|^2 \longrightarrow \min,$$

где $\mathcal{M} = \cup_{l=1}^{L} \mathcal{M}_l$, $\overline{y}(\mathcal{M}_l) = (1/|\mathcal{M}_l|) \sum_{j \in \mathcal{M}_l} y_j$, при ограничениях: (1) мощность объединенного множества $\mathcal{M}$ равна M; (2) в последовательности, образованной конкатенацией множеств $\mathcal{M}_1, \ldots, \mathcal{M}_L$, номера упорядочены по возрастанию; (3) номера из объединенного набора $\mathcal{M} = \{n_1, \ldots, n_M\}$ связаны неравенствами $T_{\min} \leqslant n_m - n_{m-1} \leqslant T_{\max} \leqslant N$, $m = 2, \ldots, M$. Задача актуальна, в частности, для помехоустойчивого анализа и распознавания сигналов. В работе построен 2-приближенный алгоритм, время работы которого равно $\mathcal{O}(LN^{L+1}(MN + q))$. При фиксированном числе L алгоритм полиномиален. Ранее существовал 2-приближенный полиномиальный алгоритм лишь для частного случая задачи, когда $L = 1$; временна́я сложность этого алгоритма равна $\mathcal{O}(N^2(MN + q))$.

Работа поддержана грантом РНФ № 16-11-10041.

[1] *Кельманов А. В., Михайлова Л. В., Хамидуллин С. А., Хандеев В. И.* Приближенный алгоритм для задачи разбиения последовательности на кластеры с ограничениями на их мощность // Труды ИММ УрО РАН, 2016 (в печати). Т. 22. № 3.

An approximation algorithm for one NP-hard problem of partitioning a sequence into clusters with restrictions on their cardinalities

Kel'manov Alexander[1,2] kelm@math.nsc.ru
Mikhailova Ludmila[1] mikh@math.nsc.ru
Khamidullin Sergey[1] kham@math.nsc.ru
Khandeev Vladimir[1,2]⋆ khandeev@math.nsc.ru
[1]Novosibirsk, Russia, IM SB RAS
[2]Novosibirsk, Russia, NGU

The authors consider the following strongly NP-hard [1]

Problem. *Given* a sequence $\mathcal{Y} = (y_1, \ldots, y_N)$ of points from $\mathbb{R}^q$ and some positive integers $T_{\min}$, $T_{\max}$, L, and M. *Find* nonempty disjoint subsets $\mathcal{M}_1, \ldots, \mathcal{M}_L$ of $\mathcal{N} = \{1, \ldots, N\}$ such that

$$\sum_{l=1}^{L} \sum_{j \in \mathcal{M}_l} \|y_j - \overline{y}(\mathcal{M}_l)\|^2 + \sum_{i \in \mathcal{N} \setminus \mathcal{M}} \|y_i\|^2 \longrightarrow \min$$

where $\mathcal{M} = \bigcup_{l=1}^{L} \mathcal{M}_l$ and $\overline{y}(\mathcal{M}_l) = \frac{1}{|\mathcal{M}_l|} \sum_{j \in \mathcal{M}_l} y_j$, under the following constraints: (i) the cardinality of $\mathcal{M}$ is equal to M; (ii) concatenation of elements of subsets $\mathcal{M}_1, \ldots, \mathcal{M}_L$ is an increasing sequence, provided that the elements of each subset are in ascending order; and (iii) the following inequalities for the elements of $\mathcal{M} = \{n_1, \ldots, n_M\}$ are satisfied: $T_{\min} \leqslant n_m - n_{m-1} \leqslant T_{\max} \leqslant N$, $m = 2, \ldots, M$.

This problem is relevant, for example, for noise-proof analysis and recognition of signals. In this work, a 2-approximation algorithm which runs in $\mathcal{O}(LN^{L+1}(MN + q))$-time is proposed. Earlier, there was a 2-approximate polynomial-time algorithm only for the case of the problem when $L = 1$; the running time of this algorithm is $\mathcal{O}(N^2(MN + q))$.

This research is funded by the Russian Science Foundation, grant 16-11-10041.

[1] Kel'manov, A., L. Mikhailova, S. Khamidullin, and V. Khandeev. 2016 (in press). An approximation algorithm for one NP-hard problem of partitioning a sequence into clusters with restrictions on their cardinalities. *Trudy IMM UrO RAN* 22(3).

Приближенный алгоритм для задачи разбиения последовательности на кластеры

Кельманов Александр Васильевич[1,2] kelm@math.nsc.ru
Михайлова Людмила Викторовна[1] mikh@math.nsc.ru
Хамидуллин Сергей Асгадуллович[1]⋆ kham@math.nsc.ru
Хандеев Владимир Ильич[1,2] khandeev@math.nsc.ru
[1]Россия, Новосибирск, ИМ СО РАН
[2]Россия, Новосибирск, НГУ

Рассматривается NP-трудная в сильном смысле [1]

Задача. *Дано:* последовательность $\mathcal{Y} = (y_1, \ldots, y_N)$ точек из $\mathbb{R}^q$, натуральные числа $T_{\min}$, $T_{\max}$ и L. *Найти:* непустые непересекающиеся подмножества $\mathcal{M}_1, \ldots, \mathcal{M}_L$ множества $\mathcal{N} = \{1, \ldots, N\}$ номеров элементов последовательности $\mathcal{Y}$ такие, что

$$\sum_{l=1}^{L} \sum_{j \in \mathcal{M}_l} \|y_j - \overline{y}(\mathcal{M}_l)\|^2 + \sum_{i \in \mathcal{N} \setminus \mathcal{M}} \|y_i\|^2 \to \min,$$

где $\mathcal{M} = \cup_{l=1}^{L} \mathcal{M}_l$, $\overline{y}(\mathcal{M}_l) = (1/|\mathcal{M}_l|) \sum_{j \in \mathcal{M}_l} y_j$, при ограничениях: (1) в последовательности, образованной конкатенацией множеств $\mathcal{M}_1, \ldots, \mathcal{M}_L$, номера упорядочены по возрастанию; (2) номера из объединенного набора $\mathcal{M} = \{n_1, \ldots, n_M\}$ связаны неравенствами $T_{\min} \leqslant n_m - n_{m-1} \leqslant T_{\max} \leqslant N$, $m = 2, \ldots, M$, где мощность M набора $\mathcal{M}$ предполагается неизвестной.

Рассматриваемая задача индуцируется, в частности, проблемами помехоустойчивого анализа временны́х рядов.

В настоящей работе предложен алгоритм, позволяющий находить 2-приближенное решение за время $\mathcal{O}(LN^{L+1}(N+q))$, полиномиальное при фиксированном числе L. Ранее существовал 2-приближенный полиномиальный алгоритм лишь для частного случая задачи, когда $L = 1$; временна́я сложность этого алгоритма равна $\mathcal{O}(N^2(N+q))$.

Работа поддержана грантами РФФИ № 15-01-00462, № 16-31-00186-мол-а и № 16-07-00168.

[1] *Кельманов А. В., Михайлова Л. В., Хамидуллин С. А., Хандеев В. И.* Приближенный алгоритм для задачи разбиения последовательности на кластеры // Ж. вычисл. мат. мат. физ., 2016 (в печати).

An approximation algorithm for a problem of partitioning a sequence into clusters

Kel'manov Alexander[1,2] kelm@math.nsc.ru
Mikhailova Ludmila[1] mikh@math.nsc.ru
Khamidullin Sergey[1]⋆ kham@math.nsc.ru
Khandeev Vladimir[1,2] khandeev@math.nsc.ru
[1]Novosibirsk, Russia, IM SB RAS
[2]Novosibirsk, Russia, NSU

The authors consider the following strongly NP-hard [1]

Problem. *Given* a sequence $\mathcal{Y} = (y_1, \ldots, y_N)$ of points from $\mathbb{R}^q$ and some positive integers $T_{\min}$, $T_{\max}$, and L. *Find* nonempty disjoint subsets $\mathcal{M}_1, \ldots, \mathcal{M}_L$ of $\mathcal{N} = \{1, \ldots, N\}$ such that

$$\sum_{l=1}^{L} \sum_{j \in \mathcal{M}_l} \|y_j - \overline{y}(\mathcal{M}_l)\|^2 + \sum_{i \in \mathcal{N} \setminus \mathcal{M}} \|y_i\|^2 \to \min$$

where $\mathcal{M} = \cup_{l=1}^{L} \mathcal{M}_l$ and $\overline{y}(\mathcal{M}_l) = (1/|\mathcal{M}_l|) \sum_{j \in \mathcal{M}_l} y_j$ under the following constraints: (i) concatenation of elements of subsets $\mathcal{M}_1, \ldots, \mathcal{M}_L$ is an increasing sequence, provided that the elements of each subset are in ascending order; and (ii) the following inequalities for the elements of $\mathcal{M} = \{n_1, \ldots, n_M\}$ are satisfied: $T_{\min} \leqslant n_m - n_{m-1} \leqslant T_{\max} \leqslant N$, $m = 2, \ldots, M$ (the cardinality M of $\mathcal{M}$ assumed to be unknown).

The problem is related, in particular, to the noise-proof analysis of time series.

In this work, the authors present an algorithm that allows to find a 2-approximate solution of the problem in $\mathcal{O}(LN^{L+1}(N+q))$-time, which is polynomial if L is fixed. Earlier, a 2-approximation polynomial-time algorithm having $\mathcal{O}(N^2(N+q))$ running time was presented only for the case of the problem when $L = 1$.

This research is funded by the Russian Foundation for Basic Research, grants 15-01-00462, 16-31-00186-mol-a, and 16-07-00168.

[1] Kel'manov, A., L. Mikhailova, S. Khamidullin, and V. Khandeev. 2016 (in press). An approximation algorithm for a problem of partitioning a sequence into clusters. *Comput. Math. Math. Phys.*

Аппроксимационная схема для задачи сбалансированной 2-кластеризации при ограничениях на мощность кластеров

Кельманов Александр Васильевич[1,2] kelm@math.nsc.ru
Моткова Анна Владимировна[2]⋆ anitamo@mail.ru
[1]Россия, Новосибирск, ИМ СО РАН
[2]Россия, Новосибирск, НГУ

Рассматривается NP-трудная в сильном смысле [1]

Задача. *Дано*: множество $\mathcal{Y} = \{y_1, \ldots, y_N\}$ точек из $\mathbb{R}^q$ и натуральное число M. *Найти* разбиение множества $\mathcal{Y}$ на кластеры $\mathcal{C}$ и $\mathcal{Y} \setminus \mathcal{C}$ такие, что

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

где $\overline{y}(\mathcal{C}) = (1/|\mathcal{C}|) \sum_{y \in \mathcal{C}} y$ — геометрический центр (центроид) $\mathcal{C}$ при ограничении $|\mathcal{C}| = M$.

Исследование мотивировано слабой изученностью задачи в алгоритмическом плане и ее актуальностью для многих приложений, среди которых, в частности, проблемы кластерного анализа данных, проблемы интерпретации данных, проблемы геометрии, статистические проблемы совместного оценивания и проверки гипотез по неоднородным выборкам и др. В настоящей работе построен приближенный алгоритм решения задачи. Для заданной относительной погрешности ε этот алгоритм позволяет находить $(1 + \varepsilon)$-приближенное решение за время $\mathcal{O}\left(qN^2(\sqrt{2q/\varepsilon} + 1)^q\right)$. В случае, когда размерность q пространства ограничена константой, время работы алгоритма оценивается величиной $\mathcal{O}\left(N^2 (1/\varepsilon)^{q/2}\right)$ и он реализует полностью полиномиальную аппроксимационную схему (FPTAS — fully polynomial-time approximation scheme). Работа поддержана грантами РФФИ №№ 15-01-00462, 16-31-00186-mol-a и 16-07-00168.

[1] *Кельманов А. В., Моткова А. В.* Аппроксимационная схема для квадратичной евклидовой задачи сбалансированной 2-кластеризации // Дискретный анализ и исследование операций, 2016 (в печати).

An approximation scheme for a balanced 2-clustering with restrictions on the cardinalities of clusters

Kel'manov Alexander[1,2] kelm@math.nsc.ru
Motkova Anna[2]⋆ anitamo@mail.ru
[1]Novosibirsk, Russia, IM SB RAS
[2]Novosibirsk, Russia, NSU

The authors consider the following strongly NP-hard [1]

Problem. *Given* a set $\mathcal{Y} = \{y_1, \ldots, y_N\}$ of points from $\mathbb{R}^q$ and a positive integer M. *Find* a partition of $\mathcal{Y}$ into two nonempty clusters $\mathcal{C}$ and $\mathcal{Y} \setminus \mathcal{C}$ such that

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

where $\overline{y}(\mathcal{C}) = (1/|\mathcal{C}|) \sum_{y \in \mathcal{C}} y$ is the geometric center (centroid) of $\mathcal{C}$, subject to constraint $|\mathcal{C}| = M$.

The research is motivated by insufficient study of the problem from an algorithmic direction and its importance in some applications including geometry, cluster analysis, statistical problems of joint evaluation and hypotheses testing with heterogeneous samples, data interpretation problem, etc.

In this work, the authors present an approximation algorithm that allows to find a $(1+\varepsilon)$-approximate solution in $\mathcal{O}(qN^2(\sqrt{2q/\varepsilon}+1)^q)$ time for a given relative error ε. If the space dimension is bounded by a constant, this algorithm implements a fully polynomial-time approximation scheme (FPTAS).

This research is funded by the Russian Foundation for Basic Research, grants 15-01-00462, 16-31-00186-mol-a, and 16-07-00168.

[1] Kel'manov, A., and A. Motkova. 2016 (in press). An approximation scheme for a quadratic euclidean balanced 2-clustering problem. *J. Appl. Ind. Math.*

Приближенная схема для задачи поиска подпоследовательности

Кельманов Александр Васильевич[1,2] kelm@math.nsc.ru
Романченко Семен Михайлович[1]⋆ rsm@math.nsc.ru
Хамидуллин Сергей Асгадуллович[1] kham@math.nsc.ru
[1]Россия, Новосибирск, ИМ СО РАН
[2]Россия, Новосибирск, НГУ

Рассматривается NP-трудная в сильном смысле [1]

Задача. *Дано*: последовательность $\mathcal{Y} = (y_1, \ldots, y_N)$ точек из $\mathbb{R}^q$, натуральные числа $T_{\min}$, $T_{\max}$ и M. *Найти*: подмножество $\mathcal{M} = \{n_1, \ldots, n_M\} \subseteq \{1, \ldots, N\}$ такое, что

$$\sum_{j \in \mathcal{M}} \|y_j - \overline{y}(\mathcal{M})\|^2 \to \min,$$

где $\overline{y}(\mathcal{M}) = (1/|\mathcal{M}|) \sum_{i \in \mathcal{M}} y_i$, при ограничениях

$$1 \leqslant T_{\min} \leqslant n_m - n_{m-1} \leqslant T_{\max} \leqslant N,\ m = 2, \ldots, M,$$

на элементы из набора $(n_1, \ldots, n_M)$.

Задача актуальна, в частности, для помехоустойчивого анализа и распознавания сигналов.

В работе построен приближенный алгоритм, который при заданной относительной погрешности ε позволяет находить $(1+\varepsilon)$-приближенное решение задачи за время $\mathcal{O}(N^2(M(T_{\max} - T_{\min} + +1)+q)(\sqrt{2q/\varepsilon}+1)^q)$. При фиксированной размерности пространства временная сложность алгоритма равна $\mathcal{O}(MN^3(1/\varepsilon)^{q/2})$ и он реализует полностью полиномиальную приближенную схему (FPTAS).

Работа поддержана грантами РФФИ № 15-01-00462, № 16-31-00186-мол-а и № 16-07-00168.

[1] *Kel'manov A., Romanchenko S., Khamidullin S.* Fully polynomial-time approximation scheme for a problem of finding a subsequence // 9th Conference (International) on Discrete Optimization and Operations Research Proceedings. Vladivostok, Russky Island, Russia, 2016 (in press).

An approximation scheme for a problem of finding a subsequence

Kel'manov Alexander[1,2] kelm@math.nsc.ru
Romanchenko Semyon[1]⋆ rsm@math.nsc.ru
Khamidullin Sergey[1] kham@math.nsc.ru
[1]Novosibirsk, Russia, IM SB RAS
[2]Novosibirsk, Russia, NSU

The authors consider the following strongly NP-hard [1]

Problem. Given a sequence $\mathcal{Y} = (y_1, \ldots, y_N)$ of points from $\mathbb{R}^q$, and some positive integer numbers $T_{\min}$, $T_{\max}$ and M. Find a subset $\mathcal{M} = \{n_1, \ldots, n_M\} \subseteq \{1, \ldots, N\}$ such that

$$\sum_{j \in \mathcal{M}} \|y_j - \overline{y}(\mathcal{M})\|^2 \to \min,$$

where $\overline{y}(\mathcal{M}) = (1/|\mathcal{M}|) \sum_{i \in \mathcal{M}} y_i$, under constraints

$$1 \leqslant T_{\min} \leqslant n_m - n_{m-1} \leqslant T_{\max} \leqslant N,\ m = 2, \ldots, M,$$

on the elements of $(n_1, \ldots, n_M)$.

This problem is relevant, for example, for noise-proof analysis and recognition of signals.

The main result of this work is an approximation algorithm which allows to find a $(1 + \varepsilon)$-approximate solution for arbitrary relative error ε in $\mathcal{O}(N^2(M(T_{\max} - T_{\min} + 1) + q)(\sqrt{2q/\varepsilon} + 1)^q)$ time. If the dimension q of the space is fixed, then the time complexity of the algorithm is equal to $\mathcal{O}(MN^3(1/\varepsilon)^{q/2})$, and it implements a fully polynomial-time approximation scheme (FPTAS).

This research is funded by the Russian Foundation for Basic Research, grants 15-01-00462, 16-31-00186-mol-a, and 16-07-00168.

[1] Kel'manov, A., S. Romanchenko, and S. Khamidullin. 2016 (in press). Fully polynomial-time approximation scheme for a problem of finding a subsequence. *9th Conference (International) on Discrete Optimization and Operations Research Proceedings*. Vladivostok, Russky Island, Russia.

Об одном подходе к доказательству фасетности опорных неравенств

Симанчев Руслан Юрьевич[1,2]⋆ osiman@rambler.ru
Уразова Инна Владимировна[2] urazovainn@mail.ru
[1]Россия, Омск, ОНЦ СО РАН
[2]Россия, Омск, ОмГУ

Пусть $P \subset R^n$ — выпуклый многогранник. Линейное неравенство называется опорным к многограннику P, если оно выполняется для любой точки из P и хотя бы для одной точки из P оно выполняется как равенство. Всякое опорное неравенство порождает грань многогранника. Грани, размерности которых на единицу меньше размерности самого многогранника, называются фасетными. Неравенство, порождающее фасету многогранника, называется фасетным. Важность фасетных неравенств обусловливается их применением в алгоритмах отсечения при решении оптимизационных задач. Как показывают экспериментальные результаты, такие отсечения являются наиболее «сильными». Кроме того, каждое фасетное неравенство (с точностью до эквивалентности) присутствует в любом линейном описании выпуклого многогранника.

В связи с вышесказанным становится актуальным вопрос о подходах к поиску фасетных неравенств. Как правило, доказательство фасетности существенно опирается на комбинаторные свойства многогранника. В настоящей работе обсуждается подход к доказательству фасетности опорного неравенства, основанный на комбинаторных свойствах множества допустимых решений оптимизационной задачи. Получены ряд необходимых условий и достаточное условие фасетности опорного неравенства в терминах допустимого множества задачи. Данный подход оказался эффективным при анализе многогранника задачи о минимальном связном k-факторе и многогранника задачи аппроксимации графа [1].

[1] *Simanchev R. Yu., Urazova I. V.* On the facets of combinatorial polytopes. — Lecture notes in computer science ser. — Springer, 2016 (in press).

An approach to the proof of facetness of support inequalities

Simanchev Ruslan[1,2]⋆ osiman@rambler.ru
Urazova Inna[2] urazovainn@mail.ru
[1]Omsk, Russia, OSC SB RAS
[2]Omsk, Russia, OmSU n.a. F. M. Dostoevskiy

Let $P \subset R^n$ is a convex polytope. Linear inequality is called a support to the polytope P, if it is true for any point from P and for at least one point of P it holds as an equality. Any support inequality generates a face of the polytope. Faces with dimension that to one less than the dimension of the polytope are called the facets. Inequality that generates facet of a polytope is called the facet inequality. Importance of facet inequalities is caused by their use in cutting plane algorithms for solving optimization problems. As shown by the experimental results, such cutting planes are more strong. In addition, each facet inequality is presented in any linear description of a convex polytope (up to equivalence).

In connection with the above, the question of the approach to finding facet inequalities becomes actual. Usually, facetness proof is essentially based on the combinatorial properties of the polytope. The approach to the proof of facetness of support inequalities, based on the combinatorial properties of the feasible solutions set of optimization problem, is discussed. A number of necessary conditions and sufficient condition for the facetness of support inequality in terms of the feasible set are obtained. This approach to the analysis of the minimum connected k-factor problem polytope and graph approximation problem polytope was effective [1].

[1] Simanchev, R. Yu., and I. V. Urazova. 2016 (in press). On the facets of combinatorial polytopes. Lecture notes in computer science ser. Springer.

Методы обработки космических изображений для оценки эмиссий малых газовых компонент и аэрозолей при природных пожарах

Бондур Валерий Григорьевич[1]⋆ vgbondur@aerocosmos.info
Мурынин Александр Борисович[1,2] amurynin@bk.ru
Гордо Кристина Араратовна[1] office@aerocosmos.info
[1]Россия, Москва, НИИ «Аэрокосмос»
[2]Россия, Москва, ФИЦ ИУ РАН

Природные пожары являются глобальным источником газовой и аэрозольной эмиссии в атмосферу и считаются одним из основных факторов экологического риска для окружающей среды. Они вызывают дисбаланс климатической системы, в том числе приводя к повышению температуры поверхности планеты, парниковому эффекту и изменению радиационного баланса атмосферы. Учитывая территориальные особенности и характерный пространственный масштаб Северной Евразии, для выявления очагов пожаров и расчета объемов эмиссий вредных газов и аэрозолей в воздушную среду требуется применение методов и средств спутникового мониторинга. В настоящей работе описана система оперативного космического мониторинга и методы обработки больших потоков космических изображений, используемые для обнаружения очагов возгорания, и для оценки их последствий. Приводятся результаты космического мониторинга природных пожаров и оценены площади, пройденные огнем, а также объемы эмиссий углеводородсодержащих газов (CO_2, CO) и аэрозолей (PM2.5) для различных регионов в различные месяцы за период с 2010 по 2016 гг. Выявлены особенности сезонной повторяемости природных пожаров и объемов эмиссий вредных газов и мелкодисперсных аэрозолей на исследуемых территориях.

Работа поддержана Минобрнауки России (уникальный идентификатор проекта RFMEFI58314X0003).

[1] *Бондур В. Г.* Космический мониторинг эмиссий малых газовых компонент и аэрозолей при природных пожарах в России // Исследование Земли из космоса, 2015. №6. С. 21–35.

Satellite imagery processing methods for the estimation of trace gas and aerosol emissions due to wildfires

Bondur Valery[1]⋆ vgbondur@aerocosmos.info
Murynin Alexander[1,2] amurynin@bk.ru
Gordo Kristina[1] office@aerocosmos.info
[1]Moscow, Russia, ISR "Aerocosmos"
[2]Moscow, Russia, FRC CSC RAS

Wildfires are global sources of gas and aerosol emissions into the atmosphere and considered as one of the main factors of environmental risk. Wildfires cause disbalance of the climate system, including raised temperatures of the planet's surface, greenhouse effect, and change in the atmosphere's radiation balance. Taking into account terrain features and characteristic spatial scale of North Eurasia, detection of fire origins and calculation of volumes of harmful gas and aerosol emissions into the atmosphere requires application of satellite monitoring methods and assets. This paper describes the system of online satellite monitoring and methods of processing of large satellite imagery flows used to detect fire origins and to assess fire effects. The results of wildfire satellite monitoring and assessments of burned out areas as well as carbon bearing gas (CO_2, CO) and aerosol (PM2.5) emission volumes are given here for different regions in various months between 2010 and 2016. The features of wildfire seasonal repetition and harmful gas and fine aerosol emission volumes for the areas of interest have been revealed [1].

The research is supported by the Ministry of Education and Science of the Russian Federation (unique project identifier RFMEFI58314X0003).

[1] Bondur, V. 2015. Space-borne monitoring of trace gas and aerosol emissions during wildfires in Russia. *Issledovanie Zemli iz Cosmosa* 6:21–35.

Устранение комбинированного шума на растровых изображениях

Двоенко Сергей Данилович[1]⋆ dsd@tsu.tula.ru
Данг Нгок Хоанг Тхань[1,2] myhoangthanh@yahoo.com
[1]Россия, Тула, ТулГУ
[2]Вьетнам, Хюэ, Хюэсский индустриальный колледж

Реальные шумы можно эффективно смоделировать смесью распределений, например, пуассоновского и гауссовского шумов. Смесь таких шумов обычно наблюдается на изображениях с электронного микроскопа, на аэрокосмических снимках и др.

Для устранения смеси пуассон-гауссовских шумов разработаны различные методы, например: масштабного градиента, альтернативной минимизации, PURE-LET, обобщенного преобразования Энскомба. Общая характеристика перечисленных методов заключается в том, что они достаточно сложны: как теоретически, так и в практической реализации. Известно, что повышенная сложность теоретически строго обоснованных моделей устранения смеси пуассон-гауссовских шумов обычно приводит к тому, что они оказываются многопараметрическими. Эта особенность часто способна заметно снизить качество обработки изображений в условиях, когда не удается хорошо оценить параметры модели.

Альтернативой могут быть упрощенные модели устранения смеси таких шумов, которые позволяют построить простые алгоритмы при сохранении высокого качества обработки.

Рассмотрена задача устранения комбинированного пуассон-гауссовского шума в предположении, что параметры распределений уже идентифицированы и оцениваются лишь пропорции их вкладов в общий шум. Предложенная модель устранения шума основана на полной вариации функции яркости изображения [1].

Работа частично поддержана грантами РФФИ №16-07-01039 и №15-07-02228.

[1] *Thanh D. N. H., Dvoenko S. D.* A method of total variation to remove the mixed Poisson–Gaussian noise // Pattern Recogn. Image Anal., 2016. Vol. 26. No. 2. P. 285–293. doi: 10.1134/S1054661816020231.

Removing of combined noise in raster images

Dvoenko Sergey[1]⋆ dsd@tsu.tula.ru
Dang Thanh[1,2] myhoangthanh@yahoo.com
[1]Tula, Russia, TulSU
[2]Hue, Vietnam, Hue Industrial College

Real noise can be effectively represented by a mixture of distributions, for example, by Poisson and Gaussian noises. Such mixture of noises is usually arised in electronic microscopy, aerospace images, etc.

Different techniques have been developed to remove the Poisson–Gaussian noise mixture, such as: scaled gradient, alternating minimization, PURE-LET, and generalized Anscombe transformation. The general characteristics of these methods is that all of them are sufficiently complicated, both in theoretical and in practical sense.

It is known that the increased complexity of the strong theoretically based models to remove the Poisson–Gaussian noise mixture usually results in multiple model parameters. This peculiarity can significantly reduce the quality of image processing under the condition when the model parameters have not been properly evaluated.

The alternative consists in simplified models of the noise mixture removing to provide simplified algorithms providing the data processing high quality.

The problem of the combined Poisson-Gaussian noise removing is under investigation with the assumption that the distribution parameters are already identified and only the ratios of them in the resulted noise need to be evaluated. The proposed denoising model is based on the total variation of the image intensity function.

This research is funded partially by the Russian Foundation for Basic Research, grants Nos. 16-07-01039 and 15-07-02228.

[1] Thanh, D. N. H., and S. D. Dvoenko. 2016. A method of total variation to remove the mixed Poisson–Gaussian noise. *Pattern Recogn. Image Anal.* 26(2):285–293. doi: 10.1134/S1054661816020231.

Оценка качества изображений при повышении разрешения на основе пространственного спектрального синтеза

Игнатьев Владимир Юрьевич[1,2]⋆ vladimir.ignatiev.mipt@gmail.com
Матвеев Иван Алекссевич[1] ivanmatveev@mail.ru
Мурынин Александр Борисович[1,2] amurynin@bk.ru
Трекин Алексей Николаевич[2] alexey.trekin@gmail.com
[1]Россия, Москва, ФИЦ ИУ РАН
[2]Россия, Москва, НИИ "Аэрокосмос"

Рассмотрены два метода улучшения изображений с использованием спектральных представлений. Первый подход основан на предположении, что доступна информация о деталях высокого пространственного разрешения, задаваемая дополнительным опорным изображением. Второй подход не требует привлечения дополнительной информации. Изображение высокого разрешения синтезируется на основе аналитического продолжения спектра исходного изображения в область высоких пространственных частот. Проведено исследование по выбору численной меры сходства (различия) изображений в задаче оценки качества повышения пространственного разрешения с помощью разработанных методов. Получены результаты по поиску оптимальных параметров спектрального синтеза в заданном пространственном разрешении. Сравниваются результаты оценки качества изображений, улучшенных с помощью интерполяции Ланцоша, и разработанными методами с оптимальными параметрами [1].

