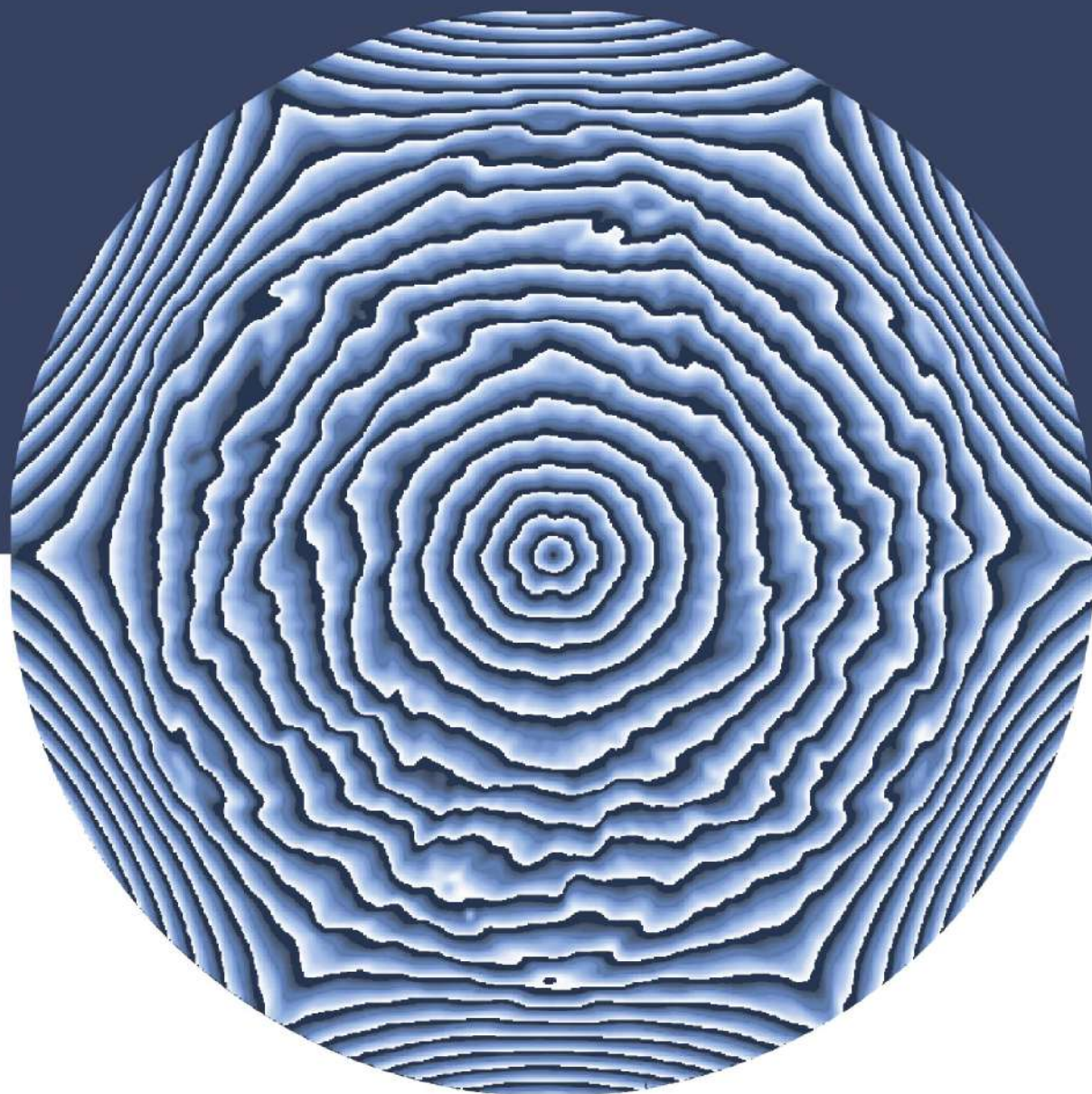


17-21 апреля, Самара, Россия

**Сборник трудов  
ИТНТ**

**2023**



IX Международная конференция и молодёжная школа  
**«Информационные технологии  
и нанотехнологии»**

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ  
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ  
«САМАРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ  
УНИВЕРСИТЕТ ИМЕНИ АКАДЕМИКА С.П. КОРОЛЕВА»  
(САМАРСКИЙ УНИВЕРСИТЕТ)