Работа выполнена при поддержке Министерства образования и науки Российской Федерации в рамках ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2014-2020 годы» (уникальный идентификатор проекта RFMEFI57414X0086) Работа поддержана грантами РФФИ № 14-05-91759 и № 16-51-55019.

[1] *Мурынин А. Б., Трекин А. Н., Игнатьев В. Ю., Матвеев И. А.* Оценка качества изображений при повышении разрешения на основе пространственного спектрального синтеза // Вестник МГТУ им. Н. Э. Баумана. Серия «Естественные науки». — М.:МГТУ им. Н. Э. Баумана, 2016 (в печати).

Image quality evaluation for resampling methods based on spatial spectrum extrapolation

Ignatiev Vladimir[1,2]⋆ vladimir.ignatiev.mipt@gmail.com
Matveev Ivan[1] ivanmatveev@mail.ru
Murynin Alexander[1,2] amurynin@bk.ru
Trekin Alexey[2] alexey.trekin@gmail.com
[1]Moscow, Russia, FRC CSC RAS
[2]Moscow, Russia, ISR "Aerocosmos"

Two methods of image enhancement using spectral representations are considered. The first approach is based on the assumption that the required information about details is obtained from the additional reference image at the high spatial resolution. High-resolution image is constructed using a combination of spatial spectra of the main and reference images. The second approach does not require the use of additional external information (reference image). High-resolution image is synthesized based on the analytic continuation of the original image spectrum to the region of high spatial frequencies. A study on the selection of a numerical measure of image similarity (difference) in the quality assessment problem is carried out. The results on the finding optimal parameters of spectral synthesis at a given spatial resolution are presented. The results of evaluation of image quality enhanced by Lanczos interpolation are compared with the developed methods with the optimal settings [1].

This research is supported by the Russian Foundation for Basic Research, grants Nos. 14-05-91759 and 16-51-55019. This research is funded by the Russian Ministry of Education and Science (Project identificator RFMEFI57414X0086).

[1] Murynin, A., A. Trekin, V. Ignatiev, and I. Matveev. 2016 (in press). Image quality evaluation for resampling methods based on spatial spectrum extrapolation. *Herald of the Bauman Moscow State Technical University*. Natural sciences ser. Moscow: Bauman University Publishing House.

Поиск отличий на последовательностях изображений в сложных сценах

Вишняков Борис Ваисович*⋆* vishnyakov@gosniias.ru
Сидякин Сергей Владимирович sersid@gosniias.ru
Рослов Николай Игоревич nroslov@gosniias.ru
Визильтер Юрий Валентинович viz@gosniias.ru
Россия, Москва, ФГУП «ГосНИИАС»

За последнее десятилетие было предложено большое число методов и подходов, посвященных проблеме поиска отличий на видео. Подавляющее большинство среди них было разработано для обнаружения движущихся объектов в системах видеонаблюдения и основывалось на техниках вычитания фона для сегментации сцены на передний план и фон. При этом каждый метод на практике сталкивается с серьезными и практически всегда нерешаемыми проблемами: высокая вариативность условий съемки, постоянные изменения освещенности разной продолжительности, шумы видеодатчиков, наличии динамических объектов заднего плана, например раскачивающихся веток, кустов.

В работе предлагается новый подход к детектированию отличий на последовательностях изображений в сложных условиях съемки. Этот подход основывается на взаимных компаративных фильтрах и нормализации фона. Основные преимущества предлагаемого алгоритма — отсутствие подсчета преобразования яркостей пикселов и предварительной сегментации изображения на области постоянной яркости. Предложенный подход может быть скомбинирован с методом моделирования фона, неустойчивым к изменению освещенности, с целью повышения качества работы последнего. В работе приведены показатели качества разработанного метода на общедоступной базе GTILT [1].

Работа поддержана грантами РФФИ №№ 15-07-09362 А, 15-07-01323 А и 16-57-52042 МНТ_а.

[1] *Вишняков Б. В., Сидякин С. В., Рослов Н. И., Визильтер Ю. В.* Поиск отличий на последовательностях изображений в сложных сценах // Вестник компьютерных и информационных технологий. — М.: Спектр, 2016 (в печати).

Change detection in the sequences of images in complex scenes

Vishnyakov Boris⋆ vishnyakov@gosniias.ru
Sidyakin Sergey sersid@gosniias.ru
Roslov Nikolay nroslov@gosniias.ru
Vizilter Yuri viz@gosniias.ru

Moscow, Russia, FGUP "GosNIIAS"

Over the last decade, a large number of methods and approaches for dealing with the problem of finding changes in the video were proposed. The vast majority of them have been developed for moving objects detection in video surveillance systems and rely on background subtraction techniques for segmentation of the scene into the foreground and the background. In addition, each method faces serious and sometimes unsolvable problems: the high variation of the record shooting conditions, constant illumination changes of different duration, matrix noise, the presence of dynamic objects in the background, for example, swaying branches, bushes.

The present authors propose a new approach to the detection of changes in the sequences of images in difficult shooting conditions. This approach is based on mutual comparative filters and normalizing the background. The proposed algorithm of allocation of basic differences is simple and efficient. Its main advantage is that it does not use the mapping of pixel brightness, which is often used by other methods, and does not require prior segmentation into regions of constant brightness. The proposed approach can be combined with almost any method of background modeling that is not robust to changing light conditions to improve the overall quality. The quality levels of the developed method have been calculated and compared with the baseline quality levels of the basic methods of background modeling or change detection using the publicly available dataset GTILT [1].

This research is funded by the Russian Foundation for Basic Research, grants Nos. 15-07-09362 A, 15-07-01323 A, and 16-57-52042 MNT_a.

[1] Vishnyakov, B., S. Sidyakin, N. Roslov, and Yu. Vizilter. 2016 (in press). Change detection for the sequences of images in complex scenes. *Herald of computer and information technologies*. Moscow: Spektr.

Идентификация лиц в реальном времени с использованием верточной нейронной сети и хэширующего леса

Горбацевич Владимир Сергеевич* gvs@gosniias.ru
Визильтер Юрий Валентинович viz@gosniias.ru
Воротников Андрей Валерьевич andronzord@gmail.com
Костромов Никита Алексеевич nikita-kostromov@yandex.ru
Россия, Москва, ФГУП «ГосНИИАС»

В работе [1] предлагается метод построения биометрического шаблона с использованием сверточной нейронной сети и хэширующего леса (CNHF — convolutional network with hashing forest). Метод состоит из двух этапов: на первом происходит обучение сверточной нейронной сети (CNN — convolutional neural network), далее к полученным дескрипторам применяется хеширующее преобразование с использованием нового предложенного метода хэширующего леса (BHF — boosted hashing forest). Метод BHF является обобщением метода Boosted SSC (Similarity Sensitive Coding) для решения задачи построения оптимального хеша, учитывающего специфику одновременно задач верификации и идентификации лиц. Метод CNHF обучен на базе лиц CASIA-WebFace, а его тестирование произведено на базе лиц LFW (Labeled Faces in the Wild). Для вложения Хэмминга (CBHF) полученный 200-битный (25 байт) биометрический шаблон показывает качество верификации 0,963, 2000-битный шаблон — 0,9814 на LFW. Метод CNHF с 7-битными деревьями 2000×7 достигает уровня 0,93 в идентификации (rank-1) относительно базовых результатов CNN в 0,899. Метод CNHF формирует описания лиц с частотой 40 fps на CPU Core i7 и 120 fps с использованием GPU GeForce GTX 650.

Работа поддержана грантом РНФ (проект № 16-11-00082).

[1] *Визильтер Ю. В., Горбацевич В. С., Воротников А. В., Костромов Н. А.* Идентификация лиц в реальном времени с использованием сверточной нейронной сети и хэширующего леса // ВКИТ, 2016 (в печати).

Real-time face identification via convolutional neural network and boosted hashing forest

Gorbatsevich Vladimir ⋆ gvs@gosniias.ru
Vizilter Yuri viz@gosniias.ru
Vorotnikov Andrew andronzord@gmail.com
Kostromov Nikita nikita-kostromov@yandex.ru
Moscow, Russia, FGUP "GosNIIAS"

In [1], the family of real-time face representations is obtained via Convolutional Network with Hashing Forest (CNHF). First, the convolutional neural network (CNN) is taught, then CNN is transformed to the multiple convolution architecture, and finally, the output hashing transform is taught via new Boosted Hashing Forest (BHF) technique. This BHF generalizes the Boosted SSC (Similarity Sensitive Coding) approach for hashing learning with joint optimization of biometric face verification and identification. The method CNHF is trained on CASIA-WebFace dataset and evaluated on LFW (Labeled Faces in the Wild) dataset. The output of single CNN with 0.97 is taught on LFW. For Hamming embedding, CBHF-200 bit (25 byte) code with 0.963 and 2000-bit code with 0.981 on LFW have been got. The method CNHF with 2000×7 bit hashing trees achieves 0.93 rank1 on LFW relative to basic CNN 0.899 rank1. The method CNHF generates templates at the rate of 40+ fps with CPU Core-i7 and 120+ fps with GPU GeForce GTX 650.

This work is supported by the Russian Science Foundation (project No. 16-11-00082).

[1] Vizilter, Yu., V. Gorbatsevich, A. Vorotnikov, and N. Kostromov. 2016 (in press). Real-time face identification via CNN and boosted hashing forest. *Vestnik Komp'yuternykh i Informatsionnykh Tekhnologiy* [Herald of Computer and Information Technologies).

Применение теоретико-информационного критерия качества для сегментации изображений

Мурашов Дмитрий Михайлович d_murashov@mail.ru
Россия, Москва, ФИЦ ИУ РАН

Рассматривается задача разработки метода обеспечения наилучшего качества сегментации цифровых изображений [1]. Метод ориентирован на применение модифицированного суперпиксельного алгоритма сегментации.

В известных работах для оценки качества сегментации использовался «взвешенный показатель недостоверности», вычисляемый через значения нормализованной взаимной информации цветовых каналов входного и сегментированного изображений. Зависимость показателя недостоверности от параметра алгоритма сегментации монотонна, что потребовало обучения алгоритма и разработки итерационной процедуры выбора параметра.

В данной работе в качестве критерия для оптимизации качества сегментации предлагается применять меру избыточности информации. Такой критерий обеспечивает лучший результат с точки зрения визуального восприятия. Показано, что предложенный способ построения меры избыточности позволил получить экстремальные свойства. Эксперимент, проведенный на изображениях из базы Berkeley Segmentation Dataset, подтвердил, что сегментированное изображение, соответствующее минимуму меры избыточности, дает минимальное различие по теоретико-информационной мере при сравнении с исходным изображением. Кроме того, выбранный с помощью предложенного критерия вариант сегментации дает наибольшее сходство с эталонами, имеющимися в базе.

Работа поддержана грантами РФФИ № 15-07-09324 и № 15-07-07516.

[1] *Мурашов Д. М.* Применение теоретико-информационного подхода для сегментации изображений // Машинное обучение и анализ данных, 2016 (в печати).

Application of information-theoretical performance criterion for image segmentation

Murashov Dmitry d_murashov@mail.ru
Moscow, Russia, FRC CSC RAS

A problem of segmentation quality of digital images is considered [1]. The developed technique is based on the information-theoretical approach and applied to a modified superpixel segmentation algorithm.

In one of the conventional techniques, weighted uncertainty index is used for measuring segmentation quality. The index is calculated using normalized mutual information of color channels in given and segmented images. The uncertainty index varies monotonously depending on parameter of the segmentation algorithm. It caused application of learning technique and iterative procedure for choosing parameter value.

In this work, information redundancy measure is proposed as a criterion for optimizing segmentation quality. This criterion provides the best result in terms of visual perception. It is shown that the proposed method of constructing the redundancy measure provides it with extremal properties. An experiment was conducted using the images from the Berkeley Segmentation Dataset. The experiment confirmed that the segmented image corresponding to a minimum of redundancy measure produces the minimum difference in the information-theoretical dissimilarity measure when compared with the original image. In addition, the segmented image that was selected using the proposed criteria, gives the highest similarity with the groundtruth segmentations, available in the database.

This research is funded by the Russian Foundation for Basic Research, grants Nos. 15-07-09324 and 15-07-07516.

[1] Murashov, D. 2016 (in press). Application of information-theoretical approach for image segmentation. *Machine Learning Data Anal.*

Оптимальный выбор параметров для восстановления спектров морского волнения по аэрокосмическим изображениям

Мурынин Александр Борисович[1,2] amurynin@bk.ru
Бондур Валерий Григорьевич[1]⋆ vgbondur@aerocosmos.info
Игнатьев Владимир Юрьевич[1,2] vladimir.ignatiev.mipt@gmail.com

[1]Россия, Москва, НИИ «Аэрокосмос»
[2]Россия, Москва, ФИЦ ИУ РАН

Рассматривается проблема восстановления спектров морской поверхности по аэрокосмическим изображениям в широком спектральном диапазоне длин волн. В рамках описанной нелинейной модели поля яркости, регистрируемого аппаратурой дистанционного зондирования, предложена модификация восстанавливающего оператора, действующего во всей пространственно-спектральной области. Описан итерационный процесс выбора оптимальных значений параметров модифицированного оператора с использованием подспутниковых измерений для валидации. Представлены результаты проверки работоспособности построенного оператора для различных условий регистрации изображений морской поверхности[1].

Работа выполнена при поддержке Министерства образования и науки Российской Федерации (идентификаторы проектов №2015/Н8, RFMEFI57714X0110).

[1] *Бондур В. Г., Дулов В. А., Игнатьев В. Ю., Мурынин А. Б.* Восстановление спектров морского волнения по спектрам космических изображений в широком диапазоне частот // Известия РАН. Физика атмосферы и океана, 2016 (в печати). Т. 52.

Parameters optimization in the problem of sea-wave spectra recovery by airspace images

Murynin Alexander[1,2] amurynin@bk.ru
Bondur Valery[1]⋆ vgbondur@aerocosmos.info
Ignatiev Vladimir[1,2] vladimir.ignatiev.mipt@gmail.com
[1]Moscow, Russia, ISR "Aerocosmos"
[2]Moscow, Russia, FRC CSC RAS

The paper considers the problem of the retrieving of sea surface spectra from aerospace images over a wide wavelength range. A modified recover operator defined in the whole spatiospectral domain is proposed. This operator is developed taking into consideration the nonlinear model of the brightness field recorded by remote sensing equipment. The iterative process of selecting the optimal values of the parameters of the modified operator using ground truth measurements to validate is described. The results of the performance test of the operator constructed are discussed for different conditions of the sea surface images registration [1].

This research is supported by the Russian Ministry of Education and Science (Project identificators 2015/H8, RFMEFI57714X0110).

[1] Bondur, V., V. Dulov, A. Murynin, and V. Ignatiev. 2016 (in press). Retrieving sea wave spectra using satellite imagery spectra in a wide range of frequencies. *Izv. RAS. Atmospheric and Oceanic Physics* 52.

Максимально правдоподобный поиск ближайшего соседа в интеллектуальных системах классификации изображений

Савченко Андрей Владимирович avsavchenko@hse.ru
Россия, Н. Новгород, НИУ ВШЭ

В [1, гл. 4] исследуется задача повышения вычислительной эффективности методов классификации изображений при наличии большой базы данных эталонных объектов. Предложен новый алгоритм приближенного поиска ближайшего соседа на основе сведения задачи к проверке статистических гипотез об однородности выделенных признаков. Построена итеративная процедура поиска, в которой следующий эталон из базы данных выбирается по принципу максимума правдоподобия (совместного распределения расстояний между входным изображением и проверенными на предыдущих шагах эталонами). Для оценки этого правдоподобия для большого числа признаков использовалось асимптотически нормальное распределение рассогласований между изображениями. При наличии параллельной среды выполнения (кластер машин, многоядерный процессор и т. п.) представлена параллельная реализация предложенного алгоритма, в которой все эталоны распределяются на множество доступных узлов (например, выполняемых потоков). Узел, первым нашедший эталон, удовлетворяющий условию останова, посылает команду останова остальным потокам. Для наилучшего присвоения эталонов каждому узлу использовалась модификация метода кластеризации GAAC (group average agglomerative clustering), в которой размеры кластеров фиксируются для достижения наилучшей степени параллелизма и ищется не минимум, а максимум суммы расстояний. Приведены результаты экспериментальных исследований для систем идентификации лиц. Показано, что предложенный алгоритм позволяет в несколько раз снизить время классификации по сравнению с известными методами приближенного поиска ближайших соседей.

Работа выполнена при поддержке Лаборатории алгоритмов и технологий анализа сетевых структур НИУ ВШЭ.

[1] *Savchenko A. V.* Search techniques in intelligent classification systems. — Springer International Publishing, 2016. 83 p.

Maximal likelihood approximate nearest neighbor search techniques in intelligent image classification systems

Savchenko Andrey avsavchenko@hse.ru

N. Novgorod, Russia, HSE

In [1, ch. 4], the problem of insufficient performance of image classification methods has been studied in the case of large databases. The novel approximate nearest neighbor algorithm is proposed by considering the recognition task as a testing of statistical hypothesis for homogeneity of extracted image features. The iterative search procedure has been explored, in which the next instance from the training database has been chosen to maximize the likelihood (conditional joint probability density of the distances between the input image and previously checked instances). The asymptotically normal distribution of the dissimilarity measure has been used for high-dimensional feature vectors. The author demonstrated the possibility to implement this search procedure in the parallel environment, if there are several tasks (CPU cores, nodes in a cluster, and machines in distributed environment) which can be executed in parallel. An efficiency of the proposed algorithm is increased, when the reference objects are distant to each other. Hence, an adaptive choice of distant clusters was proposed. In this algorithm, the distances between each reference object and all other instances in each cluster are summarized. This algorithm is quite similar to the GAAC (group-average agglomerative clustering), but it looks for a maximum sum of distances (not minimum, as in the GAAC). Also, the sizes of each cluster are chosen identical. The experimental results in face recognition prove that the proposed algorithm is much more effective for the medium-sized databases, than the brute force and the known approximate nearest neighbor methods.

The work is supported by Laboratory of Algorithms and Technologies for Network Analysis, National Research University Higher School of Economics.

[1] Savchenko, A. V. 2016. *Search techniques in intelligent classification systems*. Springer International Publishing. 83 p.

Объектно-ориентированная классификация в задаче распознавания подстилающей поверхности в арктических экосистемах

Трекин Алексей Николаевич[1]⋆ alexey.trekin@gmail.com
Мурынин Александр Борисович[1,2] amurynin@bk.ru
Матвеев Иван Алексеевич[1,2] matveev@ccas.ru
Игнатьев Владимир Юрьевич[1,2] vladimir.ignatiev.mipt@gmail.com

[1]Россия, Москва, НИИ «Аэрокосмос»
[2]Россия, Москва, ФИЦ ИУ РАН

Разработан метод распознавания типов земной поверхности по космическим изображениям с использованием объектно-ориентированной классификации. Классификация происходит в два этапа. На первом этапе объектной классификации происходит сегментация изображения путем семантической сегментации с использованием Марковских случайных полей. На втором этапе производится классификация полученных на первом этапе объектов байесовским классификатором.

Работоспособность метода проверена на космических изображениях Landsat 8, и приведены результаты двух вариантов метода в сравнении с поточечной классификацией. Метод дает широкие возможности по улучшению результатов путем замены или улучшения алгоритмов кластеризации и классификации, каждый из которых можно развивать независимо [1].

Работа выполнена при поддержке Министерства образования и науки Российской Федерации в рамках ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2014–2020 годы» (уникальный идентификатор проекта RFMEFI57414X0086).

[1] *Гурченков А. А., Мурынин А. Б., Трекин А. Н., Игнатьев В. Ю.* Метод объектно-ориентированной классификации объектов подстилающей поверхности в задаче аэрокосмического мониторинга состояния импактных районов Арктики // Вестник Московского гос. техн. ун-та имени Н. Э. Баумана. Естественные науки. — М.: Изд-во МГТУ им. Н. Э. Баумана, 2016 (в печати).

Object-oriented classification for recognition of earth surface in Arctic ecosystems

Trekin Alexey[1]⋆ alexey.trekin@gmail.com
Murynin Alexander[1,2] amurynin@bk.ru
Matveev Ivan[1,2] matveev@ccas.ru
Ignatiev Vladimir[1,2] vladimir.ignatiev.mipt@gmail.com
[1]Moscow, Russia, ISR "Aerocosmos"
[2]Moscow, Russia, FRC CSC RAS

The paper proposes a method for recognition of earth surface types in space images using object-oriented classification. The classification is conducted in two stages: Markov stochastic segmentation for object extraction and supervised classification of the objects.

At the first stage of the object-oriented classification, the image segmentation is conducted. Each region is characterized by homogenous inner structure and significant difference to the adjacent regions. Segmentation uses Markov random fields technique. At the second stage, the regions gained are classified by supervised Bayesian classificator as one of the previously specified surface types.

The method is tested with Landsat 8 multispectral space images, the results of two variants of the method are compared with pointwise classification.

While the current results require further evaluation and development, the method gives wide opportunities for the results enhancement with replacement or improvement of either clustering or classification techniques [1].

This research is funded by the Russian Ministry of Education and Science (Project identificator RFMEFI57414X0086).

[1] Gurchenkov, A., A. Murynin, A. Trekin, and V. Ignatiev. 2016 (in press). Object-oriented classification for recognition of earth surface in Arctic ecosystems. *Herald of the Bauman Moscow Stare Technical University. Natural Sciences*. Moscow: Bauman University Publs.

Новый метод интеллектуального анализа и распознавания трехмерных изображений: описание и примеры

Федотов Николай Гаврилович[1] fedotov@pnzgu.ru
Сёмов Алексей Александрович[2]⋆ mathematik_aleksey@mail.ru
Моисеев Александр Владимирович[3] moigus@mail.ru
[1]Россия, Пенза, Пензенский государственный университет
[2]Россия, Пенза, ООО «Комэрф»
[3]Россия, Пенза, ПензГТУ

Предложен новый подход к распознаванию трехмерных (3D) объектов. Дано математическое описание гипертрейс-преобразования, разработанного на основе стохастической геометрии и функционального анализа. Проанализированы принципы интеллектуального анализа 3D изображений, построенные на его основе. Одной из интеллектуальных способностей предлагаемого метода является конструирование гипертриплетных признаков разной структуры («длинные» и «короткие» признаки). Разные типы признаков находят свое применение в принципах интеллектуального анализа и распознавания 3D изображений (верифицируемость и фальсифицируемость изображений). В работе дано описание теоретических примеров построения «длинных» и «коротких» признаков изображений. Обосновано их различие и особенности практического применения. Гипертрейс-преобразование имеет уникальную способность, аналогичную возможности человеческой зрительной системы, когда при достаточно беглом взгляде человек может быстро отличить друг от друга два пространственных объекта. Данное обстоятельство повышает скорость работы сканирующей системы и надежность всей системы распознавания изображений в целом, улучшая интеллектуальные способности гипертрейс-преобразования [1].

Работа поддержана грантом РФФИ № 15-07-04484.

[1] *Федотов Н. Г., Семов А. А., Моисеев А. В.* Новый метод интеллектуального анализа и распознавания 3D изображений: описание и примеры // Машинное обучение и анализ данных, 2016 (в печати).

New method for three-dimensional images intelligent analysis and recognition: Description and examples

Fedotov Nikolay[1] fedotov@pnzgu.ru
Syemov Aleksey[2]⋆ mathematik_aleksey@mail.ru
Moiseev Alexandr[3] moigus@mail.ru
[1]Penza, Russia, Penza State University
[2]Penza, Russia, Ltd "Comearth"
[3]Penza, Russia, PenzSTU

A new approach to the three-dimensional (3D) objects' recognition is proposed. A detailed mathematical description of method developed on the above approach basis is shown. Hypertrace transform technique scan is described and the scanning element choice is substantiated. The principles of 3D images intellectual analysis and recognition built on its basis are analyzed.

The suggested method is based on the stochastic geometry and functional analysis. Hypertrace transform has many advantages and data mining capabilities. For example, one of the suggested method intellectual capabilities is the construction of different structure hypertriplet features ("long" and "short" features). Different types of features are reflected in the principles of 3D images intelligent analysis and recognition (verifiability and falsifiability of images).

Due to the limited volume and conceptual orientation of article, the practical results are missing. The theoretical examples description of verification of "long" features and falsification of "short" features of images is given. Their differences and practical application specificities are substantiated.

Hypertrace transform has a unique ability which is a similar possibility of human visual system when at sufficiently brief glance, people quickly can distinguish from each other two spatial objects. This fact increases the scanning system speed and the image recognition system reliability in general, improving the intellectual abilities hypertrace transform [1].

This research is funded by the Russian Foundation for Basic Research, grant 15-07-04484.

[1] Fedotov, N. G., A. A. Syemov, and A. V. Moiseev. 2016 (in press). Feature space minimization of 3D image recognition based on stochastic geometry and functional analysis. *Machine Learning Data Anal.*

Метод детектирования кисти руки на основе одноклассового классификатора и скелетных графов

Грачева Инесса Александровна[1]⋆ gia1509@mail.ru
Копылов Андрей Валериевич[1] And.Kopylov@gmail.com
Середин Олег Сергеевич[1] oseredin@yandex.ru
Кушнир Олеся Александровна[1] kushnir-olesya@rambler.ru
Ларин Александр Олегович[2] ekzebox@gmail.com
[1]Россия, Тула, ТулГУ
[2]Россия, Москва, МФТИ

Точное и надежное детектирование изображения кисти руки человека в видеопотоке является необходимым и критически важным этапом при построении систем бесконтактного взаимодействия человека и технических устройств, например в задачах распознавания жестов или биометрической идентификации.

Предложенный метод детектирования кисти руки в видеопотоке на основе одноклассового классификатора, вероятностной гамма-нормальной модели и скелетных графов описан в работе [1]. Первоначальная сегментация участков кожи выполняется с помощью модифицированной версии одноклассового пиксельного классификатора, обученного фрагментом изображения части лица, и не требующего формирования обучающей выборки для построения модели фона. Результатом классификации является степень принадлежности к классу интереса. Уточнение первоначальной сегментации осуществляется за счет согласования локальных решений и привлечения дополнительной информации о структуре изображения. Для этого применяется специальный фильтр со свойствами переноса структуры на основе вероятностной гамма-нормальной модели. Для принятия окончательного решения о том, что найденный фрагмент является изображением кисти человека, используется метод сравнения бинарных изображений на основе их скелетных графов. Работа поддержана грантами РФФИ №№ 14-07-00527, 16-57-52042 и 16-07-01039.

[1] *Kopylov A., Seredin O., Kushnir O., Gracheva I., Larin A.* Background-invariant robust hand detection using one-class color segmentation and skeleton description // J. Pattern Recogn. Image Anal., 2016 (в печати).

Background-invariant robust hand detection using one-class color segmentation and skeleton description

Gracheva Inessa[1]⋆ gia1509@mail.ru
Kopylov Andrey[1] And.Kopylov@gmail.com
Seredin Oleg[1] oseredin@yandex.ru
Kushnir Olesia[1] kushnir-olesya@rambler.ru
Larin Alexander[2] ekzebox@gmail.com
[1]Tula, Russia, TulSU
[2]Moscow, Russia, MIPT

Accurate and reliable hand detection on video streams is a necessary and crucial step in the human–computer interaction systems, such as the gesture recognition or biometric identification.

The hand detection method is suggested on the basis of one-class color segmentation, probabilistic gamma-normal model, and skeleton description [1]. The initial segmentation of the skin is performed using the modified one-class pixel-wise classifier, which is trained by the face fragment image and does not require the formation of a training set to construct the background model. The classification result is the degree of belonging to the interest class. Improving the initial segmentation is performed by combining the local decisions and additional image structure information. For this, the special filter with structure-transferring properties on the basis of probability gamma-normal model has been used. The final decision that detected fragment is the image of the human hand is done using the method of comparing the binary images on the basis of their skeleton descriptions.

This research is funded by the Russian Foundation for Basic Research, grants 14-07-00527, 16-57-52042, and 16-07-01039.

[1] Kopylov A., Seredin O., Kushnir O., Gracheva I., Larin A. 2016 (in press). Background-invariant robust hand detection using one-class color segmentation and skeleton description. *J. Pattern Recogn. Image Anal.*

Морфологическая оценка сходства изображений с использованием глубоких конволюционных нейронных сетей

Лебедев Максим Алексеевич* lebedev_maxim@list.ru
Костромов Никита Алексеевич nikita-kostromov@yandex.ru
Рубис Алексей Юрьевич arcelt@mail.ru
Комаров Денис Валерьевич mrkomap@mail.ru
Выголов Олег Вячеславович o.vygolov@gosniias.ru
Визильтер Юрий Валентинович viz@gosniias.ru
Россия, Москва, ФГУП «ГосНИИАС»

Задача оценки сходства изображений одной сцены, полученных при различных условиях регистрации или в различных спектральных диапазонах, актуальна во многих практических приложениях технического зрения. Один из подходов к решению данной задачи предложен в рамках морфологии Ю. П. Пытьева [1]. Сравнение изображений производится по форме — геометрическому инварианту сцены, остающемуся неизменным при различных яркостных преобразованиях. В данной статье предлагается новый метод морфологической оценки сходства изображений с использованием морфологических коэффициентов корреляции и глубоких конволюционных нейронных сетей (DCNN — deep convolutional neural networks) [2,3]. Отличительной особенностью метода является отсутствие этапа построения мозаичных форм исходных изображений и, как следствие, более высокая робастность к шуму. В статье описаны архитектура DCNN сети и методика ее обучения. Приводятся результаты работы предложенного метода на модельных изображениях.

[1] *Pyt'ev Yu.* Morphological image analysis // Pattern Recogn. Image Anal., 1993. Vol. 3. No. 1. P. 19–28.

[2] *Vizilter Y. V., Zheltov S. Y* Geometrical correlation and matching of 2D image shapes // ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci. 2012. Vol. 1. No. 3. P. 191–196.

[3] *Krizhevsky A., Sutskever I., Hinton G. E.* ImageNet classification with deep convolutional neural networks // NIPS, 2012. Vol. 25. No. 2. P. 1106–1114.

Morphological image matching using deep convolutional neural networks

Lebedev Maksim ⋆ lebedev_maxim@list.ru
Kostromov Nikita nikita-kostromov@yandex.ru
Rubis Aleksey arcelt@mail.ru
Komarov Denis mrkomap@mail.ru
Vygolov Oleg o.vygolov@gosniias.ru
Vizilter Yuri viz@gosniias.ru

Moscow, Russia, FGUP "GosNIIAS"

Estimating similarities between two images of the same scene but obtained in different light conditions or different spectral ranges (e. g., TV and infrared) is the challenging problem in many practical technical vision tasks. One approach to solving this problem is proposed in the framework of Pyt'ev morphology [1]. Image comparison is based on structural image description (image "shape"), which is a geometrical scene invariant that remains constant in case of various image intensity transformations. In the paper, a new method of morphological image matching is proposed based on morphological correlation coefficients and deep convolutional neural networks (DCNN) [2,3]. The distinctive feature of the method is the lack of the segmentation step and, therefore, higher robustness to noise. The paper describes the architecture of the DCNN and learning methodology. The examples of image matching are shown on simulated images.

[1] Pyt'ev, Yu. 1993. Morphological image analysis. *Pattern Recogn. Image Anal.* 3(1):19–28.

[2] Vizilter, Y. V., and S. Y. Zheltov. 2012. Geometrical correlation and matching of 2D image shapes. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.* 1(3):191–196.

[3] Krizhevsky, A., I. Sutskever, and G. E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. *NIPS.* 25(2): 1106–1114.

Классификация двумерных фигур с использованием скелетно-геодезических гистограмм толщин-расстояний

Ломов Никита Александрович[1]⋆ nikita-lomov@mail.ru
Сидякин Сергей Владимирович[2] sersid@gosniias.ru
Визильтер Юрий Валентинович[2] viz@gosniias.ru
[1]Москва, Россия, МГУ
[2]Москва, Россия, ФГУП «ГосНИИАС»

Рассматривается задача классификации двумерных бинарных фигур. Для ее решения в части признакового описания предлагается использовать скелетно-геодезическую гистограмму толщин-расстояний.

Гистограмма толщин-расстояний является разновидностью морфологических гистограмм, основанных на статистиках парных расстояний между элементами фигуры. Она вычисляется на основе скелетно-геодезических расстояний и разностей толщин между парами ребер скелета фигуры, этим отличаясь от обычных геодезических гистограмм, которые вычисляются для всех точек фигуры. Переход к использованию ребер скелета, а также областей их притяжения позволяет значительно ускорить расчет скелетно-геодезических гистограмм толщин-расстояний, сохранив при этом ряд полезных свойств, присущих обычным геодезическим гистограммам.

Полученные результаты классификации подчеркивают высокий потенциал предложенного дескриптора [1].

Работа поддержана грантом РФФИ № 15-07-01323 А и грантом РНФ № 16-11-00082.

[1] *Ломов Н. А., Сидякин С. В., Визильтер Ю. В.* Классификация двумерных фигур с использованием скелетно-геодезических гистограмм толщин-расстояний // Компьютерная оптика, 2016 (в печати).

Classification of two-dimensional figures using skeleton-geodesic histograms of thicknesses and distances

Lomov Nikita[1]⋆ nikita-lomov@mail.ru
Sidyakin Sergey[2] sersid@gosniias.ru
Vizilter Yury[2] viz@gosniias.ru
[1]Russia, Moscow, MSU
[2]Russia, Moscow, FGUP "GosNIIAS"

The paper considers the shape classification task. It is proposed to use skeleton-geodesic histogram of thicknesses and distances to solve the aforementioned problem.

Skeleton-geodesic histogram of thicknesses and distances is a kind of morphological histograms, based on the statistics of pair distances between shape elements. It is computed based on skeleton-geodesic distances and thickness differences between pairs of skeleton bones of a shape. This differs from conventional geodesic histograms that are computed for all points of a figure. The switch to the skeleton bones and areas of skeleton bones attraction can significantly speed up the calculation of skeleton-geodesic histogram of thicknesses and distances, while maintaining a number of useful properties inherent in usual geodesic histograms.

Obtained classification results indicate the high potential of the proposed descriptor [1].

This research is funded by the Russian Foundation for Basic Research, grant 15-07-01323 A, and Russian Science Foundation, grant 16-11-00082.

[1] Lomov, N., S. Sidyakin, and Yu. Vizilter. 2016 (in press). Classification of two-dimensional figures using skeleton-geodesic histograms of thicknesses and distances. *Computer Optics*.

Распознавание цифровых шрифтов по изображениям на основе дискового покрытия

Местецкий Леонид Моисеевич* mestlm@mail.ru
Ломов Никита Александрович nikita-lomov@mail.ru
Россия, Москва, МГУ

Количество современных цифровых шрифтов исчисляется тысячами. Необходимость определения, каким шрифтом набран текст, возникает у дизайнеров, разработчиков шрифтов, компаний-правообладателей. Предлагается оригинальный подход к построению метрики для оценки сходства и различия шрифтов на основе измерения площади дискового покрытия изображений шрифтовых символов. Рассматривается понятие «ширина фигуры» с целью использования в качестве интегрального морфологического дескриптора формы изображений. Предлагается подход к описанию этого понятия на основе покрытия фигуры дисками определенного размера. Для распознавания шрифта изображения символов сканированного фрагмента текста аппроксимируются многоугольными фигурами.

С помощью дискового покрытия строится дескриптор формы — функция зависимости площади дискового покрытия от размера дисков. Предлагается метод аналитического вычисления этой функции для многоугольных фигур. Метод основан на использовании медиального представления фигуры в виде скелета и радиальной функции. Универсальность метода определяется возможностью аппроксимации многоугольными фигурами растровых изображений и объектов с нелинейной границей. Метод обеспечивает высокую точность и вычислительную эффективность расчета предложенного дескриптора формы. Эффективность предлагаемого подхода демонстрируется вычислительными экспериментами с коллекцией цифровых шрифтов компании Паратайп, включающей 1884 шрифтовых начертания [1].

Работа поддержана грантом РФФИ № 14-01-00716.

[1] *Ломов Н. А., Местецкий Л. М.* Площадь дискового покрытия — дескриптор формы изображения // Компьютерная оптика, 2016 (в печати).

Recognition of digital fonts from images based on the disk cover

Mestetskiy Leonid ⋆ mestlm@mail.ru
Lomov Nikita nikita-lomov@mail.ru
Moscow, Russia, MSU

Modern digital font library includes thousands of font styles. At the same time, designers, font designers, companies, and copyright holders need a tool to quickly determine what is the font of the typed text. The paper proposes an original approach to the construction of metrics to assess similarities and differences of fonts based on the measurement of the area of disk cover of font character images. The term "width of figure" is considered in order to use it as an integral descriptor of the morphological features of images. The approach to the description of this concept is based on figures cover by disks of the certain size. To recognize the font, the scanned images of text characters are approximated by polygonal figures.

Function area of disk cover as dependence of disk size is defined as a shape descriptor. An analytical method for the calculation of the area of disk cover for polygonal figure is proposed. This approach is universal as polygonal figures approximate binary raster images and objects with nonlinear boundary. The proposed method is based on the medial representation of objects consisting of the skeleton and the radial function. The method provides high accuracy and computational efficiency. Using of the proposed shape descriptor is demonstrated in the sample application to computer fonts recognition problem. The effectiveness of the proposed approach is demonstrated by computational experiments with a collection of digital fonts ParaType including 1884 typefaces [1].

This research is funded by the Russian Foundation for Basic Research, grant 14-01-00716.

[1] Lomov, N., and L. Mestetskiy. 2016 (in press). Area of the disk cover as an image shape descriptor. *Comput. Opt.*

Алгоритмы уточнения оси зеркальной симметрии, найденной методом сравнения подцепочек скелетных примитивов

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

В [1] был предложен метод поиска оси зеркальной симметрии бинарного изображения, основанный на функции сравнения подцепочек примитивов, кодирующих скелет фигуры. Предложенный метод позволяет искать ось симметрии не только идеально симметричных, но и почти симметричных (квазисимметричных) изображений за время, близкое к реальному. Для оценки симметричности фигуры относительно некоторой оси используется теоретико-множественное подобие Жаккарда, применяемое к подмножествам пикселей фигуры при делении ее осью. Зачастую ось, найденная скелетным методом, отклоняется в большей или меньшей степени от эталонной оси симметрии, определенной переборным методом из всех возможных осей, пересекающих фигуру. Поэтому предлагаются алгоритмы, позволяющие уточнить найденную быстрым скелетным методом ось, путем поиска ближайшей к ней оси с большим по мере Жаккарда значением симметричности. Экспериментальные исследования на базе изображений Flavia показывают, что предложенные алгоритмы позволяют найти эталонную ось симметрии (или отличающуюся по мере от эталонной не более чем на 2%) за время, близкое к реальному, что является существенным увеличением производительности по сравнению с любым из оптимизированных методов полного перебора, описанных в [1].

Работа поддержана грантами РФФИ № 14-07-00527 и № 16-57-52042.

[1] *Kushnir O., Fedotova S., Seredin O., Karkishchenko A.* Reflection symmetry of shapes based on skeleton primitive chains // Analysis of images, social networks and texts. — Communications in computer and information science ser. — Switzerland: Springer International Publishing, 2016 (in press).

The algorithms of adjustment of reflection symmetry axis found by the skeleton primitive subchains comparison method

Fedotova Sofia⋆ fedotova.sonya@gmail.com
Seredin Oleg oseredin@yandex.ru
Kushnir Olesia kushnir-olesya@rambler.ru
Tula, Russia, TulSU

In [1], the method of identifying a reflection symmetry axis of binary images was proposed. This method is based on comparison of skeleton primitive subchains. It allows computing the absolute or approximate symmetry axis almost in real time. For evaluation of reflection symmetry measure of a shape regarding to some axis, the set-theoretic expression of Jaccard similarity is utilized. It is applied to the subsets of pixels of the shape which are split by the axis. Often, an axis found by the subskeletons comparison method diverges more or less of the ground-truth axis found by the brute-force algorithm. Thus, the algorithms of adjustment of reflection symmetry axis found by the skeleton primitive subchains comparison method are proposed. They are based on idea of searching the axis which is located near the seed skeleton axis and has greater Jaccard similarity measure. The experimental study on the Flavia Dataset shows that proposed algorithms allow to find the ground-truth axis (or the axis which has Jaccard similarity measure less than 2% under the ground-truth axis) almost in real time. It is considerably faster than any of the optimized brute-force methods performed [1].

This research is funded by the Russian Foundation for Basic Research, grants 14-07-00527 and 16-57-52042.

[1] Kushnir, O., S. Fedotova, O. Seredin, and A. Karkishchenko. 2016 (in press). Reflection symmetry of shapes based on skeleton primitive chains. *Analysis of images, social networks and texts*. Communications in computer and information science ser. Switzerland: Springer International Publishing.

Сегментация радужной оболочки методом парных градиентов и уточнение границы зрачка на изображении глаза

Ефимов Юрий Сергеевич[1]⋆ yuri.efimov@phystech.edu
Матвеев Иван Алексеевич[2] matveev@ccas.ru
[1]Россия, Долгопрудный, МФТИ
[2]Россия, Москва, ФИЦ ИУ РАН

Выделение области радужной оболочки на изображении глаза осуществляется в два этапа. На первом этапе для поиска внешней и внутренней границ радужки, представленных как окружности, используется модификация методологии Хафа — метод парных градиентов. После обработки изображения фильтром Кэнни из пикселей получившихся границ выбираются пары, с большой вероятностью лежащие на одной окружности. Для сокращения перебора и избавления от шумов вводятся условия выбора пары с учетом локальных свойств поля яркости изображения. На втором этапе осуществляется поиск точного контура зрачка методом оптимального кругового пути с учетом построенной круговой аппроксимации его границы. Рассматриваются пути между левой и правой границами полярного представления кольцеобразной области зрачка. Вводится функция «стоимости», зависящая от их формы рассматриваемого контура и локальных свойств изображения. Путь, на котором эта функция принимает минимальное значение, соответствует искомой точной границе зрачка.

Проведен вычислительный эксперимент с целью определить оптимальные параметры предлагаемого метода и проверить его работоспособность на данных из открытых баз изображений радужки. Качество сегментации изображений радужной оболочки данной системой методов сравнимо с таковым у описанных в современной литературе аналогов [1].

Работа поддержана грантом РФФИ № 16-07-01171.

[1] *Ефимов Ю. С., Матвеев И. А.* Сегментация радужной оболочки методом парных градиентов и уточнение границы зрачка на изображении глаза // Мехатроника, автоматизация, управление, 2016 (в печати).

Iris image segmentation by paired gradient method with pupil border refinement

Efimov Yuriy[1]⋆ yuri.efimov@phystech.edu
Matveev Ivan[2] matveev@ccas.ru
[1]Dolgoprudny, Russia, MIPT
[2]Moscow, Russia, FRC CSC RAS

The proposed method of iris image segmentation consists of two main steps. The first step is rough iris center and boundary radii search using the modification of Hough methodology, named Gradient Pair method. Image is processed with Canny filter and pairs of pixels are selected from the resulting boundaries, which most likely belong to one circle-like iris area border. The selection criteria for the pair uses local properties of image intensity gradient field. The second step is the pupil boundary refinement using the circular shortest path method. Polar representation of iris inner border is considered and paths from its left border to the right are analyzed. "Cost" function depending on local image properties and the shape of the contour is introduced. The path with minimal "cost" corresponds to pupil precise boundary.

Computational experiment is performed on data from the public iris image databases to obtain the optimal parameters for Gradient Pair method and check the overall efficiency of the algorithm.

The proposed method provides the high quality of eye center and pupil boundary search. Iris segmentation accuracy is comparable to that of modern state-of-the-art methods and outperforms many of them [1].

This research is funded by the Russian Foundation for Basic Research, grant 16-07-01171.

[1] Efimov, Y., and I. Matveev. 2016 (in press). Iris image segmentation by paired gradient method with pupil border refinement. *Mechatronics Automation Control*.

Метод обнаружения позиции век при распознавании по радужной оболочке глаза на мобильном устройстве

Одиноких Глеб Андреевич[1,2]⋆ g.odinokikh@gmail.com
Гнатюк Виталий Сергеевич[1,2] vitgracer@gmail.com
Коробкин Михаил Владимирович[1,2] mikhail.korobkin@hotmail.com
Еремеев Владимир Алексеевич[1,2]
[1]Россия, Москва, ФИЦ ИУ РАН
[2]Россия, Зеленоград, Национальный исследовательский университет МИЭТ

При распознавании человека по радужке информация о положении век на изображении используется для удаления шума от век и ресниц, перекрывающих полезную область радужки, оценки качества изображения и многих других целей. Детектирование век, как правило, производится после вычислительно сложной операции нахождения границ радужки и склеры. В случае использования для распознавания мобильного устройства такой подход не всегда оправдан ввиду ограниченной производительности устройства, сложностей взаимодействия пользователя с устройством и сильно изменяющихся внешних условий окружающей среды. В данном случае информация о положении век может быть извлечена сразу после этапа детектирования зрачка и использована для определения пригодности изображения для последующих более сложных этапов алгоритма распознавания. Предложен метод определения положения век на изображении с целью оценки качества изображения и последующего определения границы радужки и века. Производительность метода была оценена в сравнении с несколькими существующими решениями с использованием четырех различных открытых баз данных радужек [1].

[1] *Одиноких Г. А.* Метод обнаружения позиции век при распознавании по радужной оболочке глаза на мобильном устройстве // Машинное обучение и анализ данных, 2016 (в печати).

Eyelid position detection method for mobile iris recognition

Odinokikh Gleb[1,2]⋆ g.odinokikh@gmail.com
Gnatyuk Vitaly[1,2] vitgracer@gmail.com
Korobkin Mikhail[1,2] mikhail.korobkin@hotmail.com
Eremeev Vladimir[1,2]

[1]Moscow, Russia, A. A. Dorodnicyn Computing Center FRC CSC RAS
[2]Zelenograd, Russia, National Research University of Electronic Technology

Information about eyelid position in an image is used during iris recognition for the eyelid and eyelash noise removal, iris image quality estimation, and other purposes. Eyelid detection is usually performed after iris-sclera boundary localization which is a fairly complex operation itself. If the authentication is working on a hand-held device, this order is not always justified, mainly, because of the device limited performance, user interaction difficulties, and highly variable environmental conditions. In this case, the eyelid position information could be used to determine whether the image should be passed for the further complex processing operations. This paper proposes a method of eyelid position detection for iris image quality estimation and further complete eyelid border localization and compares its performance with several similar existing methods on four open datasets [1].

[1] Odinokikh, G. A. 2016 (in press). Eyelid position detection method for mobile iris recognition. *J. Machine Learning Data Anal.*

Определение видимой области радужки классификатором текстур с опорным множеством

Соломатин Иван Андреевич[1]⋆ ivan.solomatin@phystech.edu
Матвеев Иван Алексеевич[1,2,3] matveev@ccas.ru
Новик Владимир Петрович[3] novikvp@mail.ru
[1]Россия, Москва, МФТИ
[2]Россия, Москва, ФИЦ ИУ РАН
[3]Россия, Москва, Iritech Inc.

Распознавание человека по изображению радужной оболочки — актуальная задача в биометрических системах. Помимо выделения радужки как кольцевой области для повышения точности распознавания определяют области затенения (веки, ресницы, тени, блики — далее ВРТБ).

На вход подаются черно-белое изображение и кольцевая область локализации радужки. Требуется выделить ВРТБ области, т. е. получить ВРТБ маску — бинарное изображение. Суть метода состоит в том, что изображение переводится в полярные координаты, затем на нем находится опорное множество для обучения классификатора, после чего результаты классификации подвергаются постобработке и переводятся в декартовы координаты [1].

Метод был реализован на C++ с использованием библиотеки OpenCV. Тестирование проводилось на изображениях радужки из баз CASIA и ICE в два этапа. На первом этапе полученная ВРТБ маска сравнивалась с экспертной маской. В качестве функции ошибки использовалась сумма относительных ошибок первого и второго рода. Для обеих баз были посчитаны средние ошибки: $\overline{E}_{\text{CASIA}} = 0{,}196$ и $\overline{E}_{\text{ICE}} = 0{,}321$. На втором этапе тестирования полученные ВРТБ маски были использованы для идентификации человека по радужке алгоритмом Либора Масека. Маска значительно улучшает качество идентификации, однако уступает экспертной разметке. Например, для базы CASIA S: $\text{EER}_{\text{без маски}} = 0{,}0065$, $\text{EER}_{\text{с маской}} = 0{,}0039$ и $\text{EER}_{\text{с экспертной маской}} = 0{,}0029$.

Работа поддержана грантом РФФИ № 16-07-01171.

[1] *Соломатин И. А., Матвеев И. А., Новик В. П.* Определение видимой области радужки классификатором текстур с опорным множеством // Известия РАН. Теория и системы управления, 2016 (в печати).

Detecting visible areas of iris by qualifier of textures with support set

Solomatin Ivan[1]⋆ ivan.solomatin@phystech.edu
Matveev Ivan[1,2,3] matveev@ccas.ru
Novik Vladimir[3] novikvp@mail.ru
[1]Moscow, Russia, MIPT
[2]Moscow, Russia, FRC CSC RAS
[3]Moscow, Russia, Iritech Inc.

A person recognition by the image of the iris is an actual problem. Usually, to increase the accuracy of recognition, the areas of occlusion are detected in addition to locating of the iris as an annular region.

The input of the algorithm consists of black and white image and an annular region of iris localization. The problem is to find regions of occlusions, i. e., to build an occlusion mask — binary image.

The method suggests to transfer image to the polar coordinates, then to find a support set on the image, teach the classifier on it and classify all the pixels into two classes "iris" and "occlusion," after that make postprocessing of the results of the classification, and, finally, convert them back to Cartesian coordinates [1].

The method was implemented in C++, using OpenCV library. Computer experiment was conducted on the iris images from CASIA and ICE databases and consisted of two steps. On the first step, the mask was compared to the expert mask. As the error function, the sum of the relative errors of the first and second type was used. Average errors were calculated for both databases: $\overline{E}_{\text{CASIA}} = 0.196$ and $\overline{E}_{\text{ICE}} = 0.321$. On the second step, the masks obtained by the algorithm were used to identify human by iris using Libor Masec algorithm. The mask significantly increases the quality of identification, but not as much as an expert one. For example, on the base CASIA S: $\text{EER}_{\text{without mask}} = 0.0065$, $\text{EER}_{\text{with mask}} = 0.0039$, and $\text{EER}_{\text{with expert mask}} = 0.0029$.

This research is funded by the Russian Foundation for Basic Research, grant 16-07-01171.

[1] Solomatin, I., I. Matveev, and V. Novik. 2016 (in press). Detecting visible areas of iris by qualifier of textures with support set. *J. Comput. Syst. Sci. Int.*

Определение области затенения радужки кластеризацией, основанной на локальных текстурных признаках

Талипов Камиль Илгизарович[1] kamiltalipov@gmail.com
Матвеев Иван Алексеевич[2]⋆ ivanmatveev@mail.ru
[1]Россия, Москва, МФТИ
[2]Россия, Москва, ФИЦ ИУ РАН

Решается задача выделения точек затенения области радужки различными объектами. Исходными данными является изображение радужки глаза человека и окружности, аппроксимирующие границы зрачок–радужка и радужка–склера. В качестве метода решения предлагается использовать расчет локальных текстурных признаков и кластеризацию полученного вектора признаков. Целью работы является построение алгоритма, выделяющего точки затенения, и исследование возможности сегментации затенений радужки без априорно заданной модели ее текстуры [1].

В качестве набора локальных тестурных признаков были использованы: первый момент яркости в окрестности 7×7; второй момент яркости в окрестности 7×7; стандартное отклонение в окрестности 3×3; перепад (разница между максимальной яркостью и минимальной) яркости в окрестности 3×3; главные компонененты с 90%-ной значимостью для матрицы данных, где рассматривались значения яркости в окрестности 7×7 точки как вектор; расстояние до зрачка, нормированное на радиус радужки; случайное марковское поле в окрестости 7×7. На данном наборе протестированы различные методы кластеризации при разных метриках. Для всех методов обнаружилось, что деление на три класса дает лучшие результаты, чем деление на два класса. Наибольшая точность распознавания в 79,2%±1,1% получена при использовании нормализованного Евклидова расстояния как метрики и k-medoids как метода кластеризации при параметре $k = 3$. Необходимо дальнейшее изучение новых локальных текстурных признаков и методов кластеризации.

Работа поддержана грантом РФФИ № 15-01-05552.

[1] *Талипов К. И., Матвеев И. А.* Определение области затенения радужки кластеризацией, основанной на локальных текстурных признаках // Машинное обучение и анализ данных, 2016 (в печати).

Eyelids and eyelash detection based on clusterization of vector of local features

Talipov Kamil[1] kamiltalipov@gmail.com
Matveev Ivan[2]⋆ ivanmatveev@mail.ru
[1]Moscow, Russia, MIPT
[2]Moscow, Russia, FRC CSC RAS

An attempt to solve the problem of extracting areas where the iris is occluded by various objects has been made. Initial data consist of an image of iris and a circle approximating the boundary between the sclera, the iris,nd the pupil. Calculation of local texure features and clusterizing the data based on the extracted information is proposed as a solution method. Two main goals of this particular work are to introduce an effective algorithm for occluded point detection and to study the possibility of their segmentation without a preset texture model [1].

As local texture features, there were used: the first moment of brightness in 7×7 area; the second moment of brightness in 7×7 area; standard deviation in 3×3 area; difference between maximal and minimum values in 3×3 area; principal components with 90 percent information; normalized distance from the point to the center of pupil; and Markov random field in 7×7 area. Different clustering algorithms were tested on this set of local texture features. For all methods, it was found that the division into three classes gives better results than the division into two classes. The best precision $79.2\% \pm 1.1\%$ was shown by k-medoids algorithm and normalized Euclidean distance as distance measure. Further study of new local texture features and clustering techniques is needed.

This research is funded by the Russian Foundation for Basic Research, grant 15-01-05552.

[1] Talipov, K., and I. Matveev. 2016 (in press). Eyelids and eyelash detection based on clusterization of vector of local features. *J. Machine Learning Data Anal.*

Быстрый алгоритм поиска границ зрачка и радужной оболочки глаза

Чигринский Виктор Владимирович[1]⋆ chigrinskiy.viktor@phystech.edu
Ефимов Юрий Сергеевич[1] yuri.efimov@phystech.edu
Матвеев Иван Алексеевич[2] matveev@ccas.ru
[1]Россия, Долгопрудный, МФТИ
[2]Россия, Москва, ФИЦ ИУ РАН

Решается задача поиска границ зрачка и радужной оболочки на изображении глаза. Определяются параметры аппроксимирующих окружностей, а именно: координаты центров и радиусы. Для решения задачи выполняется последовательность шагов. Сначала изображение подвергается морфологической обработке для избавления от бликов и посторонних мелких шумов. Обработанное изображение бинаризуется, и зрачком полагается наиболее похожая на круг компонента связности бинаризованного изображения. Затем на морфологически обработанном изображении выделяются граничные точки с помощью оператора Кэнни и по этим граничным точкам с учетом определенного центра зрачка окончательно определяется его граница. Поиск границ радужной оболочки осуществляется путем анализа плотности распределения граничных точек изображения по расстояниям до найденного центра зрачка. Максимум такой плотности соответствует большому количеству точек, удаленных от центра зрачка на расстояния, находящиеся в узком диапазоне значений. Границей радужной оболочки полагается аппроксимирующая эти точки окружность.

Проведен вычислительный эксперимент с целью определения оптимальных параметров метода и сравнения полученных результатов с результатами, получаемыми применением метода парных градиентов к решению этой же задачи [1].

Работа поддержана грантом РФФИ № 16-07-01171.

[1] *Чигринский В. В., Ефимов Ю. А., Матвеев И. А.* Быстрый алгоритм поиска границ зрачка и радужной оболочки глаза // Машинное обучение и анализ данных, 2016 (в печати).

Fast algorithm for determining pupil and iris boundaries

Chigrinskiy Viktor[1]⋆ chigrinskiy.viktor@phystech.edu
Efimov Yuriy[1] yuri.efimov@phystech.edu
Matveev Ivan[2] matveev@ccas.ru
[1]Dolgoprudny, Russia, MIPT
[2]Moscow, Russia, FRC CSC RAS

The paper presents a method of pupil and iris boundaries determining on eye images. The purpose is to find out the parameters of approximating circles, namely, coordinates of centers and radiuses. To solve the problem, several steps are implemented. First, to remove highlights and small noise, the origin image is morphologically processed. Then, the resulted image is binarized. The most similar to a circle connectivity component of the binary image indicates a pupil. After that, the edges of the image are obtained by applying the Canny edge detector to the morphologically processed image. Using these edges and just found pupil center, the pupil boundaries are determined finally. To detect iris boundaries, the density of the edge points distribution by their distances to the pupil center are used. The maximum of this density of the distribution denotes a big amount of the points that are remote from the pupil center to the distances localized in the narrow range of values. The approximating these points circle denotes the iris boundary.

The computational experiment was performed to find out the optimal parameters of the method and to compare the obtained results with the results given by the paired gradient method [1].

This research is funded by the Russian Foundation for Basic Research, grant 16-07-01171.