ИНСТИТУТ СИСТЕМ ОБРАБОТКИ ИЗОБРАЖЕНИЙ РАН –  
ФИЛИАЛ ФНИЦ "КРИСТАЛЛОГРАФИЯ И ФОТОНИКА" РАН

## ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И НАНОТЕХНОЛОГИИ (ИТНТ-2023)

Том 5. Науки о данных

*Сборник трудов по материалам  
IX Международной конференции и молодежной школы  
(г. Самара, 17-23 апреля 2023 г.)*

Одобрено редакционно-издательским советом федерального государственного автономного образовательного учреждения высшего образования «Самарский национальный исследовательский университет имени академика С.П. Королева»

© Самарский университет, 2023  
ISBN 978-5-7883-1921-6 (т. 5)  
ISBN 978-5-7883-1923-0

САМАРА  
Издательство Самарского университета  
2023

УДК 004.9  
ББК 32.973  
И741

**И741 Информационные технологии и нанотехнологии (ИТНТ-2023):** сборник трудов по материалам IX Международной конференции и молодежной школы (г. Самара, 17-23 апреля 2023 г.): в 6 томах / Министерство науки и высшего образования Российской Федерации, Самарский университет, Институт систем обработки изображений РАН – филиал ФНИЦ "Кристаллография и фотоника" РАН. – Самара: Издательство Самарского университета, 2023. – **Том 5. Науки о данных** / под редакцией к.т.н. Е.В. Гошина. – 1 CD-ROM (9,05 Мб). – Загл. с титул. экрана. – Текст. Изображение: электронный.

**ISBN 978-5-7883-1921-6 (т. 5)**  
**ISBN 978-5-7883-1923-0**

Тематика Конференции ИТНТ-2023 охватывает широкий круг областей применения информационных технологий в науке и высокотехнологичных отраслях промышленности. Одним из приоритетных направлений работы Конференции является образовательный аспект, заключающийся в предоставлении студентам и молодым ученым возможности ознакомиться с новейшими научными достижениями по тематике Конференции, а также с уникальным научным оборудованием и лабораторной базой Самарского университета, используемой для реализации современных научных проектов.

УДК 004.9  
ББК 32.973

**Минимальные системные требования:**

PC, процессор Pentium, 160 МГц; оперативная память 32 Мб;  
на винчестере 16 Мб; Microsoft Windows  
XP/Vista/7; разрешение экрана 1024x768 с глубиной цвета 16 бит;  
DVD-ROM2-х и выше, мышь; Adobe Acrobat Reader.

Редактор тома Е.В. Гошина

Выпускающий редактор В.Д. Зайцев

Подписано для тиражирования 07.07.2023.

Объем издания 9,05 Мб.

Количество носителей 1 диск.

Тираж 11 дисков.

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ  
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ  
«САМАРСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ  
УНИВЕРСИТЕТ ИМЕНИ АКАДЕМИКА С.П. КОРОЛЕВА»  
(САМАРСКИЙ УНИВЕРСИТЕТ)

443086, САМАРА, МОСКОВСКОЕ ШОССЕ, 34.

Издательство Самарского университета.  
443086, Самара, Московское шоссе, 34.