[1] Chigrinskiy, V., Y. Efimov, and I. Matveev. 2016 (in press). Fast algorithm for determining pupil and iris boundaries. *Machine Learning Data Anal.*

Динамическое выравнивание непрерывных временных рядов

Гончаров Алексей Владимирович[1]⋆
alex.goncharov@phystech.edu
Стрижов Вадим Викторович[2] strijov@ccas.ru
[1]Россия, Москва, МФТИ
[2]Россия, Москва, ФИЦ ИУ РАН

Для совместного анализа дискретных временны́х рядов, собранных датчиками, работающими с существенно различной частотой предлагается использовать их непрерывное представление. Применяется метрический метод анализа временны́х рядов, основанный на применении функции динамического выравнивания. Она находит наилучшее соответствие между двумя временны́ми рядами при их нелинейной деформации — растягивании, сжатии или смещения вдоль оси времени.

Вводится понятие функции расстояния динамического выравнивания между непрерывными временны́ми рядами, выравнивающего пути и его стоимости. В дискретном случае поиск выравнивающего пути осуществляется при помощи динамического программирования. В непрерывном же случае воспользоваться перебором невозможно, так как множество путей несчетно. Проблема поиска выравнивающего пути решена путем аппроксимации реального пути параметрической функцией. В качестве класса параметрических функций выбраны сплайны. Поиск пути сводится к поиску оптимальных параметров, задающих его приближение.

Универсальность данного подхода заключается в возможности применять различные способы аппроксимации временно́го ряда, а также аппроксимации пути наименьшей стоимости с наложением ограничений [1].

Работа выполнена при поддержке РФФИ, грант 16-07-01163.

[1] *Гончаров А. В., Стрижов В. В.* Метрическая классификация временных рядов акселерометра мобильного телефона со взвешенным выравниванием относительно центроидов классов // Информатика и ее применения, 2016. Т. 10. Вып. 2. С. 36–47. http://mi.mathnet.ru/rus/ia/v10/i2/p36.

Warping path for continuous time series alignment

Goncharov Alexey[1,2]⋆ alex.goncharov@phystech.edu
Strijov Vadim[2] strijov@ccas.ru
[1]Moscow, Russia, MIPT
[2]Moscow, Ruaaia, FRC CSC RAS

Representing discrete time series with continuous objects is a useful technique in the multiscale time series analysis because such measurements are difficult to analyze using only their discrete representation. In this paper, the metric method of time series analysis in continuous space, based on the dynamic time warping (DTW) distance measure, which performs the dynamic alignment between two time series, is presented. The DTW distance finds the best alignment between two time series if they are nonlinearly deformed relative to each other.

The DTW defines the distance between discrete objects. It is impossible to use the standard method to find the warping path in continuous space; so, the concept of DTW distance function between continuous time series, warping path between them, and its cost is introduced. The path searching problem is solved by approximating the path with parametric functions. Searching the path is equal to searching the best approximation. The versatility of this approach gives the ability to use any type for approximation of the warping path. A novel approach is introduced to analyze continuous time series. The properties of warping path and its cost are investigated in the experimental part. The metric classification problem is solved for describing the merging and splitting ability of new distance function in the space of the continuous functions. Centroids of classes are build and classification is made according to the nearest neighbor method [1].

This study was supported by the Russian Foundation for Basic Research, project 16-07-01163.

[1] Goncharov, A. V., and V. V. Strijov. 2016. Metric time series classification using weighted dinamic warping relative to centroids of classes. *Informatika i ee Primeneniya — Inform. Appl.* 10(2):36–47. Available at: http://www.ipiran.ru/journal/issues/article/19922264160204.html (accessed September 15, 2016).

Порядковые фильтры: некоторые аспекты обработки периодических сигналов

Знак Владимир Ильич znak@opg.sscc.ru
Россия, Новосибирск, ИВМиМГ СО РАН

Фильтры взвешенных порядковых статистик (weighted order statistics, WOS) обладают рядом достоинств по сравнению с наиболее распространенными — линейными фильтрами. Однако при обработке периодических сигналов на WOS фильтр накладывается ограничение: его отклик стремится к нолю, если длина выборки сигнала на входе фильтра приближается к целому числу периодов сигнала. Данное ограничение позволяют преодолевать так называемые ко-фазные порядковые фильтры [1]. Другая специфическая черта порядковых фильтров — их нелинейность, что обусловливает значительные трудности аналитической оценки их поведения. При этом их отклик зависит от ряда факторов. Таким образом, можно полагать, что качество обработки сигнала WOS фильтром есть случайная величина. В этих условиях внимание привлекает метод статистических испытаний для целей выбора наиболее эффективного проекта WOS фильтра. Итеративная обработка сигнала позволяет расширить возможности названного метода.

Периодические сигналы используются весьма широко. В рассматриваемом случае эффективность итеративной обработки сигнала в условиях селекции проектов WOS фильтров в процессе их статистических испытаний демонстрируется с привлечением данных реального эксперимента, полученных в процессе вибросейсмических исследований.

В перспективе рассматриваемая методология позволяет привлекать для выбора наиболее эффективного проекта WOS фильтра такие методы, как метод максимального правдоподобия или метод последовательных приближений.

[1] *Znak V.* Towards a statistical adaptation of order filters for processing periodic and frequency-modulated signals using a graphical interface // Pattern Recogn. Image Anal., 2015. Vol. 25. No. 2. P. 281–290. doi: 10.1134/S1054661815020273.

Order filters: Some aspects of the periodic signals processing

Znak Vladimir znak@opg.sscc.ru
Novosibirsk, Russia, ICMMG SB RAS

The weighted order statistics (WOS) filters possess a number of merits as compared to other ones, e. g., linear filters. However, the restriction is imposed on the WOS filters when processing periodic signals: input approaches an integer number of periods of a periodic signal. The given restriction allows to overcome the so-called co-phased order filters [1]. Another specific feature of the order filters is their nonlinearity that causes the considerable difficulties of an analytical estimation of their behavior. At the same time, their response depends on a number of factors. Thus, it is possible to suppose that the quality of processing a signal by the WOS filter is a casual case. In these conditions, the method of statistical trials is attractive for choosing the most effective WOS filter project. Iterative processing of a signal allows one to expand possibilities of the method described above.

Periodic signals are used sufficiently often. In the considered case, the efficiency of the iterative processing of a signal in the conditions of selecting a qualitative project of the WOS filter in the course of its statistical trials is demonstrated using the data obtained in real vibro-seismic investigations.

In the prospects, the use of such methods as the method of maximum likelihood or the method of sequential approaches are of interest with respect to selecting appropriate qualitative results.

[1] Znak, V. 2015. Towards a statistical adaptation of order filters for processing periodic and frequency-modulated signals using a graphical interface. *Pattern Recogn. Image Anal.* 25(2):281–290. doi: 10.1134/S1054661815020273.

Моделирование и анализ вариаций космических лучей в периоды повышенной солнечной и геомагнитной активности

Мандрикова Оксана Викторовна* oksanam1@mail.ru
Заляев Тимур Ленарович tim.aka.geralt@mail.ru
Полозов Юрий Александрович up_agent@mail.ru
Соловьев Игорь Сергеевич kamigsol@yandex.ru
Петропавловск-Камчатский, Россия, ИКИР ДВО РАН

Описан способ анализа вариаций космических лучей, позволяющий выделять аномальные изменения и получать количественные оценки о моментах их возникновения, временно́й длительности и интенсивности. Способ включает декомпозиции данных нейтронных мониторов на основе вейвлет-преобразования и их аппроксимацию на основе адаптивных нейронных сетей переменной структуры. На основе применения способа выполнен анализ вариаций космических лучей в периоды повышенной солнечной и геомагнитной активности и выделены аномальные изменения, возникающие за несколько часов до геомагнитных бурь, во время бурь происходили длительные и глубокие Форбуш-понижения (анализировались данные нейтронных мониторов станций Апатиты и Мыс Шмидта). Совместно с данными космических лучей анализировались вариации геомагнитного поля и ионосферные параметры, обработка которых выполнялась на основе методов, предложенных авторами [1].

Работа поддержана грантом Российского научного фонда № 14-11-00194.

[1] *Мандрикова О. В., Заляев Т. Л., Полозов Ю. А., Соловьев И. С.* Моделирование и анализ вариаций космических лучей в периоды повышенной солнечной и геомагнитной активности // Машинное обучение и анализ данных, 2016 (в печати).

Modeling and analysis of cosmic ray variations during periods of increased solar and geomagnetic activity

Mandrikova Oksana⋆ oksanam1@mail.ru
Zalyaev Timur tim.aka.geralt@mail.ru
Polozov Yuriy up_agent@mail.ru
Solov'ev Igor' kamigsol@yandex.ru

Russia, Petropavlovsk-Kamchatskiy, IKIR FEB RAS

A method for analysis of cosmic ray variations, which allows to allocate anomalous changes and obtain quantitative estimates of their occurrence time, duration, and intensity is described. The method includes decomposition of neutron monitor data based on wavelet transform and their approximation based on adaptive variable structure neural networks. By using this method, analysis of cosmic ray variations during periods of increased solar and geomagnetic activity is performed and anomalous changes that occurred a few hours before geomagnetic storms are allocated. Long and deep Forbush decreases took place during the storms (neutron monitor data from Apatity and Cape Schmidt stations were analysed). Cosmic ray data were analysed together with geomagnetic field variations and ionospheric parameters, which processing was performed on the basis of methods proposed by the authors [1].

This research is supported by the grant of the Russian Science Foundation No. 14-11-00194.

[1] Mandrikova, O., T. Zalyaev, Yu. Polozov, and I. Solov'ev. 2016 (in press). Modeling and analysis of cosmic ray variations during periods of increased solar and geomagnetic activity. *Machine Learning Data Anal.*

Порождение признаков в задаче прогнозирования набора разномасштабных временных рядов

Мотренко Анастасия Петровна[1]
anastasiya.motrenko@phystech.edu
Нейчев Радослав Георгиев[1] neychev@phystech.edu
Исаченко Роман Владимирович[1] roman.isachenko@phystech.edu
Попова Мария Сергеевна[1] maria_popova@phystech.edu
Громов Андрей Николаевич[2] agromov@forecsys.ru
Стрижов Вадим Викторович[2]⋆ strijov@ccas.ru
[1]Россия, Москва, МФТИ
[2]Россия, Москва, ФИЦ ИУ РАН

Предлагается подход к прогнозированию набора разномасштабных временны́х рядов. Рассматривается задача прогнозирования состояния устройства Интернета вещей. Устройство снабжено набором сенсоров, генерирующих временны́е ряды различного масштаба с различной частотой сэмплирования. Требуется спрогнозировать значения каждого временно́го ряда в заданном промежутке времени.

Данная задача сводится к задаче регрессии. Предлагается метод построения признакового описания для регрессионной задачи, основанный на порождении заведомо избыточного набора признаков с последующим отбором признаков. Порожденные признаки включают предысторию всех временны́х рядов из набора и их локальные преобразования. Применение предлагаемого подхода рассмотрено на примере нескольких регрессионных алгоритмов. Исследовано качество прогнозов в зависимости от горизонта прогнозирования [1].

Работа поддержана грантом РФФИ № 16-07-01160.

[1] *Нейчев Р. Г., Катруца А. М., Стрижов В.* Выбор оптимального набора признаков из мультикоррелирующего множества в задаче прогнозирования // Заводская лаборатория. Диагностика материалов, 2016. Т. 82. № 3. С. 68–74.

Feature generation for multiscale time series forecasting

Motrenko Anastasia[1] anastasiya.motrenko@phystech.edu
Neychev Radoslav[1] neychev@phystech.edu
Isachenko Roman[1] roman.isachenko@phystech.edu
Popova Maria[1] maria_popova@phystech.edu
Gromov Andrey[2] agromov@forecsys.ru
Strijov Vadim[2]⋆ strijov@ccas.ru
[1]Moscow, Russia, MIPT
[2]Moscow, Russia, FRC CSC RAS

The paper presents a framework for the massive multiscale time series forecast. The focus is on the problem of forecasting behavior of a device within the concept of Internet of things. The device is monitored by a set of sensors, which produces large amount of multiscale time series during its lifespan. These time series have various time scales since distinct sensors produce observations with various frequencies from milliseconds to weeks. The main goal is to predict the observations of a device in a given time range.

The authors propose a method of constructing efficient feature description for the corresponding regression problem. The method involves feature generation and dimensionality reduction procedures. Generated features include historical information about the target time series as well as other available time series, local transformations, and multiscale features. Several forecasting algorithms have been applied to the resulting regression problem and the quality of the forecasts has been investigated for various horizon values [1].

This research is funded by the Russian Foundation for Basic Research, grant 16-07-01160.

[1] Neychev, R., A. Katrutsa, and V. Strijov. 2016. Slecting optimal subset from a set of multicollinear features in forecasting problem. *Industrial Laboratory* 82(3):68–74.

Прогностические мультимодели разномасштабных временных рядов Интернета вещей

Нейчев Радослав Георгиев [1]⋆ neychev@phystech.edu
Мотренко Анастасия Петровна[1] anastasiya.motrenko@phystech.edu
Исаченко Роман Владимирович[1] roman.isachenko@phystech.edu
Инякин Андрей Сергеевич[2] inyakin@forecsis.ru
Стрижов Вадим Викторович[2] strijov@ccas.ru
[1]Россия, Москва, МФТИ
[2]Россия, Москва, ФИЦ ИУ РАН

Решается задача прогнозирования большого числа взаимосвязанных временны́х рядов. Их источником служат датчики мониторинга устройств Интернета вещей. Предполагается, что пространство параметров описания временны́х рядов неоднородно, выборка не является простой.

Задача построения прогноза сводится к задаче регрессии. Для получения точного и устойчивого прогноза предлагается использовать смесь экспертов — прогностических моделей для ее решения. В качестве моделей используются нейронные сети. Исследуются задачи оптимизации пространства параметров нейронных сетей, выбора нейронных сетей оптимальной сложности, выбора оптимального числа экспертов. В ходе вычислительного эксперимента сравниваются три класса моделей: смесь экспертов, градиентный бустинг, решающие деревья. Эксперимент выполнен на реальных данных, содержащих информацию о потреблении электроэнергии и погодных условиях в Польше [1].

Работа поддержана грантом РФФИ № 16-07-01158.

[1] *Нейчев Р. Г., Катруца А. М., Стрижов В. В.* Выбор оптимального набора признаков из мультикоррелирующего множества в задаче прогнозирования // Заводская лаборатория. Диагностика материалов, 2016. Т. 82. № 3. С. 68–74.

Multimodel forecasting multiscale time series in Internet of things

Neychev Radoslav[1]⋆ neychev@phystech.edu
Motrenko Anastasia[1] anastasiya.motrenko@phystech.edu
Isachenko Roman[1] roman.isachenko@phystech.edu
Inyakin Andrey[2] inyakin@forecsis.ru
Strijov Vadim[2] strijov@ccas.ru
[1]Moscow, Russia, MIPT
[2]Moscow, Russia, FRC CSC RAS

The paper presents an approach to forecasting multiple intercorrelated time series that can be generated by different sensors of devices within a concept of Internet of things. In this case, generated data are not independent and identically-distributed and there feature space has a complex structure.

The forecast construction is considered as regression problem. To solve it, the authors propose mixture of experts approach where several forecasting models are used. Neural networks are chosen as the forecasting models. The optimal structure of neural networks, their parameters, and quantity of experts are analyzed. The proposed method has been tested within computational experiment where it was compared to gradient boosting and decision tree methods. The experiment was conducted on real data containing information about electricity consumption and weather conditions in Poland [1].

This research is funded by the Russian Foundation for Basic Research, grant 16-07-01158.

[1] Neychev, R., A. Katrutsa, and V. Strijov. 2016. Selecting optimal subset from a set of multicollinear features in forecasting problem. *Industrial Laboratory* 82(3):68–74.

Поиск плавно меняющихся моделей оценки вероятности дефолта

Филипенков Николай Владимирович[1]⋆ n.filipenkov@mail.ru
Петрова Марина Алексеевна[2] marina_petrova@mail.ru
[1]Россия, Москва, CAC институт
[2]Россия, Москва, НИЯУ МИФИ

В международной банковской практике задача разработки моделей оценки вероятности дефолта является крайне актуальной. При этом различают PIT-модели (от Point-In-Time), которые характеризуют кредитный риск заемщика на текущий момент, и TTC-модели (Through-The-Cycle), оценивающие кредитный риск контрагента на протяжении всего экономического цикла.

PIT-модели используются при принятии решения о выдаче кредита на данный момент и в расчете ожидаемых потерь в соответствии со стандартом МСФО-9. TTC-модели используются при оценке регуляторного капитала, в том числе в соответствии с требованиями Базельского Комитета.

Традиционные методы преобразования PIT-моделей в TTC и наоборот обычно реализуются путем добавления дополнительных параметров калибровки. Такие подходы не позволяют выявлять структурные изменения модели оценки вероятности дефолта, например усиление одних характеристик в кризисные годы и ослабление других характеристик в тот же временной период.

В настоящей работе предлагается подход, основанный на идее поиска плавно меняющихся закономерностей, изложенной в предыдущих работах авторов. Идея подхода предполагает поиск закономерностей на каждом из участков экономического цикла, а затем объединение их в единую модель на основе мер сходства закономерностей. Таким образом, предлагается строить TTC-модель как последовательность плавно меняющихся PIT-моделей. Такой подход позволяет учитывать структурные изменения модели с течением времени [1].

Работа поддержана грантом РФФИ № 16-07-01156.

[1] *Филипенков Н. В., Петрова М. А.* Поиск плавно меняющихся моделей оценки вероятности дефолта // Машинное обучение и анализ данных, 2016 (в печати).

Building slightly changing probability of default models

Filipenkov Nikolay[1]⋆ n.filipenkov@mail.ru
Petrova Marina[2] marina_petrova@mail.ru
[1]Moscow, Russia, SAS Institute
[2]Moscow, Russia, MEPhI

Building PD (Probability of Default) models is a key area for risk management in the financial institutions worldwide. There are different types of PD models. Point-In-Time (PIT) PD-models reflect the borrower's credit risk at a certain moment of time. Through-The-Cycle (TTC) PD-models estimate the borrower's credit risk through the whole economic cycle.

The PIT models are widely used at the moment of the bank's decision to lend the money. IFRS-9 also requires banks to calculate Expected Credit Losses (ECL) based on PIT models. The TTC models are used when calculating the regulatory capital based on Basel Committee standards and local regulations.

The traditional methodologies transforming PIT-models to TTC-models and *vice versa* normally use different calibration parameters. These approaches are limited in identifying the structural changes in PD-model, e. g., the increase of information value of some variables and the decline of others during the same time period.

In this paper, the approach to build TTC PD-models is introduced based on the idea of mining slightly changing patterns developed in the authors' previous papers. The idea implies mining PIT PD-models on each segment of the economic cycle and then uniting them in one TTC PD-model based on the model similarity measure. So, TTC PD-model is combined of slightly changing PIT PD-models. This approach allows one to find the structural changes in the models through the economic cycle [1].

This research is funded by the Russian Foundation for Basic Research, grant No. 16-07-01156.

[1] Filipenkov, N. V., and M. A. Petrova. 2016 (in press). Building the changing probability of default models. *Machine Learning Data Anal.*

О роли суммирования Фейера при моделировании рельефа

Флоринский Игорь Васильевич* iflor@mail.ru
Панкратов Антон Николаевич pan@impb.ru
Пущино, Россия, ИМПБ РАН — филиал ИПМ им. М. В. Келдыша РАН»

Разработан спектрально-аналитический метод моделирования рельефа [1]. Метод предназначен для обработки цифровых моделей рельефа (ЦМР) в рамках единой схемы, включающей глобальную аппроксимацию ЦМР, генерализацию, подавление шума, а также расчет частных производных высоты и морфометрических характеристик. Метод основан на ортогональных разложениях высокого порядка с использованием полиномов Чебышёва с суммированием Фейера (СФ). Для оценки роли СФ была использована ЦМР Кении (матрица 480×481, разрешение 30"). При аппроксимации ЦМР с СФ и без СФ применялись различные наборы коэффициентов разложения (от 30 до 7000). По восстановленным ЦМР рассчитаны модели горизонтальной и вертикальной кривизн, а также получены модели невязок восстановленных ЦМР. Аппроксимации с СФ и без СФ характеризуются монотонной сходимостью, но скорость сходимости аппроксимации без СФ выше. Невязки для ЦМР, восстановленных, например, с 2880 коэффициентами разложения без СФ, не превышают 1,3% от диапазона высот в исходной ЦМР. Однако на картах кривизн, рассчитанных по ЦМР, которые были восстановлены без СФ, хорошо видны осциллирующие артефакты (явление Гиббса). Восстановленные ЦМР также имеют эти артефакты, но они выражены слабо и не видны на картах высот. Дифференцирование усиливает их выраженность на картах кривизн. Вместе с тем, эти артефакты не возникают на картах кривизн, если аппроксимация ЦМР проводилась с СФ. Таким образом, СФ, подавляя эти вычислительные шумы, является необходимым этапом моделирования рельефа.

Работа поддержана грантом РФФИ № 15-07-02484.

[1] *Florinsky I. V., Pankratov A. N.* A universal spectral analytical method for digital terrain modeling // Int. J. Geogr. Inf. Sci., 2016, Vol. 30. doi: 10.1080/13658816.2016.1188932.

On the role of the Fejér summation in terrain modeling

Florinsky Igor ⋆ iflor@mail.ru
Pankratov Anton pan@impb.ru
Russia, Pushchino, IMPB RAS — Branch of KIAM RAS

A spectral analytical method for terrain modeling is developed [1]. The method is intended for the processing of digital elevation models (DEMs) within a single framework, including DEM global approximation, denoising, generalization, as well as calculating the partial derivatives of elevation and morphometric variables. The method is based on high-order orthogonal expansions using the Chebyshev polynomials with the Fejér summation (FS). To evaluate a role of FS, a DEM of Kenya was used (the matrix 480×481, resolution 30"). Various sets of expansion coefficients (from 30 to 7000) were evaluated to approximate DEMs with and without FS. The models of horizontal and vertical curvatures were computed from the reconstructed DEMs. Residual models for the reconstructed DEMs were also calculated. Approximations with and without FS are marked by a distinct monotonic convergence, but the convergence rate of the FS-free approximation is higher. The residuals for the DEM reconstructed with, for example, 2880 expansion coefficients without FS does not exceed 1.3% of the elevation range in the initial DEM. However, pronounced oscillatory artifacts (the Gibbs phenomenon) emerged on the maps of curvatures derived from the FS-free approximated DEMs. The reconstructed DEMs include these oscillations but they are weakly expressed and cannot be seen on the elevation maps. Differentiation amplifies their manifestation on the curvature maps. However, such artifacts do not appear on the curvature maps if the DEM approximation included FS. Thus, FS suppressing this computational noise is the necessary stage of the terrain modeling.

This research is funded by the Russian Foundation for Basic Research, grant 15-07-02484.

[1] Florinsky, I. V., and A. N. Pankratov. 2016. A universal spectral analytical method for digital terrain modeling. *Int. J. Geogr. Inf. Sci.* 30. doi: 10.1080/13658816.2016.1188932.

Анализ гиперсинхронизации структур головного мозга во время эпилептических разрядов на основе конических представлений сигнала электроэнцефалограммы

Анциперов Вячеслав Евгеньевич[1,2]⋆ antciperov@cplire.ru
Обухов Юрий Владимирович[1] obukhov@cplire.ru
[1]Россия, Москва, ИРЭ им. В. А. Котельникова РАН
[2]Россия, Долгопрудный, МФТИ

Обсуждаются результаты применения основанного на многомасштабном корреляционном анализе подхода к задачам оценки ритмов головного мозга [1]. Для анализа эпилептических разрядов, находящихся в центре внимания работы, предложенное ранее однопараметрическое (частотное) представление расширено до двухпараметрического (масштабно-частотного). Это сделано на основе обобщенной модели эпилептических разрядов как класса широкополосных импульсных сигналов с переменным ритмом следования и, возможно, изменяющейся формой импульсов.

Синтезированное для целей анализа электроэнцефалограмм (ЭЭГ) масштабно-частотно-временно́е представление является квадратичным представлением конического типа. Семейство распределений конического типа характеризуется свойством эффективного подавления интерференций между частотными компонентами. Также обнаружено, что синтезированное в работе представление хорошо подчеркивают нестационарности и переходные процессы, связанные, в частности, с перестройками спектрального состава сигнала.

Приведены результаты применения предложенного подхода к анализу реальных ЭЭГ-записей, содержащих фрагменты эпилептических разрядов.

Работа выполнена в рамках проекта «Разработка частотно-временны́х моделей судорожной электрической активности мозга» программы Призидиума РАН №I.33П.

[1] *Анциперов В. Е., Обухов Ю. В., Кузнецова Г. Д., Гнездицкий В. В.* Анализ гиперсинхронизации структур головного мозга во время эпилептических разрядов на основе специальных конических представлений ЭЭГ сигнала // Ж. радиоэлектроники, 2014. №11. 20 с. http://jre.cplire.ru/jre/nov14/18/text.pdf.

Analysis of brain structures hypersynchronization during the epileptic discharges on the basis of conical kernel representations of electroencephalogram signal

Antsiperov Viacheslav[1,2]⋆ antciperov@cplire.ru
Obukhov Yury[2] obukhov@cplire.ru
[1]Moscow, Russia, Kotel'nikov IRE RAS
[2]Dolgoprudny, Russia, MIPT

The paper is devoted to the developing of a new approach to the problem of cortical rhythms analysis [1]. For the analysis of epileptic discharges, which are in the focus of the work, one-parameter (frequency) representation is extended to the two-parameter (scale–frequency) representation. This was done on the basis of a generalized model of epileptic discharges as a class of broadband pulse signals with variable rhythm and, perhaps, repeatedly changing pulse waveform.

Synthesized for the aim of electroencephalogram (EEG) analysis, a scale–frequency–time representation is a quadratic cone kernel. It belongs to a great family of distributions characterized by the effective reduction of spurious interference (RID) between frequency components. The authors have also discovered that synthesized representation can emphasize unsteadiness and transients associated, in particular, with the restructuring of the signal specter.

Also, the results of the analysis of real, containing the fragments of an epileptic discharges EEG records, are presented.

The work is done in the framework of project "Development of time–frequency models of convulsive electrical activity of the brain" of the Presidium RAS program No. I.33P.

[1] Antsiperov, V. E., Y. V. Obukhov, G. D. Kuznetsova, and V. V. Gnezditskiy. 2014. Analysis of brain structures hypersynchronization during the epileptic discharges on the basis of conical kernel representations of EEG signal. *J. Radioelectronics* 11. 20 p. Available at: http://jre.cplire.ru/jre/nov14/18/text.pdf (accessed July 26, 2016).

Анализ магнитоэнцефалографических данных с применением методов спектрального анализа и теории случайных полей

Буторина Анна Валерьевна[1]⋆ armature@yandex.ru
Литвак Владимир[2] v.litvak@ucl.ac.uk
[1]Россия, Москва, МЭГ-центр, МГППУ
[2]Великобритания, Лондон, Велкам Траст Центр Нейроимаджинга — УКЛ

В данной работе авторы использовали спектральный анализ с мультитаперами и теорию случайных полей для поправки на множественные сравнения, чтобы показать наличие высокочастотной гамма-активности в МЭГ-сигнале в психофизиологическом эксперименте по изучению эффекта зеркальной руки. Магнитоэнцефалография (МЭГ) — это неинвазивный метод для регистрации слабых магнитных полей, порождаемых активностью нейронов в мозге человека. Из инвазивных работ известно, что высокочастотная активность (40–100 Гц) возникает очень локально в областях коры головного мозга, контролирующих реальные движения конечностей. Эффект зеркальной руки заключается в том, что перед человеком устанавливается зеркало так, что оно загораживает одну руку, а вторая рука отражается в зеркале на месте первой. Рука, спрятанная зеркалом, остается неподвижной, а наблюдение за движением видимой руки и ее отражения в зеркале вызывает у испытуемого ощущение движения обоими руками. Для навигации в пространстве и частоте авторы использовали данные реального движения рукой. При помощи мультитаперов сигнал был переведен в пространственно-частотное представление и, используя теорию случайных полей, определены сенсоры и частотная полоса, где сигнал при движении был неслучайно больше, чем до начала движения. Это были сенсоры над сенсомоторной корой и частотная полоса 55–85 Гц. При сравнении сигналов, отфильтрованных в полученной высокочастотной полосе от движения одним пальцем и от движения этим же пальцем, но с зеркалом, была получена статистически значимая разница сразу после начала движения [1].

[1] *Butorina A., Prokofyev A., Nazarova M., Litvak V., Stroganova T.* The mirror illusion induces high gamma oscillations in the absence of movement // NeuroImage, 2014. Vol. 103. P. 181–191.

Use of spectral analysis and random field theory for magnetoencephalography data analysis

Butorina Anna[1]⋆ armature@yandex.ru
Litvak Vladimir[2] v.litvak@ucl.ac.uk
[1]Moscow, Russia, MEG-center, MSUPE
[2]London, U.K., Wellcome Trust Centre for Neuroimaging — UCL

Multitaper spectral analysis and random field theory were used to test whether mirror hand illusion induced high gamma oscillation. Magnetoencephalography is a noninvasive technique for mapping brain activity by recording magnetic fields produced by electrical currents occurring naturally in the brain. In invasive researches, high gamma power (40–100 Hz) responses following the movements of somatotopically defined regions in the sensorimotor cortex consistent with maps generated by cortical electrical stimulation. Mirror hand phenomenon refers to the illusory percept of moving a hand while moving the opposite hand and viewing its reflection in a mirror. To induce the illusion, the mirror is placed sagittally giving the impression that the stationary hand is performing the task. The real movement data were used to define the spatiospectral region of interest. For efficient spectral estimation from a relatively small number of trials, a multitaper spectral analysis was used. Then, random field theory was used to solve the multiple comparison problem which is common in functional imaging. As a result, 55–85-hertz high gamma activity was maximal at the pair of sensorimotor sensors. Those sensors were used for the subsequent analysis. Significant differences in high gamma power between unilateral movement and unilateral movement with mirror were found in postmovement time window [1].

[1] Butorina, A., A. Prokofyev, M. Nazarova, V. Litvak, and T. Stroganova. 2014. The mirror illusion induces high gamma oscillations in the absence of movement. *NeuroImage* 103:181–191.

Алгоритм детектирования эпилептических разрядов и сонных веретен у крыс в раннем посттравматическом периоде

Кершнер Иван Андреевич[1,2]⋆ ivan.kershner@gmail.com
Обухов Юрий Владимирович[1] yuvobukhov@mail.ru
Комольцев Илья Геральдович[3] outaudiofillin@gmail.com
[1]Москва, Россия, ИРЭ им. В. А. Котельникова РАН
[2]Москва, Россия, МФТИ
[3]Москва, Россия, ИВНДиНФ РАН

Проводилось исследование многосуточных записей электроэнцефалограмм (ЭЭГ) крыс, полученных с вживленных электродов до и после черепно-мозговой травмы.

В качестве объектов исследования были выбраны эпилептические разряды и сонные веретена, возникающие на короткое время в ЭЭГ. Для их изучения нейрофизиологами вручную были отобраны характерные эпилептические разряды и сонные веретена. Для них был разработан алгоритм детектирования, который позволяет выделить событие (эпилептический разряд или сонное веретено) из фона. После детектирования собиралась информация об этом событии: частотный диапазон, длительность и максимальная спектральная плотность мощности, в котором он находится.

Полученные данные в дальнейшем могут служить опорой для автоматического распознавания эпилептических разрядов среди сонных веретен, а также для нахождения событий на длительной записи ЭЭГ.

Исследование выполнено за счет гранта Российского научного фонда (проект № 16-11-10258).

Detection algorithm of epileptic discharges and sleep spindles in rats in early posttraumatic period

Kershner Ivan[1,2]⋆ ivan.kershner@gmail.com
Obukhov Yury[1] yuvobukhov@mail.ru
Komoltsev Ilya[3] outaudiofillin@gmail.com
[1]Russia, Moscow, Kotel'nikov IRE RAS
[2]Russia, Moscow, MIPT
[3]Russia, Moscow, IHNA and NPh RAS

Multiple recordings of the electroencephalograms (EEG) of rats obtained from implanted electrodes before and after traumatic brain injury was conducted.

Epileptic discharges and sleep spindles occured for a short time in the EEG were chosen as research objects. Typical epileptic discharges and sleep spindles were manually selected by neurophysiologists. Detection algorithm which allows selecting event (epileptic discharge or sleep spindles) from the background was created by this sample. After the process of event detection, information about them such as frequency range, duration, and maximum of power spectral density were collected.

The obtained data can further serve as a basis for automatic recognition of epileptic discharges among sleep spindles as well as for finding events from long EEG recordings.

This research is funded by the Russian Science Foundation, grant 16-11-10258.

Множественный дискриминантный анализ для распознавания биосигналов в частотной области

Манило Людмила Алексеевна* lmanilo@yandex.ru
Немирко Анатолий Павлович apn-bs@yandex.ru
Россия, Санкт-Петербург, СПбГЭТУ «ЛЭТИ»

Рассмотрен метод распознавания электрокардиосигналов по спектральному описанию с применением множественного дискриминантного анализа. Обосновывается выбор весовых функций, приближающих критерий Фишера к оценке точности классификации объектов. Для задачи распознавания биосигналов по параметрам нормированного спектра получена формула вычисления весовых коэффициентов. Особое внимание уделено процедуре преобразования пространства спектральных признаков, которая позволяет сократить его размерность и за счет этого упростить алгоритм построения решающих функций. Приводятся результаты экспериментальных исследований, направленных на распознавание разных видов опасных аритмий. В частности, решается задача обнаружения по электрокардиограмме желудочковой фибрилляции, а также распознавания аритмий, являющихся предвестниками опасных нарушений. Показано, что использование дискриминанта Фишера с учетом весовых функций позволяет уменьшить ошибки классификации. Результаты исследований использованы при решении практических задач кардионаблюдения [1].

Работа поддержана грантами РФФИ 15-07-01790 и 16-01-00159 и медицинским проектом CardioQVARK — кардиограмма с помощью телефона (www.cardioqvark.ru).

[1] *Манило Л. А.* Множественный дискриминантный анализ для распознавания электрокардиосигналов в частотной области // Биомедицинская радиоэлектроника, 2016 (в печати). № 9.

Multiple discriminant analysis for recognition of biosignals in frequency domain

Manilo Ludmila* lmanilo@yandex.ru
Nemirko Anatolii apn-bs@yandex.ru
Saint Petersburg, Russia, ETU

The problem of linear decision functions creation for several classes of biomedical signals recognition is considered. Importance of the solution for the problem of dangerous arrhythmias detection at early stages of their occurrence is shown. The conclusion about the expediency of the electrocardiogram analysis in frequency domain with application of the multiple discriminant analysis is drawn. For the purpose of approach of Fischer's criterion to the groups of signals classification accuracy, it is suggested to use weight functions. For the problem of electrocardiogram recognition using the parameters of a normalized spectrum the formula of weight coefficients calculation is received. The special attention is paid to the procedure of spectral feature space transformation which allows to reduce its dimension and to simplify the algorithm of decision functions creation. The results of the experimental research on different types of dangerous arrhythmias recognition are given. In particular, the problems of ventricular fibrillation detection and also the recognition of arrhythmias which are harbingers of dangerous violations by electrocardiogram are solved. It is shown that use of the Fischer's discriminant taking into account the weight functions allows to reduce classification errors. The conclusion about high reliability of ventricular fibrillation detection according to the electrocardiogram spectral description is drawn [1].

This research is funded by the Russian Foundation for Basic Research, grants 15-07-01790 and 16-01-00159 and by medical project CardioQVARK — Cardiogram by Phone (www.cardioqvark.ru).

[1] Manilo, L. 2016 (in press). Multiple discriminant analysis for recognition of electrocardiosignals in frequency domain. *Biomed. Radioelectronics* 9.

Метрическая классификация раннего паркинсонизма в пространстве электрофизиологических признаков

Обухов Константин Юрьевич[1]⋆ k.obukhov@mail.ru
Малюта Инна Александровна[1] inna.maliuta@mail.ru
Обухов Юрий Владимирович[2] yuvobukhov@mail.ru
[1]Москва, Россия, МФТИ
[2]Москва, Россия, ИРЭ им. В. А. Котельникова РАН

Описаны новые методы анализа динамики электроэнцефалограмм ЭЭГ, электромиограмм (ЭМГ) и тремора, получены при их помощи новые маркеры диагностики ранней стадии болезни Паркинсона (БП) и представлены результаты классификации раннего паркинсонизма по этим маркерам.

Проведено ЭЭГ-обследование 73 людей, из них 42 пациента с различными формами БП и 31 человек группы контроля. Также проведены совместные обследования ЭЭГ, ЭМГ и тремора другой группы, состоящей из 31 пациента с БП и 18 человек из контрольной группы.

Анализ сигналов выполнялся с помощью вейвлет-преобразования Морле. Для количественного анализа динамики частотно-временны́х распределений локальных максимумов использовались динамические и интегральные гистограммы экстремумов. Для классификации БП на начальной стадии использовалась логистическая модель бинарной классификации в пространстве ЭЭГ признаков БП, в которой зависимая переменная принимает множество значений $Y = \{0, 1\}$, где 0 означает «здоров», а 1 — «болен». Совпадение диагнозов начальной стадии БП по набору признаков ЭЭГ с клиническими диагнозами составляет 78%, а по набору признаков ЭЭГ и тремора превышает 90% [1].

Работа поддержана грантом РФФИ № 15-07-07846.

[1] *Сушкова О. С., Габова А. В., Карабанов А. В., Кершнер И. А., Обухов К. Ю., Обухов Ю. В.* Метод частотно временного анализа совместных измерений электроэнцефалограмм, электромиограмм и механического тремора при болезни Паркинсона // Радиотехника и электроника, 2015. Т. 60. № 10. С. 1064–1072.

Classification of early stage Parkinson's disease based on electroencephalography feature set

Obukhov Konstantin[1]⋆ k.obukhov@mail.ru
Maliuta Inna[1] inna.maliuta@mail.ru
Obukhov Yury[2] yuvobukhov@mail.ru
[1]Russia, Moscow, MIPT
[2]Russia, Moscow, Kotel'nikov IRE RAS

The new methods of dynamical analysis of electroencephalography (EEG), electromyography (EMG), and tremor are presented. According to these methods, the new features of early stage Parkinson's disease (PD) were extracted and disease classification model based on these features was built.

The EEG analysis of 73 people was performed, among which there were 42 patients with various forms of PD and 31 people from the control group. Moreover, joint analysis of EEG, EMG, and tremor was performed based on other group, which contained 31 patients with PD and 18 people from the control group.

The signal analysis was done via wavelet transform with Morlet mother function. Dynamical and integral histograms were used for quantitative analysis of time-frequency dynamics of the local extrema. Binary logistic regression model was built as a disease classification model in EEG features space. The binary target variable from the set of values $Y = \{0, 1\}$ was used where 0 indicates nondisease observation and 1 indicates a disease observation. The overlap of EEG-based diagnosis and clinical diagnosis is 78%, and in case of EEG- and tremor-based diagnosis is higher than 90% [1].

This research is funded by the Russian Foundation for Basic Research, grant 15-07-07846.

[1] Sushkova, O., A. Gabova, A. Karabanov, I. Kershner, K. Obukhov, and Yu. Obukhov. 2015. Time-frequency analysis of simultaneous measurements of electroencephalograms, electromyograms, and mechanical tremor under Parkinson disease. *J. Comm. Technol. Electronics* 60(10):1064–1072.

Об одном подходе к классификации сонных веретен и эпилептических разрядов в электроэнцефалограмме после черепно-мозговых травм

Обухов Константин Юрьевич[1] k.obukhov@mail.ru
Малюта Инна Александровна[1]⋆ inna.maliuta@mail.ru
Обухов Юрий Владимирович[2] yuvobukhov@mail.ru
[1]Россия, Москва, МФТИ
[2]Россия, Москва, ИРЭ им. В. А. Котельникова РАН

Описан новый метод различения сонных веретен и эпилептических разрядов в сигналах электроэнцефалограмм (ЭЭГ) после черепно-мозговых травм у экспериментальных животных (крыс), являющихся клинически релевантной моделью травматического повреждения мозга. Для детектирования сонных веретен и разрядов используются хребты вейвлет спектрограмм. Определение схожих сигналов сводится к их кластеризации методом k-means в N-мерном пространство, где каждое измерение служит соответствующей временно́й компоненте мощности, либо частоты сигнала. Центроиды кластеров могут быть обратно отображены в характерный для кластера сигнал. Выявлено, что различие в кривых центроид мощности наиболее характерно в третьем и четвертом каналах после травмы. За метрику дисперсии предложено взять отношение среднеквадратичного отклонения частоты к среднему значению частоты для каждого из сигналов. По записям, произведенным через один день после травмы, выявлено различие в распределении метрики. После проведения кластеризации по мощности сигналов было получено два кластера, распределения относительного отклонения частоты в которых практически не перекрывались в третьем и четвертом каналах [1].

Исследование выполнено за счет гранта Российского научного фонда, проект № 16-11-10258.

[1] *Кершнер И. А., Комольцев И. Г., Малюта И. А., Обухов К. Ю., Обухов Ю. В., Тишкина А. О.* Об одном подходе к детектированию и классификации сонных веретен и эпилептических разрядов в ЭЭГ после черепно-мозговых травм // Pattern Recogn. Image Anal., 2016 (в печати).

The classification method of sleep spindles and epilepsy seizures in electroencephalograms after traumatic brain injury

Obukhov Konstantin[1] k.obukhov@mail.ru
Maliuta Inna[1]⋆ inna.maliuta@mail.ru
Obukhov Yury[2] yuvobukhov@mail.ru
[1]Moscow, Russia, MIPT
[2]Moscow, Russia, Kotel'nikov IRE RAS

The new method of electroencephalography (EEG) signal classification between sleep spindles and epilepsy seizures after traumatic brain injury is described. The EEG data was measured on experimental rats. The traumatic brain injury can be considered as a relevant clinical model of post-traumatic epilepsy. For the detection of sleep spindles and seizures the wavelet spectrogram ridges were used.

The identification of similar signals starts from the mapping of time-dependent data to N-space dimension, where each space is dedicated to a specific value of signal power or frequency. In this N-dimensional space, k-means clustering algorithm is used to cluster the data in predefined number of classes. The centroids of these clusters can be reversely transformed to time-dependent data of signals power or frequency. It was found that the difference in the shape of centroids is more significant in the 3rd and 4th channels after the brain injury. The relation of frequency standard deviation to mean was selected as a metric of this variance for each signal. According to the initial data before injury, there is no significant difference between sleep spindles and seizures. However, the variance becomes visible after the brain injury. The significant difference between spindles and seizures was found after clustering in signals power. The two clusters were built, and the relative deviation distributions did not overlap in the pair of clusters in the 3rd and 4th channels [1].

This research is funded by the Russian Science Foundation, project No. 16-11-10258.

[1] Kershner, I., I. Komoltsev, I. Maliuta, K. Obukhov, Yu. Obukhov, and A. Tishkina. 2016 (in press). The discovery and classification method of sleep spindels and epilepsy sezures in EEG after traumatic brain injury. *Pattern Recogn. Image Anal.*

Методы интеллектуального анализа квазипериодических биосигналов в задачах оценки состояния человека-оператора

Покровская Ирина Вячеславовна* ivp750@mail.ru
Гучук Владимир Всеволодович polma@bk.ru
Десова Аэлита Арсеньевна achern@ipu.ru
Дорофеюк Александр Александрович daa2@mail.ru
Россия, Москва, ИПУ РАН

Работа посвящена оценке психофизилогического состояния человека-оператора, которое во многом определяется степенью его утомления. Разработан методологический подход, состоящий из трех этапов: (1) формирование критерия исследуемого состояния оператора для конкретной области его деятельности; (2) определение набора информативных показателей, характеризующих состояние оператора; (3) построение решающего правила для определения типа состояния оператора, при этом существенно используется пульсовой сигнал лучевой артерии (ПС) оператора [1]. Для решения проблемы (1) разработана методика многовариантной экспертизы, для решения проблемы (2) авторы используют имеющийся задел по выявлению информативных показателей из ПС, для решения проблемы (3) используются современные схемы и процедуры распознавания образов. Описаны методы выделения функционально-значимых элементов ПС, а также алгоритмы спектрального анализа колебательных компонент динамических рядов ПС.

В настоящее время проводятся работы по сбору массива данных работы человека-оператора в экспериментальных условиях с регулируемой нагрузкой и сложностью, на базе которого будут уточнены критерии оценки состояния оператора.

Работа выполнена при частичной финансовой поддержке РФФИ, проекты №№ 14-07-00463-а, 15-07-06713-а и 16-07-00896-а.

[1] *Дорофеюк А. А., Гучук В. В., Десова А. А., Дорофеюк Ю. А.* Оценка работоспособности человека-оператора по информации из пульсового сигнала лучевой артерии // Управление в интеллектуальных, эргатических и организационных системах. — Ростов-на-Дону: Изд-во ЮФУ, 2013. Т. 3. С. 173–179.

Mining methods of quasi-periodic biosignals analysis in the assessment tasks of the human operator

Pokrovskaya Irina* ivp750@mail.ru
Guchuk Vladimir polma@bk.ru
Desova Aelita adesova@mail.ru
Dorofeyuk Alexander daa2@mail.ru
Moscow, Russia, ICS RAS

The paper is devoted to assessment of human operator psychophysiological status, which in significant part is determined by the degree of his fatigue. The methodological approach to the solution of this problem was developed, which includes three main stages: (*i*) formation of the criterion of the studied operator state for a specific field of his activity; (*ii*) definition of an informative indicators set that can adequately characterize the operator state; and (*iii*) the decision rules construction to determine the class (type) of the operator situation. The implementation of this approach takes place in the direction of an operator radial artery pulse signal (PS) maximum use [1]. For issue (*i*), the methodology for conducting a multivariate examination among experts in this field was developed; for issue (*ii*), the authors use the existing backlog to identify informative indicators of the PS in medical diagnostics solving problems; and for issue (*iii*), the authors use the modern schemes and procedures of pattern recognition. The methods of functionally significant elements of the PS isolation as well as algorithms for spectral analysis of the time series PS oscillatory components are described.

Currently, the data array of the work of the human operator in the experimental conditions with variable load and complexity are collected. On the basis of this array, the criteria for assessing the state of the operator, previously obtained by expertise, will be specified.

This research is executed at partial financial support of the Russian Foundation for Basic Research, grants Nos. 14-07-00463-a, 15-07-06713-a, and 16-07-00896-a.

[1] Dorofeyuk, A. A., V. V. Guchuk, A. A. Desova, and Y. A. Dorofeyuk. 2013. Assessment of the human operator productivity according to the radial artery pulse signal. *Management in an intelligent, ergonomic and organizational systems*. Rostov-on-Don: SFU Publs. 3:173–179.

Парциальные спектры спонтанной активности головного мозга человека

Рыкунов Станислав Дмитриевич[1]⋆ stanislavrykunov@gmail.com
Бойко Анна Ивановна[1] a.boyko@list.ru
Сычев Вячеслав Викторович[1] sychyov@yahoo.com
Устинин Михаил Николаевич[1] ustinin@impb.ru
Рыкунова Елена Дмитриевна[2] alenarykunova@gmail.com
[1]Пущино, Россия, ИМБП РАН
[2]Долгопрудный, Россия, МФТИ

Создан новый метод для вычисления спектральных характеристик различных отделов головного мозга. Этот метод объединяет два типа пространственных данных: (1) функциональную томограмму, представляющую собой трехмерное распределение мощности электрических источников, и (2) анатомическую структуру мозга, представленную магнитно-резонансной томограммой. В данной работе функциональная томограмма рассчитывается по данным многоканальной магнитной энцефалографии. В функциональной томограмме каждой элементарной осцилляции сопоставлено ее пространственное положение на дискретной сетке. Пространственная структура отдела мозга определяется с помощью аннотированной сегментации магнитно-резонансной томограммы. Парциальный спектр отдела головного мозга формируется из частот, локализованных в этом отделе. Разработано программное обеспечение, реализующее этот метод. Выполнен анализ парциальных спектров альфа-ритма [1].

Работа выполнена при поддержке РФФИ (проекты №№ 16-07-00937, 16-07-01000 и 14-07-00636) и Программы фундаментальных исследований Президиума РАН № I.33П.

[1] *Рыкунов С. Д., Устинин М. Н., Полянин А. Г., Сычев В. В., Линас Р. Р.* Комплекс программ для расчета парциальных спектров головного мозга человека // Математическая биология и биоинформатика, 2016 (в печати). Т. 11. № 1.

Partial spectroscopy of the human brain spontaneous activity

Rykunov Stanislav[1]⋆ stanislavrykunov@gmail.com
Boyko Anna[1] a.boyko@list.ru
Sychev Vyacheslav[1] sychyov@yahoo.com
Ustinin Mikhail[1] ustinin@impb.ru
Rykunova Elena[2] alenarykunova@gmail.com
[1]Russia, Pushchino, IMPB RAS
[2]Russia, Dolgoprudniy, MIPT

The new methodology was developed to calculate the spectral characteristics of various compartments of the human brain. This technology combines two types of the spatial data: (*i*) functional tomogram presenting spatial distribution of the electric sources; and (*ii*) anatomical structure of the brain as determined by the magnetic resonance imaging. Presently, the functional tomogram is calculated from the multichannel magnetoencephalograms. In the functional tomogram, unique spatial location corresponds to each elementary oscillation. Spatial structure of the brain compartment is generated by the segmentation of magnetic resonance image. The partial spectrum is composed by the selection of frequencies belonging to this compartment. The software implementing this methodology was developed and applied to partial spectral analysis of the alpha rhythm [1].

The research was partly supported by the Russian Foundation for Basic Research (grants Nos. 16-07-00937, 16-07-01000, and 14-07-00636) and by the Program of the Presidium of RAS No. I.33P.

[1] Rykunov, S. D., M. N. Ustinin, A. G. Polyanin, V. V. Sychev, and R. R. Lin'as. 2016 (in press.) Software for the partial spectroscopy of the human brain. *Math. Biol. Bioinformatics* 11(1).

Статистически значимое уменьшение количества бета-всплесков у пациентов на ранней стадии болезни Паркинсона

Сушкова Ольга Сергеевна[1]⋆ o.sushkova@mail.ru
Морозов Алексей Александрович[1,2] morozov@cplire.ru
Габова Александра Васильевна[3] agabova@yandex.ru
Бугаёв Александр Степанович[1] bugaev@cos.ru

[1]Москва, Россия, ИРЭ им. В.А. Котельникова РАН
[2]Москва, Россия, ФГБОУ ВО МГППУ
[3]Москва, Россия, ИВНДиНФ РАН

Разработан метод количественной оценки всплескообразной электрической активности мозга на основе вейвлет-анализа и непараметрической статистики. При помощи разработанного метода была обнаружена новая нейрофизиологическая закономерность у пациентов на ранних стадиях болезни Паркинсона. Сравнение разработанного метода с другими методами обработки электроэнцефалограмм (ЭЭГ), а именно: с вейвлет-спектрограммами на основе комплексного вейвлета Морле и стандартными спектрами Фурье, показало, что эти методы являются взаимодополняющими методами анализа. В частности, с помощью стандартных методов анализа было обнаружено статистически значимое увеличение мощности у пациентов с болезнью Паркинсона в отведениях C3 и C4 в диапазоне альфа, а с помощью метода оценки всплескообразной электрической активности мозга было обнаружено, что количество всплесков в частотном диапазоне бета в этих отведениях у пациентов статистически значимо уменьшено по сравнению с контрольной группой [1].

Работа поддержана грантами РФФИ №№ 16-37-00426 и 15-07-07846.

[1] *Сушкова О. С., Морозов А. А., Габова А. В.* Статистически значимое уменьшение количества бета-вспышек у пациентов на ранней стадии болезни Паркинсона // Нелинейный мир, 2016. Т. 14. №1. С. 59–60. http://www.radiotec.ru/catalog.php?cat=jr11&art=17554.

A statistically significant decrease of the quantity of beta wave packets in de novo Parkinson's disease

Sushkova Olga[1]⋆ o.sushkova@mail.ru
Morozov Alexei[1,2] morozov@cplire.ru
Gabova Alexandra[3] agabova@yandex.ru
Bugaev Alexandr[1] bugaev@cos.ru

[1]Russia, Moscow, Kotel'nikov IRE RAS
[2]Russia, Moscow, MSUPE
[3]Russia, Moscow, IHNA and NPh RAS

A method of analysis of electroencephalogram (EEG) wave packets based on wavelets and nonparametric statistics is developed. The method reveals a new statistically significant difference between a group of de novo Parkinson's disease patients and a control group. The method is compared with standard methods based on Fourier spectra and complex Morlet wavelets by the example of Parkinson's disease experimental data. The authors demonstrate that these methods are complementary, that is, the standard methods and the wave packets analysis method reveal sufficiently different effects in the EEG data. It means that the standard methods indicate a significant increase of the power spectral density in the Parkinson's disease in the C3 and C4 electrodes in the alpha frequency band and the method of wave packets analysis indicates a significant decrease of the quantity of the wave packets in these electrodes in the nearby beta frequency range [1].

This research is funded by the Russian Foundation for Basic Research, grants 16–37–00426 and 15–07–07846.

[1] Sushkova, O., A. Morozov, and A. Gabova. 2016. Statistically significant decrease of the quantity of beta peaks in de novo Parkinson's disease. *Nonlinear World* 14(1):59–60. Available at: http://www.radiotec.ru/catalog.php?cat=jr11&art=17554 (accessed July 26, 2016).

Анализ данных в пространстве «частота-паттерн» для оценки функциональной структуры тела человека по внешнему магнитному полю

Устинин Михаил Николаевич* ustinin@impb.ru
Рыкунов Станислав Дмитриевич stanislavrykunov@gmail.com
Бойко Анна Ивановна a.boyko@list.ru
Сычев Вячеслав Викторович sychyov@yahoo.com

Пущино, Россия, ИМПБ РАН

Предложен новый метод для реконструкции электрической активности человеческого тела по многоканальным записям внешнего магнитного поля. Преобразование Фурье полного временно́го ряда приводит к детальному спектру, содержащему десятки тысяч частотных компонент. Для каждой частоты восстанавливается многоканальный временной ряд и проводится оценка когерентности. Если когерентность ниже некоторого порога, методом анализа независимых компонент выделяются когерентные составляющие, каждая из которых имеет постоянный паттерн магнитного поля. Локализация всех источников по паттернам приводит к построению функциональной томограммы — распределения энергии спектра в пространстве. Метод был верифицирован на симулированных данных и на физическом фантоме, после чего использовался для восстановления функциональной структуры мозга, сердца и скелетных мышц. Полученные результаты разумно интерпретируются анатомически [1].

Работа выполнена при поддержке РФФИ (проекты №№ 16-07-00937, 16-07-01000 и 14-07-00636) и Программы фундаментальных исследований Президиума РАН № I.33П.

[1] *Llinás R. R., Ustinin M. N., Rykunov S. D., Boyko A. I., Sychev V. V., Walton K. D., Rabello G. M., Garcia J.* Reconstruction of human brain spontaneous activity based on frequency-pattern analysis of magnetoencephalography data // Front. Neurosci., 2015. Vol. 9. P. 373. http://journal.frontiersin.org/article/10.3389/fnins.2015.00373/full.

Frequency-pattern data analysis to estimate the functional structure of the human body from external magnetic field

Ustinin Mikhail* ustinin@impb.ru
Rykunov Stanislav stanislavrykunov@gmail.com
Boyko Anna a.boyko@list.ru
Sychev Vyacheslav sychyov@yahoo.com
Russia, Pushchino, IMPB RAS

A new method was proposed to reconstruct the electric activity of the human body from multichannel recordings of external magnetic field. The Fourier transform of the long time series produces detailed spectrum, containing tens of thousands of frequency components. For each frequency, the multichannel time series are reconstructed and the coherence is estimated. If this coherence is lower than some threshold, then coherent components are extracted by the independent component analysis. Each coherent component manifests the constant magnetic field pattern. Localization of all sources from those patterns gives the functional tomogram — spatial distribution of spectral energy. The method was verified on simulated data and on physical phantom. Then, it was used to reconstruct the functional structure of the brain, heart, and skeletal muscles. The results obtained can be reasonably interpreted anatomically, leading to the conclusion about good prospects of the proposed method in diagnostics [1].

The research was partly supported by the Russian Foundation for Basic Research (grants Nos. 16-07-00937, 16-07-01000, and 14-07-00636) and by the Program of the Presidium of RAS No. I.33P.

[1] Llinás, R. R., M. N. Ustinin, S. D. Rykunov, A. I. Boyko, V. V. Sychev, K. D. Walton, G. M. Rabello, and J. Garcia. 2015. Reconstruction of human brain spontaneous activity based on frequency-pattern analysis of magnetoencephalography data. *Front. Neurosci.* 9:373. Available at: `http://journal.frontiersin.org/article/10.3389/fnins.2015.00373/full` (accessed July 26, 2016).

Внутренние расстояния спиральных пар в белковых молекулах

Куликова Людмила Ивановна[1]⋆ likulikova@mail.ru
Тихонов Дмитрий Анатольевич[1] dmitry.tikhonov@gmail.com
Ефимов Александр Васильевич[2] efimov@protres.ru
[1]Россия, Пущино, ИМПБ РАН — филиал ИПМ им. М. В. Келдыша РАН
[2]Россия, Пущино, ИБ РАН

Проведен статистический анализ распределения межспиральных расстояний в парах связанных между собой перетяжками спиралей в пространственных структурах белковых молекул. Полученное по определенным правилам множество спиральных пар всех белковых молекул, включенных в Protein Data Bank, разбито на три подмножества по критерию пересечения проекций спиралей на параллельные плоскости, проходящие через оси спиралей. Показано, что распределения расстояний для спиральных пар, проекции которых не имеют пересечений, имеют более дальнодействующий характер, чем те, проекции которых пересекаются. При помощи регрессионного анализа исследован характер распределений, в частности показано, что в подмножестве без пересечений распределения различных расстояний между осями спиралей относятся к гамма-распределениям. Показано, что подмножества пар с пересечением имеют с большой вероятностью малое отношение минимального расстояния к межплоскостному, в отличие от подмножества пар без пересечения, где наблюдается противоположная картина. Обосновывается вывод о том, что спиральные пары, проекции спиралей которых пересекаются, дополнительно стабилизируются за счет внутренних взаимодействий [1].