# Оглавление

Предисловие	6-10
1. Вторые моменты очередей в системах массового обслуживания с групповыми пуассоновскими потоками Б.Я. Лихтциндер, В. И. Моисеев, А.Ю. Привалов	050192
2. Эффективная Распределенная Обработка Больших Данных на Основе Наименьшего Информационного Пространства П.В. Голубцов	050252
3. Численная идентификация граничных условий в модели реакции-диффузии Д.В. Галушкина, А.Н. Кувшинова, Ю.В. Цыганова	050322
4. Идентификация параметров моделей дискретных стохастических систем с мультипликативными и аддитивными шумами А.В. Цыганов, Ю.В. Цыганова, А.В. Голубков	050682
5. Предсказание метеорологических величин с помощью гибридного метода обработки временных рядов Е.А. Черных, Н.Е. Шапкина, П.В. Голубцов	050912
6. Построение алгоритма аннотирования русскоязычных текстовых данных социальных сетей с использованием переносимого обучения Д.С. Баканов, А.В. Куприянов	050942
7. Математическое моделирование вольт-амперной характеристики мемристора с учетом его неоднородности Д.В. Продан	051012
8. Квадратно-корневой алгоритм вычисления отношения правдоподобия в задаче обнаружения изменения и идентификации режима движения А.В. Голубков, Ю.В. Цыганова, А.В. Цыганов	051112
9. Организация высокопроизводительных вычислений для исследования живучести энергетических инфраструктур А.В. Еделев, С.А. Горский, А.Г. Феоктистов, И.В. Бычков	051132
10. Метод анализа нестационарных сигналов на основе декомпозиции данных и вейвлет-преобразования Б.С. Мандрикова, О.И. Есиков	051162
11. Построение и идентификация параметров дискретной стохастической модели годового хода температуры воздуха М.А. Шугурова, А.В. Цыганов	051192
12. Метод обнаружения аномалий в природных данных на основе нейронных сетей и вейвлет-фильтрации О.В. Мандрикова, Ю.А. Полозов	051232
13. Применение метода активных контуров в задачах цефалометрии Ю.Ж. Пчелкина, Р.А. Парингер, А.В. Куприянов, П.Е. Савельева	051442
14. Технология автоматизированного интеллектуального отбора информативных признаков для задачи классификации областей натуральных гиперспектральных изображений М.И. Хотилин	051642
15. Быстрая одноклассовая SVM классификация для большой обучающей совокупности М.Ю. Курбаков, В. В. Сулимова	051742
16. О логической классификации целочисленных данных Е. В. Дюкова, Г. О. Масляков, А. П. Дюкова	051812
17. Анализ влияния различных аспектов личности студента на академическую успеваемость Н.В. Пустовалова, Т.В. Авдеевко	051922

18. Выявление проблемных вопросов по социально-направленным тематикам на основе данных открытых источников  
О.К. Головнин, А.В. Кривошеев, И.Н. Дубинина, П.В. Ситников, А.В. Иващенко 051972
19. Использование свойств вейвлет-преобразования в задачах поиска закономерностей  
Е.А.Нелюбина, В.В.Рязанов, А.П.Виноградов 052072
20. Алгоритм обнаружения и выделения сигналов в сильно зашумленных потоках данных  
В.А. Засов, М.В. Ромкин 052582
21. Сравнение эффективности методов машинного обучения в задаче оценки стоимости недвижимости  
Е.О. Агафонова, А.А. Белоусов 053032
22. Annotation of mathematical formulas in PDF documents  
K. Nikolaev, O. Nevzorova 053052
23. Способ темпоральной интерполяции толщины подвергающейся коррозии стенки газопровода согласованной с физической моделью  
Р.Р. Габбасов, Р.А. Парингер 053592

# ПРЕДИСЛОВИЕ

Конференция ИТНТ-2023 проводится с целью предоставления возможности научных дискуссий и обсуждения результатов фундаментальных и прикладных исследований в области информационных технологий и нанотехнологий, привлечения молодежи в сферу передовых научных исследований, обмена опытом научно-образовательной деятельности при подготовке ИТНТ-специалистов.

Тематика Конференции ИТНТ-2023 охватывает широкий круг областей применения информационных технологий в науке и высокотехнологичных отраслях промышленности.

Основными направлениями работы Конференции ИТНТ-2023 являются:

Компьютерная оптика и нанофотоника

- дифракционная оптика;
- планарные оптические структуры;
- гиперспектральные системы;
- нанофотоника;
- системы оптической сенсорики, передачи и обработки информации;
- сингулярная оптика.

Информационные технологии дистанционного зондирования Земли

- информационные технологии в проектировании космических аппаратов дистанционного зондирования Земли и полезных нагрузок для них;
- программные и математические решения для управления движением космических аппаратов наблюдения;
- программные и аппаратные средства для получения, обработки и анализа данных, получаемых с космических аппаратов дистанционного зондирования Земли;
- математическое моделирование процессов функционирования космических аппаратов дистанционного зондирования Земли;
- современные проектные решения для создания космических аппаратов мониторинга Земли и околоземного пространства и их группировок, в том числе на базе аппаратов типа CubeSat;
- системы дистанционного зондирования Земли на основе БПЛА.

Распознавание, обработка и анализ изображений

- математические методы цифровой обработки изображений и распознавания образов
- трёхмерное зрение
- биометрические системы на основе изображений
- геоинформационные системы и технологии
- защита и верификация мультимедиа

Искусственный интеллект

- новые подходы, тренды и фундаментальные результаты в сфере искусственного интеллекта и его приложениях к распознаванию образов и анализу изображений, обработке текстов, речевой информации;
- нейросетевые методы и глубокое обучение;
- прикладные технологии искусственного интеллекта в обработке изображений, беспилотном транспорте, производственных и сельскохозяйственных приложениях, медицинских приложениях, экологии, мониторинге окружающей среды и других;
- программные технологии для решения задач искусственного интеллекта – фреймворки, библиотеки, открытые инициативы и сообщества;
- мультидисциплинарные аспекты искусственного интеллекта и машинного обучения.

Науки о данных

- Компьютерные науки:
  - инженерия данных;
  - визуализация данных;
  - математические методы анализа данных;
  - программные платформы и библиотеки для работы с данными;
  - аппаратные средства хранения и обработки данных;
  - высокопроизводительные, параллельные и облачные вычисления, технологии обработки больших данных;
  - базы данных, инструменты и языки для работы с базами данных.
- Прикладные задачи интеллектуального анализа данных:
  - решение актуальных прикладных задач.

Информационные технологии в биомедицине

- математические методы обработки биомедицинских данных, сигналов, изображений, биомедицинская визуализация;
- интеллектуальный анализ биомедицинских данных, системы поддержки принятия врачебных решений;
- технологии искусственного интеллекта в обработке биомедицинских данных, нейронные сети и глубокое обучение в биомедицинских приложениях;
- технологии дополненной и виртуальной реальности в биомедицинских приложениях;

- медицинские информационные системы, системы удаленного взаимодействия и мониторинга, телемедицина, интернет-медицина;
- терапевтические и диагностические системы, импланты, искусственные органы, биомедицинские датчики, медицинское оборудование, медицинский интернет вещей;
- математическое моделирование биофизических процессов.

Одним из приоритетных направлений работы Конференции ИТНТ-2023 является образовательный аспект, заключающийся в предоставлении студентам и молодым ученым возможности ознакомиться с новейшими научными достижениями по [тематикам](#) Конференции, а также с уникальным научным оборудованием и лабораторной базой [Самарского университета](#), используемыми для реализации современных научных проектов.

В рамках Конференции проводится [Молодежная школа](#), где молодые ученые и студенты получают возможность повысить свой профессиональный уровень и опубликовать свои научные результаты.

В данный сборник трудов вошли материалы по 6 основным направлениям Конференции:

Том 1. Компьютерная оптика и нанофотоника (под редакцией к.ф.-м.н. Е.С. Козловой)

Том 2. Информационные технологии дистанционного зондирования Земли (под редакцией к.т.н. И.С. Ткаченко)

Том 3. Распознавание, обработка и анализ изображений (под редакцией д.т.н. В.В. Сергеева)

Том 4. Искусственный интеллект (под редакцией д.т.н. А.В. Никонорова)

Том 5. Науки о данных (под редакцией к.т.н. Е.В. Гошина).

Том 6. Информационные технологии в биомедицине (под редакцией д.ф.-м.н. В.П. Захарова).

Выпускающий редактор томов 1-6: В.Д. Зайцев.

Официальный сайт Конференции ИТНТ-2023: <http://itnt-conf.org/>

# ОРГАНИЗАТОРЫ

- Институт систем обработки изображений РАН (ИСОИ РАН) – филиал ФНИЦ «Кристаллография и фотоника» РАН, г. Самара, Россия;
- Самарский национальный исследовательский университет имени академика С.П. Королева (Самарский университет), г. Самара, Россия.

## *Организационный комитет*

### Председатель

Богатырёв В.Д. – д.э.н., профессор, ректор Самарского университета имени академика С.П. Королева, г. Самара, Россия.

### Заместители председателя

Казанский Н.Л. – д.ф.-м.н., проф., руководитель ИСОИ РАН, г. Самара, Россия;

Сергеев В.В. – д.т.н., проф., заведующий кафедрой геоинформатики и информационной безопасности Самарского университета имени академика С.П. Королева, г. Самара, Россия;

Куприянов А.В. – д.т.н., доцент, исполнительный директор института информатики и кибернетики Самарского университета имени академика С.П. Королева, г. Самара, Россия.

### Ответственный секретарь

Христофорова Ю.А. – к.ф.-м.н., Самарский университет имени академика С.П. Королева, г. Самара, Россия.