Работа выполнена при поддержке грантов РФФИ №№ 16-01-00692, 14-07-00924 и 15-29-07063.

[1] *Тихонов Д. А., Куликова Л. И., Ефимов А. В.* Статистический анализ внутренних расстояний спиральных пар в белковых молекулах // Математическая биология и биоинформатика, 2016. Т. 11. № 2. С. 170–190. doi: 10.17537/2016.11.170.

Internal distances of helical pairs in protein molecules

Kulikova Liudmila[1]⋆ likulikova@mail.ru
Tikhonov Dmitrii[1] dmitry.tikhonov@gmail.com
Efimov Aleksandr[2] efimov@protres.ru
[1]Pushchino, Russia, IMPB RAS — Branch of KIAM RAS
[2]Pushchino, Russia, IPR RAS

A statistical analysis of interhelical distances in pairs of connected helices found in known proteins has been performed. In accordance with the certain rules, a database of the pairs found in the Protein Data Bank has been compiled. This set was subdivided into three subsets according to criterion of crossing helix projections on the parallel plane passing through the axis of the helix. It was shown that the distribution of distances between the pairs of helices whose projections are not crossed has a more long-range nature than those whose projections are overlapped. Using the regression analysis, the nature of distributions is investigated. In particular, it is shown that the distributions of interhelical distances in the subset of pairs of helices without intersections belong to the gamma distributions. It is also shown that the subset of the pairs with crossing projections have a smaller ratio of the minimal distance between the helical axes to the interplanar distance that is contrast to the set without crossing projections. It was concluded that the helical pairs with crossing projections are additionally stabilized by internal interactions [1].

This work was supported by the Russian Foundation for Basic Research, grants Nos. 16-01-00692, 14-07-00924, and 15-29-07063.

[1] Tikhonov, D., L. Kulikova, and A. Efimov. 2016. Statistical analysis of the internal distances of helical pairs in protein molecules. *Math. Biol. Bioinform.* 11(2):170–190. doi: 10.17537/2016.11.170.

Построение метрик на множестве биомолекулярных последовательностей

Сулимова Валентина Вячеславовна[1]⋆ vsulimova@yandex.ru
Середин Олег Сергеевич[1] oseredin@yandex.ru
Моттль Вадим Вячеславович[2] vmottl@yandex.ru
[1]Россия, Тула, ТулГУ
[2]Россия, Москва, ФИЦ ИУ РАН

Для анализа биомолекулярных последовательностей необходимо уметь сравнивать их между собой. С точки зрения передовых методов анализа данных наиболее предпочтительным способом сравнения являются меры несходства, обладающие свойствами метрики. С точки зрения молекулярной биологии важно, чтобы способ сравнения учитывал биологические особенности объектов сравнения. Кроме того, важна вычислительная эффективность и возможность использовать в дальнейшем удобные инструменты анализа данных. Ни один из известных способов сравнения биомолекулярных последовательностей не обладает всеми требуемыми свойствами.

В данной работе предлагается простой способ построения метрик на множестве биомолекулярных последовательностей. Он, как и традиционные способы сравнения биомолекулярных последовательностей, основывается на поиске их оптимального парного выравнивания и модели мутационных замен аминокислот в ходе эволюции.

В данной работе приводится доказательство того, что предложенные меры несходства обладают свойствами метрики, что позволяет использовать их в передовых методах анализа данных, сохраняющих вычислительные достоинства SVM (support vector machine), но не требующих введения признаков последовательностей и(или) скалярного произведения. Результаты экспериментов подтверждают адекватность предложенных метрик прикладным задачам на примере классификации мембранных гликопротеинов [1].

Работа поддержана грантом РФФИ № 15-07-08967.

[1] *Сулимова В. В., Середин В. В., Моттль В. В.* Метрики на основе оптимального выравнивания биомолекулярных последовательностей // Машинное обучение и анализ данных, 2016 (в печати).

Construction metrics for biomolecular sequences

Sulimova Valentina[1]⋆ vsulimova@yandex.ru
Seredin Oleg[1] oseredin@yandex.ru
Mottl Vadim[2] vmottl@yandex.ru
[1]Tula, Russia, TulSU
[2]Moscow, Russia, FRC CSC RAS

For biomolecular sequences analysis, it is important to have an appropriate way for comparing them. From the point of view of the advanced methods of data analysis, the most preferred way for comparing objects is a dissimilarity measure, possessing metric's properties. From the other side, from the point of view of the molecular biology, it is important to take into account biological features of the compared objects. Besides, the computational effectiveness and the possibility of futher using convinient instruments of data analysis are also important.

There are a number of ways for comparing biomolecular sequences, though no one of them possess the all required properties.

This paper proposes a simple way for computing metrics for biomolecular sequences.

The proposed approach, following the traditional ways for biomolecular sequences comparing, is based on finding an optimal pairwise alignment and on the model of mutual changes of amino acids at the process of evolution.

In this paper, it has been proven that the proposed dissimilarity measure is a metric. So, it can be used at the advanced methods of data analysis, saving computational advantages of SVM (support vector machine) without introducing features of objects or(and) an inner product. The experimental results confirm usability of the proposed metric for membrane glicoprotein classification [1].

This research is funded by the Russian Foundation for Basic Research, grant 15-07-08967.

[1] Sulimova, V., O. Seredin, and V. Mottl. 2016 (in press). Metrics on the basis of optimal alignment of biomolecular sequences. *Machine Learning Data Anal.*

Автоматизированная технология обнаружения участков латентной периодичности в последовательностях ДНК

Чалей Мария Борисовна[1] maramaria@yandex.ru
Кутыркин Владимир Андреевич[2] vkutyrkin@yandex.ru
Тюльбашева Гаянэ Эдуардовна[1] gayat@multiline.ru
Теплухина Елена Ивановна[1] elena.teplo@yandex.ru
Назипова Нафиса Наиловна[1]⋆ nnn@impb.psn.ru
[1]Пущино, Россия, ИМПБ РАН
[2]Москва, Россия, МГТУ им. Н. Э. Баумана

Латентная профильная периодичность обощает понятие неточного тандемного повтора (ТП). Точный ТП — это текстовая строка, представляющая собой чередование копий подстроки, которая называется паттерном периодичности. В неточных ТП в копиях паттерна допускается наличие вставок, выпадений и замен символов. Когда речь идет о латентной профильной периодичности, искажения в каждой позиции копий паттерна описываются некоторым случайным распределением [1]. Для обнаружения районов латентной периодичности в нуклеотидных последовательностях был разработан спектрально-статистический подход, достоверно детектирующий наличие гетерогенности в копиях паттерна. Подход использует χ^2-статистику для проверки гипотезы о гомогенности ДНК на уровне значимости, характерном для неточных ТП, которые являются гетерогенными последовательностями. Однако наличие значимой гетерогенности является необходимым, но не достаточным условием латентной периодичности. Проблема надежности в поиске приближенных ТП при условии небольшого размера выборки решается с помощью стохастической модели проявления гетерогенности. Достоверные и неизбыточные данные о скрытых ТП в ряде геномов были получены в результате применения автоматической вычислительной процедуры и помещены в БД HeteroGenome (`http://www.jcbi.ru/lp_baze`).

Работа поддержана грантом РФФИ № 15-07-05783.

[1] *Chaley M., Kutyrkin V., Tulbasheva G., Teplukhina E., Nazipova N.* HeteroGenome: Database of genome periodicity // Database: The Journal of Biological Databases and Curation, 2014. Vol. 2014. P. 1–18. doi: `http://dx.doi.org/10.1093/database/bau040`.

Automated technology for revealing latent periodicities in DNA sequences

Chaley Mariya[1] maramaria@yandex.ru
Kutyrkin Vladimir[2] vkutyrkin@yandex.ru
Tulbasheva Gayane[1] gayat@multiline.ru
Teplukhina Elena[1] elena.teplo@yandex.ru
Nazipova Nafisa[1]⋆ nnn@impb.psn.ru
[1]Russia, Pushchino, IMPB RAS
[2]Russia, Moscow, BMSTU

Notion of latent profile periodicity expands on the notion of approximate tandem repeat (TR). Perfect TR is a textual string which consists of sequential copies of its substring called a periodicity pattern. In approximate TR, a number of characters in pattern copies are distorted by insertion/deletions and point mutations. If latent profile periodicity exists in DNA sequence, distortion of the characters in each position of pattern copies occurs in accordance with the corresponding probability distribution [1]. The spectral-statistical approach was developed which reveals latent periodicity by detecting significant heterogeneity at the pattern copies of an analyzed nucleotide sequence. The approach employs χ^2-statistics for testing homogeneity in DNA sequence at significance level characteristic for approximate TRs which sequences are obviously the heterogenic ones. However, a significant heterogeneity is a necessary, but insufficient condition for determining latent periodicity. The authors searched for latent periodicity by identifying significant heterogeneities in overlapping windows of various sizes in analysis of multiple-scanned DNA sequences with variable steps. The problem of reliability in locating approximate TRs under the condition of small sample size is resolved using the stochastic model of heterogeneity manifestation. This model allows employing the additional statistical tests to check the hypothesis of heterogeneity presence in DNA sequence. Reliable approximate TRs were revealed and collected in the HeteroGenome database (`http://www.jcbi.ru/lp_baze`).

The research is funded by the RFBR, grant No. 15-07-05783.

[1] Chaley, M., V. Kutyrkin, G. Tulbasheva, E. Teplukhina, and N. Nazipova. 2014. HeteroGenome: Database of genome periodicity. *Database: The Journal of Biological Databases and Curation* 2014:1–18. doi: `http://dx.doi.org/10.1093/database/bau040`.

Сопоставление и интеграция подходов к дешифровке древнерусских знаменных песнопений

Бахмутова Ирина Владимировна bakh@math.nsc.ru
Гусев Владимир Дмитриевич gusev@math.nsc.ru
Мирошниченко Любовь Александровна* luba@math.nsc.ru
Титкова Татьяна Николаевна titkova@math.nsc.ru
Россия, Новосибирск, ИМ СО РАН

Проблема нотолинейной реконструкции (дешифровки) древнерусских церковных песнопений в знаменной форме записи является одной из наиболее актуальных в музыкальной медиевистике. Рассматриваются наиболее перспективные для решения указанной проблемы компьютерно-ориентированные подходы. Предпочтение отдается предложенному авторами подходу, основанному на использовании внутригласовых инвариантов (ВИ) — цепочек знамен, характеризующихся минимальным уровнем неоднозначности. Словари инвариантов формируются на рукописных двознаменниках конца XVII – начала XVIII вв., являющихся своего рода билингвами знаменного распева. В них песнопения представлены параллельно в знаменной и нотолинейной форме записи и синхронизованы со стихотворным текстом. Сделан вывод о целесообразности формирования словарей ВИ и КВИ (внутригласовых квазиинвариантов) по отдельным типам певческих книг (Октоихи, Праздники, Ирмологии и т. п.) и необходимости согласования типов дешифруемого песнопения и используемого словаря. Это повышает точность дешифровки [1].

Получены оценки эффективности разных подходов на контрольном материале, не задействованном при обучении. Рассмотрена возможность интеграции разных подходов, что позволяет обеспечить в среднем 65–80-процентную дешифруемость по разным гласам (классам песнопений).

Работа поддержана грантом РФФИ № 16-07-00812.

[1] *Бахмутова И. В., Гусев В. Д. Мирошниченко Л. А., Титкова Т. Н.* Сопоставление и интеграция подходов к дешифровке древнерусских знаменных песнопений // Машинное обучение и анализ данных, 2016 (в печати).

Comparison and integration of approaches to deciphering of ancient Russian hymnals

Bakhmutova Irina bakh@math.nsc.ru
Gusev Vladimir gusev@math.nsc.ru
Miroshnichenko Lubov ⋆ luba@math.nsc.ru
Titkova Tatyana Titkova@math.nsc.ru
Novosibirsk, Russia, IM SB RAS

The problem of noted reconstruction (deciphering) of ancient Russian hymnals presented in neumatic (znamenny) form is the most vital in musical medieval history. The most promising computer-oriented approaches to the solution of this problem are considered. Preference is given to the approach, suggested by the present authors, that is based on use of echos-specific invariants — the neume chains characterized by minimal level of ambiguity. The dictionaries of invariants are formed on the base of dvoeyznamenniks of the XVII–XVIII centuries that are some kind of bilinguas of znamenny chant. Here, the hymnals are presented in parallel: in neume, in note, and synchronized with verses. The conclusion has been done that it is very reasonable to form the dictionaries of invariants and quasi-invariants by separate types of cantorum manuscripts (Octoechos, Holidays, Hirmologion, etc.) and it is necessary to fit the type of deciphering hymnal to the type of the used dictionary. This increases the faithfulness of deciphering [1]. The performance evaluations of various approaches are obtained on checking material not related to learning. The possibility of integration of various approaches was considered that provided the deciphering 65%–80% (on average) for different ichoses (classes of hymnals).

This research is funded by the Russian Foundation for Basic Research, grant 16-07-00812.

[1] Bakhmutova, I., V. Gusev, L. Miroshnichenko, and T. Titkova. 2016 (in press). Comparison and integration of approaches to deciphering of ancient Russian hymnals. *Machine Learning Data Anal.*

Презентация программного средства для встраивания и извлечения скрытых сообщений в аудиофайлах

Жарких Александр Александрович[1]⋆ zharkikh090107@mail.ru
Горбунов Алексей Владимирович[2] lergex@gmail.com
[1]Россия, Мурманск, ФГБОУ ВПО «МГТУ»
[2]Россия, Мурманск, МФ ФГБУ ЦСМС

Цель данного сообщения — презентация стеганографического метода, алгоритмов его реализации и программного средства для встраивания и извлечения. Метод базируется на модификации вектора отсчетов аудиосигнала, представленного в формате импульсно-кодовой модуляции. Сообщение представляет собой вектор отсчетов сигнала, полученного путем двоичной частотной манипуляции из вектора бит сообщения.

На начальном этапе аудиофайл-контейнер специальным образом подготавливается к встраиванию. Контейнер разбивается на отрезки. На каждом отрезке осуществляется режекторная фильтрация. Скрываемое сообщение представляется в виде бинарной последовательности. Далее сообщение модулируется методом частотной манипуляции. Затем полученные отсчеты модулированного сообщения суммируются с отсчетами контейнера. В итоге на выходе формируется стегоконтейнер.

Для извлечения сообщения предлагается использовать следующий способ. Аудиофайл-стегоконтейнер также разбивается на отрезки. На каждом отрезке осуществляется полосовая фильтрация. Таким образом, выделяется модулированное сообщение. Для получения скрытого сообщения используется корреляционный метод демодуляции.

На основе изложенных алгоритмов на языке C# было создано программное средство, позволяющее осуществлять встраивание скрытых сообщений в аудиофайлы, а также их извлечение и прочтение [1].

[1] *Жарких А. А., Горбунов А. В.* Реализация пакета программ для встраивания и извлечения скрытых сообщений в аудиофайлах // Машинное обучение и анализ данных, 2016 (в печати).

Presentation of software tool for embedding and extracting of the hidden messages in audio files

Zharkikh Aleksandr[1]⋆ zharkikh090107@mail.ru
Gorbunov Aleksei[2] lergex@gmail.com
[1]Murmansk, Russia, FSBEI HE "MSTU"
[2]Murmansk, Russia, MB FGFI CFMC

The purpose of this contribution is the presentation of the steganographic method, its implementation algorithms, and software tool for embedding and extracting. The method is based on the modification vector of audio samples represented by a pulse-code modulated format. Message is a vector of signal samples obtained by a binary frequency shift keying from bit vector message.

At the initial stage, an audio file container is specially prepared for embedding. The container is divided into segments. The container is divided into segments. Each segment uses a notch filter. Concealed message is represented as a binary sequence. The duration of embedding one bit of the message is assessed. Then, a message is modulated by frequency shift keying. Then, the received reports of the modulated messages are added to the samples of the container. As a result, at the output, stego container is generated.

To retrieve the messages from the container, the following method is proposed to use. Stego audio file is also divided into segments. Each segment uses a bandpass filter. Thus, the modulated message is allocated. Further, the duration of the transmission of one bit of the message is assessed. The modulated message is divided into segments of a single bit transmission, within each of which the frequency is assessed. At this frequency, a decision is made on which bit is transmitted. For this purpose, the correlation demodulation method is used.

On the basis of the described algorithms, a software tool has been created in the language C#, which allows one to embed the hidden messages in audio files as well as the extraction and interpretation.

[1] Zharkikh, A., and A. Gorbunov. 2016 (in press). Implementation of the software package for embedding and extracting hidden messages in audio files. *Machine Learning Data Anal.*

Выбор решений при распознавании эмоций по речи

Кальян Виктор Петрович[1,2] vkalyan@mail.ru
Кальян Анастасия Викторовна[2]⋆ nastya-kalyan@yandex.ru
[1]Россия, Москва, ФИЦ ИУ РАН
[2]Россия, Москва, РУДН

Описан опыт выбора решений в системе распознавания эмоционального состояния человека по речи в отношении правдивости и искренности говорящего. Анализируется информативность измерительной базы распознавания на основании паралингвистических, артикуляционных и экстралингвистических особенностей речи с учетом индивидуальных эмоционально-смысловых коннотаций испытуемого, описываются алгоритмы распознавания эмоций по речи, осуществляется выбор из множества решений и их верификация в отношении искренности и правдивости говорящего с учетом ситуативного контекста.

Анализ эмоциональных проявлений с их эмоционально-смысловыми коннотациями без соотнесения с реконструируемой картиной происшествия и ситуацией дознания показывает неоднозначность этих коннотаций, что может привести к существенным ошибкам распознавания.

Предложена стратегии выбора решений распознавания эмоционального состояния человека по речи в связанной системе темпорально-акустических, эмоционально-смысловых и ситуационных зависимостей. При настоящем подходе верификация смыслов эмоциональных речевых реакций становится возможной благодаря именно сопоставлению с ситуативным контекстом [1].

Работа поддержана грантом РФФИ № 11-01-00900.

[1] *Кальян В. П.* Выбор решений при распознавании эмоций по речи // Машинное обучение и анализ данных, 2016 (в печати).

The choice of decisions at recognition of emotions on the speech

Kalyan Viktor[1,2] vkalyan@mail.ru
Kalyan Anastasia[2]⋆ nastya-kalyan@yandex.ru
[1]Moscow, Russia, FRC CSC RAS
[2]Moscow, Russia, PFUR

An experience of the choice of decisions in system of recognition of an emotional condition of the person on the speech concerning truthfulness and sincerity of speaking is described. Informational content of measuring base of recognition on the basis of paralinguistic, articulation, and extralinguistic features of the speech taking into account individual emotional and semantic connotations of the examinee is analyzed, algorithms of recognition of emotions according to the speech are described, and the choice from a set of decisions and their verification concerning sincerity and truthfulness speaking taking into account a situational context is carried out.

The analysis of emotional manifestations with their emotional and semantic connotations without correlation with the reconstructed picture of incident and a situation of inquiry shows ambiguity of these connotations that can lead to essential errors of recognition.

The strategy of the choice of solutions of recognition of an emotional condition of the person on the speech in the related system of temporal and acoustic, emotional and semantic, and situational dependences is suggested. At the present approach, verification of meanings of emotional speech reactions becomes possible thanks to comparison to a situational context [1].

The work is supported by the Russian Foundation for Basic Research, grant No. 11-01-00900.

[1] Kalyan, V. 2016 (in press). The choice of decisions at recognition. *Machine Learning Data Anal.*

Неявная модель вариативности произношения

Чучупал Владимир Яковлевич v.chuchupal@gmail.com
Россия, Москва, ФИЦ ИУ РАН

Вариативность произношения слов и словосочетаний в естественной разговорной речи является одним из основных источников ошибок при автоматическом распознавании речи. Поэтому использование моделей вариативности произношения представляется важным направлением повышения эффективности работы систем распознавания речи. Рассматривается проблема моделирования вариативности произношения, которая вызвана нечеткой, неполной артикуляцией, например в результате нарушения синхронизации работы органов речеобразования. Предлагается использовать неявные произносительные модели, основанные на комбинировании акустических моделей соседних звуков. Комбинирование может заключаться в сглаживании или интерполяции параметров акустических моделей текущих звуков параметрами соседних моделей. Степень проявления вариативности зависит от синтаксического и просодического контекста звука, поэтому предлагается выбирать значения параметров интерполяции в зависимости от наличия позиционных, фонетических, синтаксических и просодических признаков. Подход с комбинированием акустических моделей на основе сглаживания их параметров был описан в научной литературе, однако автору неизвестны исследования с комбинированием соседних акустических моделей, а также где бы параметры комбинирования зависели от текущих контекстных признаков. Численные эксперименты на корпусах с читаемой и разговорной речью показали справедливость предположений о целесообразности использования интерполяционных моделей и существования зависимости параметров сглаживания моделей от наличия позиционных и синтаксических признаков.

Работа поддержана грантом РФФИ № 14-01-00607.

[1] *Чучупал В. Я.* Неявная модель вариативности произношения // Машинное обучение и анализ данных, 2016 (в печати).

Implicit pronunciation variation model

Chuchupal Vladimir v.chuchupal@gmail.com
Moscow, Russia, FRC CSC RAS

Variations in pronunciation of words in natural speech are considered as one of the main sources of speech recognition errors. This is the reason for development and implementation of the advanced pronunciation models in modern automatic speech recognition systems. The paper considers the pronunciation variations caused by a fuzzy or incomplete articulation that is frequently observed in spontaneous speech. The implicit pronunciation model is proposed that is implemented in the form of a combination of the acoustical models of the adjacent phones. Such a model can be realized by smoothing or interpolation of the corresponding model parameters. Also, it is proposed to use the context-dependent interpolation, so that the values of the smoothing parameters are conditioned by the current position, syntax, and prosodic contexts of the phoneme. The pronunciation modeling approach on the base of combination of acoustical models (including the interpolation) has already been discussed in literature; however, the method based on the combination of the adjacent models with the use of the context-dependent smoothing parameters has not already been published as far as the author knows. The numerical experiments on the databases that contained both a readable and spontaneous speech showed that the implementation of the described acoustic model combination with the variable smoothing parameters could bring the gain in performance of a recognition system [1].

This research is funded by the Russian Foundation for Basic Research, grant 14-01-00607.

[1] Chuchupal, V. 2016 (in press). Implicit pronunciation variation model. *Machine Learning Data Anal.*

Аддитивная регуляризация тематических моделей для поиска этничного дискурса в социальных медиа

Апишев Мурат Азаматович[2] great-mel@yandex.ru
Кольцов Сергей Николаевич[1] skoltsov@hse.ru
Кольцова Елена Юрьевна[1] ekoltsova@hse.ru
Николенко Сергей Игоревич[1,3] sergey@logic.pdmi.ras.ru
Воронцов Константин Вячеславович[4]⋆ vokov@forecsys.ru
[1]Россия, Санкт-Петербург, НИУ ВШЭ
[2]Россия, Москва, МГУ
[3]Россия, Санкт-Петербург, МИАН РАН
[4]Россия, Москва, МФТИ

В последнее время технологии анализа больших текстовых коллекций все чаще используются для социологических исследований в Интернете. Одна из таких задач — выявление тематики этничного дискурса для раннего обнаружения межнациональных конфликтов в социальных медиа. Чаще всего для этого используются байесовские вероятностные тематические модели на основе *латентного размещения Дирихле* (LDA). Однако *аддитивная регуляризация тематических моделей* (ARTM) предоставляет больше гибкости для комбинирования критериев и источников данных, управления качеством и интерпретируемостью тем.

В данной работе используется библиотека с открытым кодом BigARTM и определяется проблемно-ориентированная комбинация регуляризаторов, чтобы найти как можно больше этно-релевантных тем. Наиболее важным является регуляризатор, использующий словарь из нескольких сотен этнонимов, вокруг которых концентрируются этно-релевантные темы. Эксперименты показывают, что регуляризованные модели лучше подходят для выявления большого числа редких тем, поскольку находят больше тем лучшего качества.

Работа поддержана грантом Российского научного фонда 15-18-00091.

[1] *Apishev V., Koltcov S., Koltsova O., Nikolenko S., Vorontsov K.* Additive regularization for topic modeling in sociological studies of user-generated text content // 15th Mexican Conference (International) on Artificial Intelligence, 2016.

Additively regularized topic model for searching ethnical discourse in social media

Apishev Murat[2] great-mel@yandex.ru
Koltsov Sergei[1] skoltsov@hse.ru
Koltsova Olessia[1] ekoltsova@hse.ru
Nikolenko Sergey[1,3] sergey@logic.pdmi.ras.ru
Vorontsov Konstantin[4]⋆ vokov@forecsys.ru

[1]St. Petersburg, Russia, HSE
[2]Moscow, Russia, MSU
[3]St. Petersburg, Russia, Steklov Mathematical Institute of RAS
[4]Moscow, Russia, MIPT

Recently, social studies of the Internet have started to adopt various techniques of large-scale text mining for a variety of goals. One of such goals is the unsupervised discovery of topics related to ethnicity for early detection of ethnic conflicts emergent in social media. Probabilistic topic modeling used for such goals usually employs Bayesian inference for one of the numerous extensions of the *latent Dirichlet allocation* (LDA) model that has been widely popular over the last decade. However, recent research suggests that a non-Bayesian approach of *additive regularization of topic models* (ARTM) results in more control over the topics purity and interpretability, more flexibility for combining topic models, and faster inference. In this work, ARTM framework and BigARTM open-source software are applied to a case study of mining ethnic content from the Russian-language blogosphere. A problem-specific combination of regularizers for ARTM is introduced and ARTM is compared with LDA. The most important regularizer uses a vocabulary of a few hundred ethnonyms as seed words for ethnic-related topics. One may conclude that ARTM is better suitable for mining rare topics, such as those on ethnicity, since it obviously finds larger numbers of relevant topics of higher or comparable quality. This work was supported by the Russian Science Foundation grant no. 15-18-00091.

[1] Apishev, V., S. Koltcov, O. Koltsova, S. Nikolenko, and K. Vorontsov. 2016. Additive regularization for topic modeling in sociological studies of user-generated text content. *15th Mexican Conference (International) on Artificial Intelligence*.

Вероятностная модель для сглаживания целевых метрик качества ранжирования

Волков Никита Алексеевич[1,2]⋆ nikitavolkov@yandex-team.ru
Жуковский Максим Евгеньевич[1,2] zhukmax@yandex-team.ru
[1]Россия, Москва, МФТИ
[2]Россия, Москва, ООО «Яндекс»

Задача информационного поиска — находить релевантные документы по запросу пользователя. Одной из важнейших задач информационного поиска является задача ранжирования результатов поиска. В настоящее время широко распространен подход получения функции релевантности с помощью методов машинного обучения на основе обучающей выборки. Для оценки качества принято использовать метрики качества, например $pFound$. Однако большинство из них являются дискретными, что усложняет, например, отбор признаков.

Цель данной работы — получить гладкий аналог дискретной метрики. Рассматриваются несколько вариантов определения гладкой метрики, лучший из которых экспериментально определяется по критериям гладкости и похожести на дискретную метрику. Критерий похожести, в отличие от критерия гладкости, численно выразить довольно трудно ввиду большого множества различных случаев поведения метрик, поэтому от численного описания пришлось отказаться. По результатам многочисленных экспериментов была найдена оптимальная модель гладкой метрики в соответствии с предъявляемыми требованиями [1].

[1] *Волков Н. А., Жуковский М. Е.* Вероятностная модель для сглаживания целевых метрик качества ранжирования // Машинное обучение и анализ данных, 2016 (в печати).

On a probabilistic model for smoothing discrete ranking quality metrics

Volkov Nikita[1,2]⋆ nikitavolkov@yandex-team.ru
Zhukovskii Maksim[2] zhukmax@yandex-team.ru
[1]Moscow, Russia, MIPT
[2]Moscow, Russia, Yandex LLC

The information retrieval aim is to find relevant documents by given user's query. One of the main information retrieval task is the ranking problem of searching results. Currently, the method of getting relevant function with machine learning methods based on train sample is widely distributed. The quality of ranking is assessing with quality metrics, for example, pFound. However, most of them is discrete, it is difficult for feature selection.