### Члены Организационного комитета

Архипова Д.В. – Самарский университет имени академика С.П. Королева, г. Самара, Россия;

Батаева Е.М. – Самарский университет имени академика С.П. Королева, г. Самара, Россия;

Бояркин Ю.Н. – ИСОИ РАН, г. Самара, Россия;

Гашников М.В. – к.т.н., Самарский университет имени академика С.П. Королева, г. Самара, Россия;

Душанина И.И. – Самарский университет имени академика С.П. Королева, г. Самара, Россия;

Еленев Д.В. – Ph.D, к.т.н., доцент, Самарский университет имени академика С.П. Королева, г. Самара, Россия;

Ильасова Н.Ю. – д.т.н., доцент, ИСОИ РАН, г. Самара, Россия;

Кадомина Е.А. – Самарский университет имени академика С.П. Королева, г. Самара, Россия;

Кириш Д.В. – к.т.н., Самарский университет имени академика С.П. Королева, г. Самара, Россия;

Леонова К.С. – Самарский университет имени академика С.П. Королева, г. Самара, Россия;

Логанова Л.В. – к.т.н., Самарский университет имени академика С.П. Королева, г. Самара, Россия;

Максимов А.И. – Самарский университет имени академика С.П. Королева, г. Самара, Россия;

Маркушин М.А. – Самарский университет имени академика С.П. Королева, г. Самара, Россия;

Матвеева И.А. – Самарский университет имени академика С.П. Королева, г. Самара, Россия;

Мисневич С.К. – Самарский университет имени академика С.П. Королева, г. Самара, Россия;

Паренский Н.А. – Самарский университет имени академика С.П. Королева, г. Самара, Россия;

Пашков Д.Е. – к.э.н., доцент, Самарский университет имени академика С.П. Королева, г. Самара, Россия;

Подлипов В.В. – Самарский университет имени академика С.П. Королева, г. Самара, Россия;

Попов С.Б. – д.т.н., проф., ИСОИ РАН, г. Самара, Россия;

Пресняков К.Г. – Департамент информационных технологий и связи Самарской области, г. Самара, Россия;

Сорокина Е.В. – Самарский университет имени академика С.П. Королева, г. Самара, Россия;

Стафеев С.С. – к.ф.-м.н., ИСОИ РАН, г. Самара, Россия;

Татарина С.С. – Самарский университет имени академика С.П. Королева, г. Самара, Россия;

Тиц С.Н. – к.т.н., Самарский университет имени академика С.П. Королева, г. Самара, Россия;

Фомченков С.А. – ИСОИ РАН, г. Самара, Россия;

Хасаев Г.Р. – д.э.н., проф., Самарский университет имени академика С.П. Королева, г. Самара, Россия;

Хнырева Е.С. – Самарский университет имени академика С.П. Королева, г. Самара, Россия;

Яшина В.В. – к.ф.-м.н., Федеральное государственное учреждение «Федеральный исследовательский центр «Информатика и управление» Российской академии наук».

## *Программный комитет*

### Председатель

Сойфер В.А. – академик РАН, д.т.н., проф., Самарский университет имени академика С.П. Королева, г. Самара, Россия.

### Заместитель председателя

Казанский Н.Л. – д.ф.-м.н., профессор, ИСОИ РАН, г. Самара, Россия.