The aim of this work is to get a smoothing analogue of the discrete metric. The present authors consider some different definitions of smoothing metric, the best of which has been found by the smoothing criteria and the criteria of a similarity to discrete metric by means of experiments. In comparence with smoothing criteria, it is difficult to obtain numerical description of the similarity criteria because there are a lot of different cases of metric behavior. So, it was decided not to use numerical description. As a result of the large number of experiments, an optimal model of a smoothing metric was got according to the necessary requirements.

[1] Volkov, N. A., and M. E. Zhukovskii. 2016 (in press). On a probabilistic model for smoothing discrete ranking quality metrics. *J. Machine Learning Data Anal.*

Построение иерархических тематических моделей коллекций коротких текстов

Кузьмин Арсентий Александрович[1]⋆ arsentii.kuzmin@gmail.com
Адуенко Александр Александрович[1] aduenko1@gmail.com
Стрижов Вадим Викторович[2] strijov@ccas.ru
[1]Россия, Москва, МФТИ
[2]Россия, Москва, ФИЦ ИУ РАН

Решается проблема построения иерархической тематической модели коллекции коротких текстов и верификация экспертной тематической модели [1]. Предлагаются алгоритмы выбора оптимальной метрики и отбора признаков. Анализируются способы представления документа в виде действительного вектора. Сравниваются агломеративные и дивизимные подходы при построении иерархической тематической модели.

Предлагается иерархическая взвешенная функция сходства для классификации неразмеченных документов коллекции, в которой весом каждого слова из словаря коллекции является его важность для кластеризации и классификации. Предлагается энтропийный метод оценки весов данной функции с помощью экспертной тематической модели. Предложенная функция сходства адаптируется для учета векторного представления слов с помощью языковых моделей и представляется в виде четырехслойной нейронной сети.

Предложенные методы используются при построении экспертной системы для классификации новых тезисов крупной конференции EURO с помощью экспертных тематических моделей данной конференции с 2006 по 2016 гг. Результаты сравниваются с иерархическим мультиклассовым SVM, вероятностной тематической моделью SuhiPLSA и иерархическим наивным байесовским подходом.

Работа поддержана грантом РФФИ № 16-07-01160.

[1] *Кузьмин А. А., Адуенко А. А., Стрижов А А.* Тематическая классификация тезисов крупной конференции с использованием экспертной модели // Информационные технологии, 2014. Т. 6. С. 22–26.

Hierarchical thematic modeling of short text collection

Kuzmin Arsentiy[1]⋆ arsentii.kuzmin@gmail.com
Aduenko Alexander[1] aduenko1@gmail.com
Strijov Vadim[2] strijov@ccas.ru
[1]Moscow, Russia, MIPT
[2]Moscow, Russia, FRC CSC RAS

The aim of this study is to construct and verify a hierarchical thematic model of a short text collection. The present authors consider the ways for metrics learning and features selection. Agglomerative and divisive methods to construct a hierarchical model are compared.

A hierarchical weighted similarity function is suggested for unlabeled data classification. Weights in this function are the importance values of the terms from the collection dictionary. Entropy-based approach is used to estimate these weights according to the expert model. The proposed similarity function is represented as four-level neural network to consider vector representation of the words given by a trained language model.

The proposed methods are used to construct an expert system that helps experts to classify unlabeled abstracts of the major conference EURO. The parameters of this model are estimated using expert models of EURO conference from 2006 till 2016. The results are compared with hierarchical multiclass SVM, probabilistic thematic model SuhiPLSA, and hierarchical naive Bayes approach.

This research is funded by the Russian Foundation for Basic Research, grant 16-07-01160.

[1] Kuzmin, A., A. Aduenko, and V. Strijov. 2014. Thematic classification using expert model for major conference abstracts. *Informational Technologies* 6:22–26.

Сила связи слов и оценка релевантности текста единице представления знаний в открытых тестах

Михайлов Дмитрий Владимирович ⋆ mdv74@list.ru
Козлов Александр Павлович caleo@yandex.ru
Емельянов Геннадий Мартинович Gennady.Emelyanov@novsu.ru
Великий Новгород, Россия, НовГУ

Разработка открытых тестов предполагает накопление и систематизацию экспертных знаний, исходно представляемых текстами естественного языка (ЕЯ). Помимо отбора фраз из готового текстового корпуса важнейшей составляющей здесь является формирование самого корпуса с включением в него публикаций, максимально релевантных заданным ситуациям предметной области теста и ЕЯ-формам их описания. При этом релевантность текста определяется суммарной численной оценкой значимости («силы») связи для встречающихся в его фразах сочетаний слов исходной фразы. Сортировкой по убыванию оценки значимости с последующей кластеризацией наиболее значимых выделяются связи из класса близких максимальному значению указанной оценки. Тогда для текстовых документов, максимально релевантных исходной фразе, будет найден максимум таких связей при максимальной суммарной силе всех найденных в исходной фразе связей. В работе рассматриваются варианты оценки значимости связи слов и их использование для выделения составляющих образа исходной фразы в виде слов и их сочетаний при формировании корпуса текстов по тематике теста. По сравнению с поиском совокупностей указанных составляющих на готовом синтаксически размеченном корпусе предложенный в работе метод отбора текстов позволяет в среднем в 15 раз сократить выход фраз, не релевантных исходной ни по описываемому фрагменту знания, ни по языковым формам его выражения [1].

Работа поддержана РФФИ (проект № 16-01-00004) и Минобрнауки РФ (базовая часть госзадания).

[1] *Михайлов Д. В., Козлов А. П., Емельянов Г. М.* Выделение знаний, языковых форм их выражения и оценка эффективности формирования множества тематических текстов // Компьютерная оптика, 2016. Т. 40, №4. С.572-582. http://www.computeroptics.smr.ru/KO/PDF/KO40-4/400417.pdf

Coupling strength of words and estimation of text relevance to unit of knowledge in open tests

Mikhaylov Dmitry ⋆ mdv74@list.ru
Kozlov Aleksandr caleo@yandex.ru
Emelyanov Gennady Gennady.Emelyanov@novsu.ru
Russia, Veliky Novgorod, NovSU

Development of open form test assignments requires the accumulating and ordering an expert knowledge initially represented by natural-language (NL) texts. In addition to selection of phrases from a ready text corpus, the most important task here is the formation of corpus itself with adding the publications most relevant to given situations of test subject area and forms of their NL-descriptions. In this paper the text relevance to initial phrase is defined by total numerical estimation of coupling strength of words (CSoW) from initial phrase jointly occurring in phrases of analyzed text. Sorting the word combinations by descending values of CSoW with the further clustering as the most significant, the word combinations closest to maximal value of the given estimation are selected. Then, for text documents which are maximally relevant, the maximum of such word combinations at maximal summary CSoW in initial phrase will be found. The paper considers the variants of coupling strength estimation for word combination and their application for search of distinct components which reflect the initial phrase in texts selected to the topical text corpus for given test. These components correspond to words and their combinations. In comparison with the search of such combined-together components on a ready syntactically marked corpus, the method suggested in the present paper for text selection can reduce, on average, by 15 times the output of phrases which are not relevant to initial neither at described knowledge fragment nor at its expression forms in a given NL [1]. This research is funded by the Russian Foundation for Basic Research, grant 16-01-00004, and the Ministry of Education and Science of Russia (the basic part of the state task).

[1] Mikhaylov, D., A. Kozlov, and G. Emelyanov. 2016. Extraction the knowledge and relevant linguistic means with efficiency estimation for formation of subject-oriented text set. *Comp. Opt.* 40(4): 572-582. http://www.computeroptics.smr.ru/KO/PDF/KO40-4/400417.pdf

Отбор кандидатов при поиске заимствований в коллекции документов на иностранном языке

Романов Алексей Владимирович[1,2]⋆ romanov@ap-team.ru
Хританков Антон Сергеевич[2] anton.khritankov@gmail.com
[1]Россия, Москва, ЗАО «Антиплагиат»
[2]Россия, Москва, МФТИ

Автоматическое детектирование заимствований текста на сегодняшний день является одной из актуальных задач в области анализа текстов и информационного поиска. Кросс-языковые заимствования — явление, объединяющее случаи текстовых заимствований, в которых документ-источник и документ, содержащий заимствования, представляют собой тексты на разных естественных языках. Формирование круга документов-кандидатов — один из логических этапов процесса детектирования кросс-языковых заимствований.

Предлагается алгоритм отбора кандидатов, который может быть использован для решения задачи поиска кросс-языковых заимствований. Алгоритм направлен на сглаживание неточностей перевода, производимого системой статистического машинного перевода, за счет кластеризации слов целевого языка и последующей замены слов документа на метки кластеров [1].

Описан алгоритм кластеризации, и в соответствии с ним произведена кластеризация векторных представлений английских слов, полученных с помощью дистрибутивной модели на корпусе англоязычных текстов; предложен метод отбора кандидатов, основанный на такой кластеризации; в соответствии с метриками качества информационного поиска на предварительно подготовленных данных оценено качество алгоритма отбора кандидатов и продемонстрирована эффективность предложенного метода.

[1] *Романов А. В., Кузнецова М. В., Бахтеев О. Ю., Хританков А. С.* Machine-translated text detection in a collection of Russian scientific papers // Компьютерная лингвистика и информационные технологии. — М.: РГГУ, 2016. С. 578–589. http://www.dialog-21.ru/media/3422/romanovavetal.pdf.

Candidate document retrieval for cross-lingual plagiarism detection

Romanov Alexey[1,2]⋆ romanov@ap-team.ru
Khritankov Anton[2] anton.khritankov@gmail.com
[1]Moscow, Russia, JSC Anti-Plagiat
[2]Moscow, Russia, MIPT

Automatic plagiarism detection is an important task of natural language processing and information retrieval. Cross-language plagiarism comprises cases of plagiarism from the source document in one language to the plagiarized document in another language. Candidate retrieval is one of the stages of cross-language plagiarism detection.

The present authors propose a method of candidate retrieval that can be used for cross-language plagiarism detection, given a tool that translates suspicious documents into the language of potential plagiarism sources. The method is aimed at smoothing of statistical machine translation inaccuracies by means of target language word clustering and subsequent mapping of words of the suspicious document onto corresponding cluster labels [1].

A clustering algorithm is descrtibed for English word embeddings that saves information about semantic relatedness of words. The authors propose a method of candidate retrieval, which is based on this clustering, and evaluate it on the prepared corpus of plagiarized texts in accordance with information retrieval quality metrics. Preliminary results show the feasibility of the approach.

[1] Romanov, A. V., M. V. Kuznetsova, O. Yu. Bakhteev, and A. S. Khritankov. 2016. Machine-translated text detection in a collection of Russian scientific papers. *Computational linguistics and intellectual technologies*. Moscow: RSUH. 578–589. Available at: http://www.dialog-21.ru/media/3422/romanovavetal.pdf (accessed September 29, 2016).

Мультимодальная тематическая модель текстов и изображений на основе использования их векторного представления

Смелик Николай Дмитриевич smelik@rain.ifmo.ru
Фильченков Андрей Александрович* afilchenkov@corp.ifmo.ru
Россия, Санкт-Петербург, Университет ИТМО

Совместная тематическая модель для текстов и изображений позволяет автоматически выделять темы из изображений на основе их описания и далее предлагать описания новых изображений. Целью данной работы является создание такой модели. Предложен подход к построению мультимодальной модели на основе векторного представления текстов и изображений. Векторы значимых слов строятся за счет применения Word2Vec, для изображений — как выход последнего неполносвязного слоя сверточной нейронной сети. Затем вектор изображения представляется в виде псевдодокумента, в котором словами будут векторные представления слов. Далее по коллекции псевдодокументов строится тематическая модель. Предложены алгоритм обучения тематической модели по коллекции аннотированных изображений, а также алгоритмы аннотирования нового изображения и иллюстрирования нового текста. Вычислительные эксперименты проводились на наборе данных Microsoft Common Object in Context, содержащем 21 000 изображений, с не менее чем пятью аннотациями для каждого. Эксперименты показали, что с точки зрения качества модели использование ARTM (additive regularization of topic models) дает существенный выигрыш по всем метрикам, чем PLSA (probabilistic latent semantic analysis). В задаче аннотирования изображений модель сравнивалась с моделями CORRLDA (correspondence latent Dirichlet allocation (LDA)), MIXLDA и sLDA (supervised LDA), в задаче иллюстрирования текстов — с MIXLDA. В обеих задачах предложенная модель показала более высокие результаты, чем аналоги [1].

Работа выполнена при поддержке Правительства Российской Федерации, грант 074-U01.

[1] *Смелик Н. Д., Фильченков А. А.* Мультимодальная тематическая модель текстов и изображений на основе использования их векторного представления // Machine Learning Data Anal., 2016 (в печати).

Multimodal topic model for texts and images utilizing their embeddings

Smelik Nikolay smelik@rain.ifmo.ru
***Filchenkov Andrey**⋆* afilchenkov@corp.ifmo.ru
St. Petersburg, Russia, ITMO University

A joint topic model for texts and images allows to extract image topics based on their text annotations and to suggest annotations for new images. In this paper, a novel multimodal topic model is introduced for images and texts.

The proposed model utilizes vector representation of texts and images. Vector representation for a text is based on Word2Vec embedding. Vector representation for an image is convolutional neural network feature map. Vector of image then is considered as a pseudodocument containing vectors of words instead of words. The proposed model is learnt on the resulting pseudodocument collection. An algorithm to learn the model, as well as an algorithm for image annotating and an algorithm for text illustrating with a learnt model are proposed.

For experiments, Microsoft Common Object in Context dataset is used. It contains 21,000 images, each has at least 5 annotations. The results show that usage of ARTM (additive regularization of topic models) leads to much more quality than usage of PLSA (probabilistic latent semantic analysis). The model was compared with CORRLDA (correspondence latent Dirichlet allocation (LDA)), MIXLDA, and sLDA (supervised LDA) in image annotating problem and with MIXLDA in text illustrating problem. In both cases, the proposed model shows better results [1].

The research was supported by the Government of the Russian Federation (grant 074-U01).

[1] Smelik, N., and A. Filchenkov. 2016 (in press). Multimodal topic model for texts and images utilizing their embeddings. *Machine Learning Data Anal.*

Аддитивная регуляризация мультимодальных иерархических тематических моделей

Чиркова Надежда Александровна[1,2]⋆ nadiinchi@gmail.com
Воронцов Константин Вячеславович[3] voron@forecsys.ru
[1]Россия, Москва, ЗАО «Антиплагиат»
[2]Россия, Москва, МГУ
[3]Россия, Москва, ФИЦ ИУ РАН

Вероятностные тематические модели выявляют семантику текстовых коллекций, описывая каждый документ дискретным распределением вероятностей на множестве тем. Иерархические модели делят темы на подтемы, что упрощает информационный поиск и навигацию по большим мультидисциплинарным коллекциям. В большинстве работ по иерархическому тематическому моделированию применяется байесовский вывод, что затрудняет введение тематических иерархий в тематические модели других видов. Не-байесовская аддитивная регуляризация тематических моделей (ARTM), наоборот, позволяет комбинировать любые тематические модели, если их специфические особенности формализуемы в виде критериев-регуляризаторов. Однако до сих пор иерархические модели не имели такой формализации.

В данной работе предлагаются регуляризаторы тематических иерархий, адаптируемые для широкого класса задач, в частности для тематизации мультимодальных и мультиязычных данных.

Рассматриваются иерархии, в которых каждая подтема может иметь несколько родительских, что особенно актуально для междисциплинарных коллекций научных статей. Предлагаемый подход позволяет контролировать разреженность отношения тема–подтема и автоматически определять число подтем каждой темы. При построении модели задается только число тем на каждом уровне иерархии. Аддитивная регуляризация не усложняет процесс обучения тематической модели, что делает данный подход масштабируемым на большие текстовые коллекции [1].

Работа поддержана грантами РФФИ 16-37-00498, 14-07-00847 и 14-07-00908.

[1] *Chirkova N., Vorontsov K.* Additive regularization for hierarchical multimodal topic modeling // J. Machine Learning Data Anal., 2016 (in press).

Additive regularization for hierarchical multimodal topic modeling

Chirkova Nadezhda[1,2]⋆ nadiinchi@gmail.com
Vorontsov Konstantin[3] vokov@forecsys.ru
[1]Moscow, Russia, JSC Anti-Plagiat
[2]Moscow, Russia, MSU
[3]Moscow, Russia, FRC CSC RAS

Probabilistic topic models uncover the latent semantics of text collections and represent each document by a multinomial distribution over topics. Hierarchical models divide topics into subtopics recursively, thus simplifying information retrieval, browsing, and understanding of large multidisciplinary collections. The most of existing approaches to hierarchy learning rely on Bayesian inference. This makes difficult the incorporation of topical hierarchies into other types of topic models. The authors use non-Bayesian multicriteria approach called Additive Regularization of Topic Models (ARTM), which enables to combine any topic models formalized via log-likelihood maximization with additive regularization criteria.

In this work, such formalization is proposed for topical hierarchies. Hence, one can easily adapt the hierarchical ARTM to a wide class of text mining problems, e. g., for learning topical hierarchies from multimodal and multilingual heterogeneous data of scientific digital libraries or social media.

The authors focus on topical hierarchies that allow a topic to have several parent topics, which is important for multidisciplinary collections of scientific papers. The regularization approach allows one to control the sparsity of the parent-child relation and automatically determine the number of subtopics for each topic. Before learning the hierarchy, one needs to fix the number of topics for each layer. The additive regularization does not complicate the learning algorithm; so, this approach is well scalable on large text collections [1].

This research is funded by the Russian Foundation for Basic Research, grants Nos. 16-37-00498, 14-07-00847, and 14-07-00908.

[1] Chirkova, N., and K. Vorontsov. 2016 (in press). Additive regularization for hierarchical multimodal topic modeling. *J. Machine Learning Data Anal.*

Параметрический подход к построению синтаксических деревьев для частично формализованных текстовых документов

Чувилин Кирилл Владимирович[1,2] kirill@chuvilin.pro
[1]Россия, Долгопрудный, МФТИ
[2]Россия, Протвино, ИФТИ

Работа посвящена исследованию возможности автоматического построения абстрактного синтаксического дерева для текстовых документов, формат которых не является полностью определенным стандартами или другими общими для всех документов правилами. В таких случаях нет возможности в автоматическом режиме построить синтаксический анализатор. Исследуются текстовые файлы в формате LaTeX. Актуальность их анализа обусловлена тем, что многие научные издательства и конференции используют систему верстки LaTeX, и это порождает важные прикладные задачи по автоматизации рубрикации, коррекции, сравнения, сбора статистики, отображения для WEB и т. п. При синтаксическом анализе документов в формате LaTeX требуются внешние данные о стилях. Предлагается метод их описания в формате JSON, который позволяет задавать не только информацию, необходимую для синтаксического анализа, но и метаинформацию, упрощающую дальнейший интеллектуальный анализ. Такой подход использован впервые. Описываются разработанные алгоритмы синтаксического анализа. Полученные результаты успешно применены в задачах сравнения, автоматической коррекции и рубрикации научных статей. Реализация разработанных алгоритмов доступна в виде набора библиотек, распространяемых по лицензии LGPLv3. Ключевыми особенностями предлагаемого подхода являются гибкость (в рамках рассматриваемой задачи) и простота описания параметров [1].

Работа выполнена при финансовой поддержке РФФИ, проекты № 16-37-60049 и № 16-07-01267.

[1] *Чувилин К. В.* Параметрический подход к построению синтаксических деревьев для частично формализованных текстовых документов // Машинное обучение и анализ данных, 2016 (в печати). Т. 2. № 1.

Parametric approach to the construction of syntax trees for partially formalized text documents

Chuvilin Kirill[1,2] kirill@chuvilin.pro

[1]Dolgoprudny, Russia, MIPT
[2]Protvino, Russia, ICPT

This article investigates the possibility of logical structure (abstract syntax tree) automatic construction for text documents, the format of which is not fully defined by standards or other rules common to all the documents. In such cases, there is no way to build the parser automatically. Text files in LaTeX format are used for the research. The relevance of LaTeX document analysis is due to the fact that many scientific publishings and conferences use LaTeX typesetting system, and this gives rise to important applied task of automation for categorization, correction, comparison, statistics collection, rendering for WEB, etc. The parsing of documents in LaTeX format requires additional data about styles. In this work, a method to describe them in JSON format is proposed. It allows to specify not only the information necessary to pars, but also metainformation that facilitates further data mining. This approach is used for the first time. The developed parsing algorithms are described. The results are successfully applied in the tasks of comparison, autocorrection, and categorization of scientific papers. The implementation of the developed algorithms is available as a set of libraries released under the LGPLv3. The key features of the proposed approach are flexibility (within the framework of the problem) and simplicity of the parameter descriptions [1].

The research was supported by the Russian Foundation for Basic Research (grants 16-37-60049 and 16-07-01267).

[1] *Chuvilin, K.* 2016. Parametric approach to the construction of syntax trees for partially formalized text documents. *Machine Learning Data Anal.* 2(1).

Мультимодальные тематические модели для разведочного поиска в коллективном блоге

Янина Анастасия Олеговна[1,2]⋆ yanina-n@yandex-team.ru
Воронцов Константин Вячеславович[1,2] vokov@forecsys.ru
[1]Россия, Москва, МФТИ
[2]Россия, Москва, ООО «Яндекс»

Разведочный информационный поиск нацелен на приобретение и систематизацию профессиональных знаний, в отличие от обычного поиска, отвечающего на короткие запросы массового пользователя. Исследуются методы тематического поиска документов по длинным текстовым запросам. Тематическая модель строится с помощью библиотеки с открытым кодом *BigARTM* (`bigartm.org`), основанной на *аддитивной регуляризации тематических моделей*. Она позволяет строить модели с заданными свойствами, комбинируя разнородные требования и источники данных. Тематический поиск реализуется путем сравнения сжатых тематических описаний запроса и документов. Для оценивания качества поиска разработана коллекция запросов — заданий разведочного поиска, которые сначала выполняются людьми (асессорами), затем системой тематического поиска, затем релевантность найденных ею документов снова оценивается асессорами. Данная методика позволяет, единожды сделав разметку результатов поиска, многократно оценивать качество поиска для различных тематических моделей и механизмов поиска. Эксперименты на коллекции из 132 тыс. статей коллективного блога `habrahabr.ru` показали, что тематический поиск находит больше релевантных документов, чем асессоры, сокращая среднее время поиска с получаса до секунд. Подбор тематической модели по критериям точности и полноты поиска позволил оптимизировать число тем и показать, что использование метаинформации (тегов, комментариев) улучшает качество поиска [1].

Работа поддержана грантами РФФИ 16-37-00498, 14-07-00847 и 14-07-00908.

[1] *Янина А. О., Воронцов К. В.* Мультимодальные тематические модели для разведочного поиска в коллективном блоге // Машинное обучение и анализ данных, 2016 (в печати).

Multimodal topic modeling for exploratory search in collective blog

Ianina Anastasia[1,2]⋆ yanina-n@yandex-team.ru
Vorontsov Konstantin[2] vokov@forecsys.ru
[1]Moscow, Russia, MIPT
[2]Moscow, Russia, Yandex LLC

Exploratory Search is a new paradigm in information retrieval focused on the acquisition and systematization of knowledge by professionals, unlike major Web search engines that answer short text queries of mass users. The present authors develop an exploratory search engine based on probabilistic topic modeling for seeking information thematically relevant to the long text queries. *Additive Regularization for Topic Modeling* (ARTM) is used to combine many requirements such as sparsity, diversity, and interpretability of topics and incorporate heterogeneous modalities such as authors, tags, and categories into the model. The parallelized online implementation of ARTM in open source library *BigARTM* (`bigartm.org`) has been used. The thematic search is implemented by maximizing cosine similarity between query and document both represented by their sparse distributions over topics. The present authors evaluate precision and recall of the thematic search by a two-step procedure. First, human assessors perform exploratory search tasks manually using any available search utilities (it takes them about 30 min per task in average). Second, they evaluate the relevance of search results found by the developed thematic search engine for the same tasks. The experiments on the collection of 132,000 articles from `habrahabr.ru` collective blog showed that thematic search provides comparable precision and better recall, also reducing search time from half an hour to seconds. With data labeled by assessors, the optimal number of topics was determined and it was shown that the joint use of all modalities (authors of articles, authors of comments, tags, and hub categories) significantly improves the search quality [1].

This research is funded by the Russian Foundation for Basic Research, grants 16-37-00498, 14-07-00847, and 14-07-00908.

[1] Ianina, A. O., and K. V. Vorontsov. 2016 (in press). Multimodal topic modeling for exploratory search in collective blog. *J. Machine Learning Data Anal.*

Агентное моделирование региональной эколого-экономической системы. Тематическое исследование для Республики Армения

Бекларян Левон Андреевич[1]⋆ beklar@cemi-rssi.ru
Акопов Андраник Сумбатович[2] aakopov@hse.ru
Бекларян Армен Левонович[2] abeklaryan@hse.ru
Сагателян Армен Карленович[3] ecocentr@sci.am
[1]Россия, Москва, ЦЭМИ РАН
[2]Россия, Москва, НИУ ВШЭ
[3]Республика Армения, Ереван, Центр эколого-ноосферных исследований Национальной академии наук Республики Армения

Рассматриваются актуальные вопросы моделирования эколого-экономической системы на примере Республики Армения (РА). Основываясь на методах агентного моделирования и системной динамики, создана имитационная модель эколого-экономической системы, позволившая построить Экологическую карту РА. Важной целью предлагаемого подхода является поиск сценариев рациональной модернизации предприятий, являющихся основными источниками выбросов вредных веществ, с одновременным определением эффективной стратегии государственного регулирования. Сформулирована и решена бикритериальная задача оптимизации характеристик эколого-экономической системы на примере РА [1].

Работа поддержана грантом РФФИ № 15-51-05011 Арм_а.

[1] *Beklaryan L. A., Akopov A. S., Beklaryan A. L., Saghatelyan A. K.* Agent-based simulation modelling for regional ecological-economic systems. A case study of the Republic of Armenia // Machine Learning and Data Anal., 2016 (в печати).

Agent-based simulation modeling for regional ecological-economic systems. A case study of the Republic of Armenia

Beklaryan Levon[1]⋆ beklar@cemi-rssi.ru
Akopov Andranik[2] aakopov@hse.ru
Beklaryan Armen[2] abeklaryan@hse.ru
Saghatelyan Armen[3] ecocentr@sci.am
[1]Moscow, Russia, CEMI RAS
[2]Moscow, Russia, HSE
[3]Yerevan, Republic of Armenia, CENS NAS RA

Actual problems of modeling of ecologic-economic systems on the example of the Republic of Armenia (RA) are considered. Based on methods of agent modeling and system dynamics, the simulation model of ecological-economic system, which has allowed constructing the RA Ecological Map, was created. The important purpose of the suggested approach is search of scenarios of rational modernization of the agent-enterprises, which are the main sources of emissions with simultaneous definition of effective strategy of the government regulation. The bi-criterial optimization problem for the ecological-economic system of RA is formulated and solved with the help of the developed genetic algorithm [1].

This research is funded by the Russian Foundation for Basic Research, grant 15-51-05011 Arm_a.

[1] Beklaryan, L. A., A. S. Akopov, A. L. Beklaryan, and A. K. Saghatelyan. 2016 (in press). Agent-based simulation modelling for regional ecological-economic systems. A case study of the Republic of Armenia. *Machine Learning and Data Anal.*

Применение информационно-энтропийного подхода для исследования особенностей адаптации студентов к обучению в вузе

Берестнева Ольга Григорьевна[1]⋆ ogb6@yandex.ru
Марухина Ольга Владимировна[1] Marukhina@tpu.ru
Шаропин Костантин Александрович[2] kashar@mail.ru
[1]Томск, Россия, ТПУ
[2]Москва, Россия, АНО ВПО МГЭИ

Как биологические, так и социальные системы существуют, приспосабливаются и развиваются благодаря обмену энергией, веществом и информацией с внешней средой. Для получения количественных характеристик процесса адаптации были введены энтропийные показатели состояния биосистемы. Авторами рассмотрена возможность применения данного подхода для исследования особенностей адаптации студентов к обучению в вузе. Для исследования особенностей адаптации студентов к учебному процессу был использован разработанный авторами метод моделирования адаптационных стратегий. В этом случае в качестве адаптационной функции выступает значение интегрального показателя I_{adapt} в моменты времени. При использовании интегрального критерия I_{adapt} предполагается, что существует некоторое «эталонное» («предпочтительное») состояние биосистемы, степень отклонения от которого в текущий момент времени и позволяет оценить I-критерий. Проведенные авторами исследования особенностей адаптации первокурсников к учебной деятельности показали, что в данном случае присутствуют четыре основных типа реакции организма на внешнее воздействие. Кроме четырех основных типов реакции организма человека на экстремальные воздействия окружающей среды было выявлено еще три типа адаптационных реакций [1].

Исследования выполнены при частичной финансовой поддержке Российского фонда фундаментальных исследований, проекты № 14-06-00026 и № 15-07-08922

[1] *Берестнева О.Г.*, *Марухина О.В.*, *Шаропин К.А.* Применение информационно-энтропийного подхода для исследования особенностей адаптации студентов к обучению в вузе // Науковедение, 2013. № 3. http://naukovedenie.ru/PDF/53tvn313.pdf.

Information and entropy approach in research of student's adaptation to the university training characteristics

Berestneva Olga[1]⋆ ogb6@yandex.ru
Marukhina Olga[1] Marukhina@tpu.ru
Sharopin Konstantin[2] kashar@mail.ru
[1]Russia, Tomsk, TPU
[2]Russia, Moscow, HEIM

Biological and social systems exist, adjust, and develop due to the power, substance, and information interchange with the external environment. To get the adaptation process quantity characteristics, the entropy parameters of biosystem state have been introduced. The authors considered the possibility of applying this approach to the research of students adaptation to the university training characteristics. The method of adaptation strategies modeling that was developed by the authors was used to study the characteristic properties of students adaptation to the academic activity. In this case, the adaptation function is an integral index value I_{adapt} at time periods. When the integral index I_{adapt} is used, it is supposed that there is some "reference" ("preferential") state of the biosystem and the deviation degree from this state in real time allows evaluating I-criterion. The studies of adaptation characteristic properties of the first-year students to the academic activity that were carried out by the authors showed that in this case, there are four main types of organism excitation response. In addition to the four main types of the human body's response to extreme environmental influences, there were identified three types of adaptive reactions [1].

The research is conducted with partial financial support from the Russian Foundation for Basic Research, projects Nos. 14-06-00026 and 15-07-08922.

[1] Berestneva, O. G., O. V. Marukhina, and K. A. Sharopin. 2013. Application of information and entropy approach to study the peculiarities of students adaptation to teaching at the university. *Naukovedenie* 3. Available at: http://naukovedenie.ru/PDF/53tvn313.pdf (accessed July 26, 2016).

Расчет параметров модельной гидротурбины по значению коэффициента быстроходности

Волков Юрий Степанович[1]⋆ volkov@math.nsc.ru
Богданов Владимир Васильевич[1] bogdanov@math.nsc.ru
Мирошниченко Валерий Леонидович[1] miroshn@math.nsc.ru
Салиенко Александр Евгеньевич[2] sa_cae@yahoo.com
[1]Россия, Новосибирск, ИМ СО РАН
[2]Россия, Сызрань, ОАО «Тяжмаш»

Ранее авторами разработана методика построения математической модели универсальной характеристики модельной гидротурбины по результатам энергетических испытаний [1]. Универсальная характеристика является основным документом для выбора параметров натурной гидравлической турбины (диаметр рабочего колеса, частота вращения и др.), чтобы обеспечить наиболее эффективную работу турбины при всех режимах ее эксплуатации на конкретной гидроэлектростанции. Для интегрального описания гидравлических качеств турбины по скорости вращения и пропускной способности, а также для сравнения между собой различных турбин в гидротурбостроении введен так называемый коэффициент быстроходности.

Предлагается методика расчета основных параметров гидротурбины и ее универсальной характеристики по промежуточному значению коэффициента быстроходности однотипных моделей гидротурбин без проведения энергетических испытаний. Подробное изложение методики приведено в работе [2].

Работа поддержана грантом РФФИ № 15-07-07530.

[1] *Волков Ю. С., Мирошниченко В. Л., Салиенко А. Е.* Математическое моделирование универсальной характеристики поворотно-лопастной гидротурбины // Машинное обучение и анализ данных, 2014. Т. 1. № 10. С. 1439–1450. http://jmlda.org/papers/doc/2014/no10/Volkov2014KaplanTurbine.pdf.

[2] *Волков Ю. С., Богданов В. В., Мирошниченко В. Л., Салиенко А. Е.* Расчет параметров модельной гидротурбины по значению коэффициента быстроходности // Сиб. ж. индустриальной мат., 2016 (в печати).

Calculation of the parameters of model hydraulic turbine using the value of the specific speed

Volkov Yuriy[1]⋆ volkov@math.nsc.ru
Bogdanov Vladimir[1] bogdanov@math.nsc.ru
Miroshnichenko Valery[1] miroshn@math.nsc.ru
Salienko Alexander[2] sa_cae@yahoo.com
[1]Novosibirsk, Russia, IM SB RAS
[2]Syzran, Russia, JSC Tyazhmash

Earlier, the authors developed a method of mathematical modeling of the hill diagram of model hydraulic turbine model on the power test results [1]. The hill diagram is the basic document for choice of full-scale hydraulic turbine parameters (turbine wheel diameter, rotating frequency, etc.) to ensure the most efficient operation of the turbine at all modes of its exploitation in a particular hydropower station. The integral parameter so-called specific speed was introduced in a hydroturbine constructing to characterize the properties of hydraulic turbines such as rotation speed and conveyance capacity, and also to compare different turbines.

The paper presents a method of calculating the basic parameters of hydraulic turbine and its hill diagram on the intermediate value of the specific speed of similar model hydraulic turbines without power tests. A detailed description of techniques is given in [2].

This research is funded by the Russian Foundation for Basic Research, grant 15-07-07530.

[1] Volkov, Yu. S., V. L. Miroshnichenko, and A. E. Salienko. 2014. Mathematical modeling of hill diagram for Kaplan turbine. *Machine Learning Data Anal.* 1(10):1439–1450. Available at: http://jmlda.org/papers/doc/2014/no10/Volkov2014KaplanTurbine.pdf (accessed September 27, 2016).

[2] Volkov, Yu. S., V. V. Bogdanov, V. L. Miroshnichenko, and A. E. Salienko. 2016 (in press). Calculation of parameters of model hydraulic turbine by value of the specific speed. *J. Appl. Ind. Math.*

Программный комплекс интерпретации данных возвратно-наклонного зондирования ионосферы

Грозов Виктор Петрович* grozov@iszf.irk.ru
Пономарчук Сергей Николаевич spon@iszf.irk.ru
Россия, Иркутск, ИСЗФ СО РАН

Метод вертикально-наклонного зондирования позволяет получать информацию о месте, откуда ведется зондирование, о рассеивающих областях, удаленных от него на тысячи километров. Результаты оперативной диагностики радиоканала по текущим данным возвратно-наклонного зондирования ионосферы (ВНЗ) могут быть использованы для восстановления пространственного распределения электронной концентрации в секторе зондирования ионосферы линейно частотно-модулированным (ЛЧМ) ионозондом в режиме ВНЗ, т. е. определение модов распространения в зависимости от отражения радиосигнала от ионосферных слоев. Рассмотрены методики и алгоритмы программного комплекса автоматической интерпретации радиофизической информации, получаемой на базе ЛЧМ-ионозонда Института солнечно-земной физики СО РАН, работающего в режимах ВНЗ. В рамках данной проблемы рассмотрены следующие задачи: (а) проведение предобработки для удаления шума с изображения и улучшения амплитудных характеристик; (б) сжатие данных с использованием клеточного автомата; (в) интерпретация ионограмм ВНЗ. Методика интерпретации ионограмм основана на использовании результатов моделирования частотных зависимостей минимального группового пути в режиме долгосрочного прогноза и результатов обработки экспериментальных данных. Представлены результаты оперативной диагностики коротковолнового радиоканала по текущим данным ВНЗ.

The program complex for interpretation of ionosphere backscatter sounding data

Grozov Viktor ⋆ grozov@iszf.irk.ru
Ponomarchuk Sergey spon@iszf.irk.ru
Irkutsk, Russia, ISTP SB RAS

Techniques and algorithms for program complex of automatic interpretation of the radiophysical information obtained on the basis of chirp ionosonde located at the Institute of Solar-Terrestrial Physics of the Siberian Branch of the Russian Academy of Sciences working in the mode of backscatter ionosphere sounding are considered. Within this problem, the following tasks are discussed: (a) carrying out preprocessing for removal of noise from the image and improvement of amplitude characteristics; (b) compression of data with use of the cellular automaton; and (c) backscatter ionograms interpretation and determination of propagation modes depending on reflection of radiosignal from ionospheric layers. The interpretation technique for ionograms is based on the results of frequency dependences modeling for the minimum group way in the mode of the long-term forecast and results of the experimental data processing. Also, the results of operative diagnostics for high-frequency radiochannel on the base of backscatter sounding current data are presented.

Моделирование и прогнозирование экологически обусловленной заболеваемости населения Байкальского региона

Ефимова Наталья Васильевна[1] medecolab@inbox.ru
Горнов Александр Юрьевич[2] gornov@icc.ru
Зароднюк Татьяна Сергеевна[2]⋆ tzarodnyuk@gmail.com
Аникин Антон Сергеевич[2] anton.anikin@gmail.com
[1]Россия, Ангарск, ФГБНУ ВСИМЭИ
[2]Россия, Иркутск, ИДСТУ СО РАН

Представлены результаты математического моделирования динамики здоровья населения с целью анализа и прогнозирования экологически обусловленных изменений заболеваемости при изменении факторов внешней среды. Выполнено изучение динамики заболеваемости населения и качества среды обитания промышленных центров; разработаны математические модели, учитывающие динамику факторов окружающей природной и социальной среды.

Авторами разработан программный комплекс, включающий современные конкурентоспособные нелокальные методы и вспомогательные локальные алгоритмы, с тонкими настройками алгоритмических параметров для повышения их эффективности на изучаемом классе задач. Апробация предложенных подходов проведена на примере развития на территории Иркутской области промышленных предприятий.

Построенные математические модели системы «заболеваемость – факторы окружающей среды» позволили получить достаточно точные прогнозы и разработать на их основе управляющие решения по минимизации рисков здоровью населения [1].

Исследования выполнены при частичной финансовой поддержке РФФИ, проект р_сибирь_а № 14-47-04089.

[1] Ефимова Н. В., Горнов А. Ю., Донских И. В., Зароднюк Т. С., Аникин А. С., Елфимова Т. А. Методические подходы к изучению роли факторов, формирующих заболеваемость населения крупного промышленного центра // Современные наукоемкие технологии, 2015. Т. 12. № 5. С. 781–785.

Simulation and prediction of environmentally induced morbidity of the population of the Baikal region

Efimova Natalia[1] medecolab@inbox.ru
Gornov Alexander[2] gornov@icc.ru
Zarodnyuk Tatiana[2]⋆ tzarodnyuk@gmail.com
Anikin Anton[2] anton.anikin@gmail.com
[1]Angarsk, Russia, FSBSI ESSIMER
[2]Irkutsk, Russia, ISDCT of SB RAS

The paper presents the results of mathematical modeling of the public health dynamics for the purpose of analysis and forecasting environmentally induced morbidity changes when changing environmental factors. The dynamics of morbidity and quality of habitat industrial centers have been studied; mathematical models which take into account the dynamics of the natural and social environment factors have been developed.

The authors have developed a software including advanced nonlocal techniques practices and supporting local algorithms with settings of algorithmic parameters for increase of their effectiveness in the considerable class of problems. Testing of the proposed approaches was carried out on the example of development of industrial enterprises in the Irkutsk region.

Constructed mathematical models of "morbidity – environmental factors" system allowed to get the accurate predictions and to develop the control solutions on their basis to minimize the risks to public health [1].

The research was carried out with the partial financial support of the Russian Foundation for Basic Research, project No. 14-47-04089.

[1] Efimova, N., A. Gornov, I. Donskih, T. Zarodnyuk, A. Anikin, and T. Elfimova. 2015. Methodological approaches to the study of the role of factors forming the morbidity of a large industrial center. *Modern High Technologies* 12(5):781–785.

Особенности использования технологий BIG DATA в задачах медицинской диагностики

Ильясова Наталья Юрьевна[1,2]⋆ ilyasova@smr.ru
Куприянов Александр Викторович[1,2] akupr@smr.ru
Парингер Рустам Александрович[1,2] rusparinger@gmail.com
[1]Россия, Самара, ИСОИ РАН — филиал ФНИЦ «Кристаллография и фотоника» РАН
[2]Россия, Самара, СГАУ

Изложены основные результаты исследований в области применения технологий интеллектуального анализа больших массивов данных в медицине. Описана информационная технология интеллектуального анализа различных классов диагностических изображений, основанная на методологии выделения диагностически значимой информации и построении информативных признаков [1]. Показано, что использование технологий Big Data в разрабатываемых системах медицинской диагностики позволило за счет привлечения разнородных источников диагностической информации и большего объема данных усовершенствовать обучающую выборку и повысить достоверность постановки диагноза до 0,95 (см. таблицу). Работа поддержана грантами РФФИ № 15-29-07077 и № 16-41-630761.

Повышение достоверности диагностики с BigData

Набор данных	Повышение разделимости классов	Достоверность
Рентген костей	8%	0,90
Ультразвуковое исследование почек	23%	0,87
Компьютерная томография легких	17%	0,95
Кровеносные сосуды	21%	0,94

[1] *Ильясова Н. Ю., Куприянов А. В., Попов С. Б., Парингер Р. А.* Особенности использования технологий BIG DATA в задачах медицинской диагностики // Системы высокой доступности, 2016. Т. 12. № 1. С. 45–52. http://www.radiotec.ru/catalog.php?cat=jr15&art=1773.

Particular usage characteristics of BIG DATA in medical diagnostics tasks

Ilyasova Nataly[1,2]⋆ ilyasova@smr.ru
Kupriyanov Alexandr[1,2] akupr@smr.ru
Paringer Rustam[1,2] rusparinger@gmail.com
[1]Samara, Russia, IPSI RAS — Branch of the FSRC "Crystallography and Photonics" RAS
[2]Samara, Russia, SSAU

The paper presents the main research results in the area of data mining application to medicine. A new information technology of data mining is proposed for different classes of biomedical images based on the methodology of diagnostically relevant information selection and creation of informative characteristics [1]. Application of Big Data technologies in the proposed systems of medical diagnostics has allowed to improve the learning set quality and to reduce the classification error. Based on these results, the conclusion is made, that the usage of many heterogeneous sources of diagnostic information made it possible to improve the overall quality of the diagnostics (see the table).

This research is funded by the Russian Foundation for Basic Research, grants 15-29-07077 and 16-41-630761.

Changes of classes separability and diagnostic reliability for different types of diagnostic data

Data set	Rising of classes separability	Reliability
Bones X-rays	8%	0,90
Blood vessels	21%	0,94
Ultrasonography of kidneys	23%	0,87
Computer tomography of lungs	17%	0,95

[1] Ilyasova, N. Yu., A. V. Kupriyanov, R. A. Paringer, and S. B. Popov. 2016. Particular usage characteristics of BIG DATA in medical diagnostics tasks. *Highly Available Systems* 12(1):48–55. Available at: http://www.radiotec.ru/catalog.php?cat=jr15&art=17734 (accessed September 14, 2016).

Использование методов Data Mining при принятии медицинских диагностических решений

Мокина Елена Евгеньевна Alisandra@tpu.ru
Марухина Ольга Владимировна* Marukhina@tpu.ru
Шагарова Мария Дмитриевна mds1@tpu.ru
Дубинина Ирина Анатольевна Dubinina@tpu.ru
Россия, Томск, ТПУ

Статья посвящена рассмотрению вопросов использования методов Data Mining при исследовании медицинских данных и построения системы поддержки принятия решений (СППР) на основе результатов исследования, в данном случае выявления наличия неврологических заболеваний по результирующим показателям опросников качества жизни и тревоги и депрессии.

В данном исследовании база знаний содержит логические правила в виде продукционных моделей. Метод, используемый для построения логических правил, — деревья решений.

Поддержка процесса постановки диагноза пациенту с помощью СППР включает в себя такие задачи, как анализ данных опроса пациента и сопоставление его с моделью для постановки диагноза; отслеживание изменений показателей качества жизни пациента в динамике, поскольку в ходе лечения дополнительная диагностика позволяет улучшить лечебный процесс и прогнозировать изменения показателей качества жизни.

В качестве инструментария для построения дерева решений и логических правил выбрана среда RapidMiner, представляющая собой комплексную систему, которая обладает набором алгоритмов для обработки и анализа данных, в том числе обработку больших массивов.

Исследования выполнены при частичной финансовой поддержке грантов РФФИ, проекты № 14-06-00026 и № 14-07-00675

[1] *Мокина Е. Е., Марухина О. В., Шагарова Д. В., Дубинина И. А.* Использование методов Data Mining при принятии медицинских диагностических решений // Фундаментальные исследования, 2013. № 5(2). С. 269–274.

Applying data mining techniques when making medical diagnostic decisions

Mokina Elena Alisandra@tpu.ru
Marukhina Olga ⋆ Marukhina@tpu.ru
Shagarova Maria mds1@tpu.ru
Dubinina Irina Dubinina@tpu.ru
Tomsk, Russia, TPU

The paper considers the application of Data Mining techniques for studying medical data and building the decision support system on the basis of research results being, in the present case, the detection of the neurological disorders by the result indicators of the surveys on living standard, anxiety, and depression.

Knowledge should describe the new relationship between the properties to predict the values of some attributes on the basis of others. In this research, the knowledge database contains the logical rules in the form of production models. The method used to construct the logical rules is decision trees building.

Diagnosis, including the one of neurological disorder, influences the indicators of living standards and patients anxiety level. Support of establishing diagnosis in patients with a certain nosology via the decision support system includes such tasks as: analysis of patients survey outcomes and comparison them with the model of establishing diagnosis; monitoring of the changes of a patients living standard in dynamics, while additional diagnostics enables to improve treatment and foresee the changes of living standard indicators.

RapidMiner environment was used as a tool for a decision tree building and logical rules establishing. RapidMiner is a complex system which implements the set of algorithms for data processing and analyzing, including the ones for processing large data arrays.

The research is conducted with partial financial support from the Russian Foundation for Basic Research, projects Nos. 14-06-00026 and 14-07-00675.

[1] Mokina, E. E., O. V. Marukhina, D. V. Shagarova, and I. A. Dubinina. 2013. Applying data mining techniques when making medical diagnostic decisions. *Fundamental Research* 5(2):269–274.

О проблеме введения средств распределённого многоагентного программирования в логический язык со строгой типизацией

Морозов Алексей Александрович[1,2]⋆ morozov@cplire.ru
Сушкова Ольга Сергеевна[1] o.sushkova@mail.ru
Полупанов Александр Фёдорович[1] sashap55@mail.ru
[1]Россия, Москва, ИРЭ им. В. А. Котельникова РАН
[2]Россия, Москва, ФГБОУ ВО МГППУ

Рассмотрена проблема введения средств многоагентного программирования в логический язык со строгой (сильной) типизацией. Для обеспечения корректного взаимодействия агентов предложен подход на основе комбинированной системы типов, идея которого заключается в том, что принцип статической проверки типов в логическом языке смягчается, а именно: проверка корректности использования объекта, полученного из другой программы (агента), откладывается до тех пор, пока он не понадобится для удалённого вызова предикатов. Во всех остальных случаях осуществляется статическая проверка типов. На основе предложенного подхода реализовано расширение объектно-ориентированного логического языка Акторный Пролог, поддерживающее распределённое программирование. Конечной целью создания этого расширения языка является разработка средств распределённого логического программирования для многоагентной обработки видеоинформации и распределённого интеллектуального видеонаблюдения [1].

Работа поддержана РФФИ, проект № 16-29-09626-офи_м.

[1] *Морозов А. А., Сушкова О. С., Полупанов А. Ф.* О проблеме введения средств распределённого многоагентного программирования в логический язык со строгой типизацией // Ж. радиоэлектроники, 2016. № 7. http://jre.cplire.ru/jre/jul16/9/text.pdf.

Incorporation of distributed multiagent programming means in a strongly typed logic language

Morozov Alexei[1,2]⋆ morozov@cplire.ru
Sushkova Olga[1] o.sushkova@mail.ru
Polupanov Alexander[1] sashap55@mail.ru
[1]Moscow, Russia, Kotel'nikov IRE RAS
[2]Moscow, Russia, MSUPE

The paper addresses the problem of the development of agent logic programming means. The Actor Prolog object-oriented logic language extension that supports distributed logic programming and remote predicate calls is described. This is intended for the multiagent visual surveillance system implementation, that is, for the development of logic programs (agents) that acquire, analyze the video stream semantics in real time, and communicate with each other to facilitate the analysis and share obtained information/conclusions. The Actor Prolog language has a strong type system that is an important feature of the language and is necessary for the fast and reliable executable code generation. Thus, the contradiction between the language strong type system and the idea of the software agents' independency was a problem to be resolved in the course of adapting the Actor Prolog language to the multiagent paradigm. The problem of incorporation of distributed multiagent programming means into the strongly typed logic language is considered. The approach to the multiagent interaction based on the dynamic and static typing fusion is proposed. This approach combines the advantages of the static type-checking for the high-performance code generation with the flexibility of the dynamic type-checking that is necessary for the multiagent systems programming [1].

This research is funded by the Russian Foundation for Basic Research, grant 16-29-09626.

[1] Morozov, A. A., O. S. Sushkova, and A. F. Polupanov. 2016. Incorporation of distributed multiagent programming means in a strongly typed logic language. *Zh. Radioelektroniki* [J. of Radioelectronics] 7. Available at: http://jre.cplire.ru/jre/jul16/9/text.pdf (accessed September 16, 2016).

Анализ безопасности распределенных информационных систем на основе беспризнакового распознавания образов

Руднев Дмитрий Олегович dima__rudnevl@mail.ru
Сычугов Алексей Алексеевич* xru2003@list.ru
Россия, Тула, ТулГУ

В настоящее время активно развиваются технологии распределенных вычислений и становится актуально решение задачи обеспечения информационной безопасности таких систем.

В распределенных информационных системах (РИС) невозможно управление и мониторинг средств безопасности со стороны владельца информации. Таким образом, актуальна задача разработки методов, позволяющих оценить владельцем информации элементы РИС с точки зрения безопасности и одновременно сохранить конфиденциальность сведений о средствах защиты и состоянии элементов РИС.

Понятие безопасности элемента РИС зависит от величины доверия к элементу, которая определяется вероятностью того, что данные, переданные на узел РИС, и результаты вычислений не будут скомпрометированы и искажены. Предлагается анализировать меру похожести состояния элемента на заранее заданный базис с известным значением доверия. Таким образом, к предлагаемому подходу можно применить методы беспризнакового распознавания образов.

В основу построения функции доверия предлагается заложить следующий принцип — доверие выше к тому элементу, чье будущее состояние более предсказуемо. Был поставлен численный эксперимент, который показал высокую точность распознавания для отдельных элементов.

Предложенный метод не может являться единственным способом защиты информации в РИС, но его использование позволит повысить сложность взлома информационных систем в целом [1].

Работа поддержана грантом РФФИ № 16-07-01008.

[1] *Руднев Д.О., Сычугов А.А.* Анализ безопасности распределенных информационных систем на основе беспризнакового распознавания образов // Известия ТулГУ. Технические науки, 2016. Вып. 10.

Analysis of the security of distributed information systems based on featureless pattern recognition

Rudnev Dmitry dima__rudnevl@mail.ru
Sychugov Alexey⋆ xru2003@list.ru
Tula, Russia, TulSU

There are currently actively developing distributed computing technology. It is important to find a solution to the problem of information security of such systems.

It is impossible to control and monitor the security features in the distributed information systems. Thus, task of developing methods to assess the elements of the distributed system in terms of security and at the same time preserve the confidentiality of information about elements state is actual.

The concept of security distributed system‘s element depends on trust to the element. Trust is determined by the probability that the data transmitted to the distributed system site and the results of the calculations will not be compromized and distorted. It is proposed to analyze the similarity measure of the state of an element on a predetermined basis with a known value of the trust. Thus, the methods of featureless pattern recognition can be applied to the proposed approach.

In the basis of confidence-building function, it is proposed to put the following principle: trust up to the element whose future state is more predictable. Numerical experiment was raised which showed high recognition accuracy for individual items.

The proposed method cannot be the only way to protect the information in the distributed systems, but its use will increase the difficulty of hacking information systems in general [1].

This research is funded by the Russian Foundation for Basic Research, grant 16-07-01008.

[1] Rudnev, D., and A. Sychugov. 2016. Analysis of the security of distributed information systems based on featureless pattern recognition. *News of Tula State Univ. Technical Sciences* 10.

Комплексирование данных из разнородных источников в задачах моделирования транспортных потоков

Старожилец Всеволод Михайлович[1,2]⋆ starvsevol@gmail.com
Чехович Юрий Викторович[1,2] chehovich@forecsys.ru
[1]Россия, Москва, ФИЦ ИУ РАН
[2]Россия, Долгопрудный, МФТИ

Исследуется задача агрегации данных с GPS-треков и дорожных датчиков для построения и решения разностной схемы, соответствующей выбранной математической модели транспортного потока. Рассматриваются отдельно задачи для данных с автомагистрали и с въездов и съездов.

Для данных с автомагистрали предложен метод агрегации данных GPS-треков и дорожных датчиков, основанный на построении линейной модели для скорости и числа автотранспортных средств (АТС). Критерием качества полученной модели является среднеквадратичная ошибка между оцененным числом проехавших АТС и реальным. Отметим, что полученная модель может быть использована на участках автомагистралей, на которых отсутствуют дорожные датчики. Для данных с въездов и съездов был разработан метод восстановления суммарного потока, основанный на уравнениях баланса.

Для обеих задач проведены эксперименты на реальных данных с использованием предложенных алгоритмов. Для проведения вычислительных экспериментов использовались данные с GPS-треков от сервиса Яндекс.Пробки и данные с дорожных датчиков от Центра организации дорожного движения. В качестве автомагистрали рассматривалась Московская кольцевая автомобильная дорога [1].

Работа поддержана грантом РФФИ № 14-07-00685.

[1] *Старожилец В. М., Чехович Ю. В.* Комплексирование данных из разнородных источников в задачах моделирования транспортных потоков // Машинное обучение и анализ данных, 2016 (в печати).

Aggregation of data from different sources in traffic flow tasks

Starozhilets Vsevolod[1,2]⋆ starvsevol@gmail.com
Chehovich Yury[1,2] chehovich@forecsys.ru
[1]Moscow, Russia, FRC CSC RAS
[2]Dolgoprudny, Russia, MIPT

Data aggregation problem, where data are taken from GPS-tracks and traffic detectors, has been studied. Aggregated data are used to state and solve finite differences equation corresponding to the chosen traffic flow mathematical model. The problem is separated in two ones: the first one is about highway data and the second one is about entrances and exits data.

The authors propose to use a linear model to estimate speed and number of cars taking into account the highway data from GPS-tracks and traffic detectors. The quality criteria are mean squared error and correlation coefficient. Note that the built model can be used on highway data, which do not have data from traffic detectors, but have only data from GPS-tracks. For entrances and exits data, a method to recover summary total flow has been developed. This method is based on the preservation of cars in transport network.

For both problems, on real data, the computational experiment is provided and the performance of the proposed approaches is demonstarated. Data from GPS-tracks were provided by Yandex.Traffic and data from traffic detectors were provided by Moscow traffic management center. Moscow Ring Road was used as a highway [1].

This research is funded by the Russian Foundation for Basic Research, grant 14-07-00685.

[1] Starozhilets, V., and U. Chehovich. 2016 (in press). Aggregation of data from different sources in traffic flow tasks. *Machine Learning Data Anal.*

Содержание

Contents

Авторский указатель

Author index

MachineLearning.ru

http://www.machinelearning.ru/

Профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных. Цели ресурса — сконцентрировать информацию о достижения ведущих научных школ; способствовать обмену опытом, накоплению и распространению научных знаний; предоставить площадку для виртуальных научных семинаров и обсуждений.

Журнал «Машинное обучение и анализ данных»

http://jmlda.org

Журнал Машинное обучение и анализ данных публикует новые теоретические и обзорные статьи с результатами научных исследований в области искусственного интеллекта, теоретической информатики и приложений. Цель журнала — развитие теории машинного обучения, интеллектуального анализа данных и методов проведения вычислительных экспериментов. Принимаются статьи на русском и английском языках.

Открытый конкурс на лучшую технологию дешифрирования аэрокосмической информации

http://fpi.gov.ru/activities/konkurs/spacemap

http://dataring.ru

Фонд перспективных исследований совместно с Фондом «Сколково» проводят конкурс по решению следующих задач: распознавание единичных строений на аэрофотоснимках поверхности Земли в инфракрасном диапазоне; распознавание и классификация транспортных средств на спутниковых снимках в видимом диапазоне.

Научное издание

Интеллектуализация
обработки информации
ИОИ-2016

Тезисы докладов
11-й Международной конференции
(Москва, Россия – Барселона, Испания)

Напечатано с готового оригинал-макета

Сдано в набор 01.08.16. Подписано в печать 29.09.16.
Формат 60×90/16. Бумага офсетная. Печать цифровая.
Усл.-печ. л. 14,83. Уч.-изд. л. 11,0. Тираж 150 экз.

Заказ № 946.

Издательство «ТОРУС ПРЕСС»
121614, г. Москва, ул. Крылатская 29-1-43
e-mail: torus@torus-press.ru
http://www.torus-press.ru

Отпечатано в НИПКЦ «Восход-А» с готовых файлов
Москва 109052, ул. Смирновская, д. 25, стр. 3, офис 101