Члены Программного комитета

Джалем К. – Ph.D, Центральный университет Джаркханда, г. Ранчи, Джаркханд, Индия;  
Магрупов Т.М. – д.т.н., проф., Ташкентский государственный технический университет имени Ислама Каримова, г. Ташкент, Узбекистан;  
Недзьведь А.М. – д.т.н., доцент, Белорусский государственный университет, г. Минск, Белоруссия;  
Недзьведь О.В. – к.т.н., доцент, Белорусский государственный университет, г. Минск, Белоруссия;  
О’Фаолеин Л. – проф., Национальный институт Гиндаля, г. Корк, Ирландия;  
Паскали М.А. – проф., Институт информационных наук и технологий «А. Фаэдо» (ИСТИ) Итальянского национального исследовательского совета, г. Пиза, Италия;  
Саболевски М. – проф., Польско-японская академия информационных технологий, г. Варшава, Польша;  
Сажин С. – проф., Университет Брайтона, г. Брайтон, Великобритания;  
Фан Б. – проф., Институт оптики и электроники Китайской академии наук, г. Сычуань, Китай;  
Жанг Л. – проф., Шаньдунский научно-технологический университет, г. Циндао, Китай;  
Бычков И.В. – академик РАН, д.т.н., проф., Институт динамики систем и теории управления имени В.М. Матросова Сибирского отделения РАН, г. Иркутск, Россия;  
Воеводин В.В. – чл.-корр. РАН, д.ф.-м.н., проф., Московский государственный университет имени М.В. Ломоносова, г. Москва, Россия;  
Головашкин Д.Л. – д.ф.-м.н., проф., Институт систем обработки изображений РАН – филиал ФНИЦ «Кристаллография и фотоника» РАН, г. Самара, Россия;  
Гошин Е.В. – к.т.н., доцент, Самарский национальный исследовательский университет имени академика С.П. Королева, г. Самара, Россия;  
Гуляев Ю.В. – академик РАН, д.ф.-м.н., проф., Институт радиотехники и электроники имени В.А. Котельникова РАН, г. Москва, Россия;  
Желтов С.Ю. – академик РАН, д.т.н., проф., ГосНИИ авиационных систем, г. Москва, Россия;  
Захаров В.П. – д.ф.-м.н., проф., Самарский национальный исследовательский университет имени академика С.П. Королева, г. Самара, Россия;  
Ильсова Н.Ю. – д.т.н., доцент, Институт систем обработки изображений РАН – филиал ФНИЦ «Кристаллография и фотоника» РАН, г. Самара, Россия;  
Калошин В.А. – д.ф.-м.н., проф., Институт радиотехники и электроники имени В.А. Котельникова РАН, г. Москва, Россия;  
Карпов О.Э. – академик РАН, д.м.н., проф., Национальный медико-хирургический Центр имени Н.И. Пирогова, г. Москва, Россия;  
Кистенев Ю.В. – д.ф.-м.н., проф., Томский государственный университет, г. Томск, Россия;  
Козлова Е.С. – к.ф.-м.н., Институт систем обработки изображений РАН – филиал ФНИЦ «Кристаллография и фотоника» РАН, г. Самара, Россия;  
Конов В.И. – академик РАН, д.ф.-м.н., проф., Институт общей физики имени А.М. Прохорова РАН, г. Москва, Россия;  
Котляр В.В. – д.ф.-м.н. проф., Институт систем обработки изображений РАН – филиал ФНИЦ «Кристаллография и фотоника» РАН, г. Самара, Россия;  
Кульчин Ю.Н. – академик РАН, д.ф.-м.н., проф., Институт автоматики и процессов управления Дальневосточного отделения РАН, г. Владивосток, Россия;  
Куприянов А.В. – д.т.н., доцент, Самарский национальный исследовательский университет имени академика С.П. Королева, г. Самара, Россия;  
Лабунец В.Г. – д.т.н., проф., Уральский государственный лесотехнический университет, г. Екатеринбург, Россия;  
Лупян Е.А. – д.т.н., проф., Институт космических исследований РАН, г. Москва, Россия;  
Мясников В.В. – д.ф.-м.н., проф., Самарский национальный исследовательский университет имени академика С.П. Королева, г. Самара, Россия;  
Немирко А.П. – д.т.н., проф., Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина), г. Санкт-Петербург, Россия;  
Никитов С.А. – академик РАН, д.ф.-м.н., проф., Институт радиотехники и электроники имени В.А. Котельникова РАН, г. Москва, Россия;  
Николаев Д.П. – к.ф.-м.н., Институт проблем передачи информации имени А.А. Харкевича РАН, г. Москва, Россия;  
Никоноров А.В. – д.т.н., Институт систем обработки изображений РАН – филиал ФНИЦ «Кристаллография и фотоника» РАН, г. Самара, Россия;  
Новиков Д.А. – академик РАН, д.т.н., проф., Институт проблем управления РАН, г. Москва, Россия;  
Потатуркин О.И. – д.т.н., проф., Институт автоматики и электрометрии Сибирского отделения Российской академии наук, г. Новосибирск, Россия;

Сергеев В.В. – д.т.н., проф., Самарский национальный исследовательский университет имени академика С.П. Королева, г. Самара, Россия;

Соколов И.А. – академик РАН, д.т.н., проф., Федеральный исследовательский центр «Информатика и управление» РАН;

Ткаченко И.С. – к.т.н., доцент, Самарский национальный исследовательский университет имени академика С.П. Королева, г. Самара, Россия;

Тучин В.В. – член-корр. РАН, д.ф.-м.н., проф., Саратовский национальный исследовательский государственный университет им. Н.Г. Чернышевского, г. Саратов, Россия;

Хонина С.Н. – д.ф.-м.н., проф., Институт систем обработки изображений РАН – филиал ФНИЦ «Кристаллография и фотоника» РАН, г. Самара, Россия;

Чочиа П.А. – д.т.н., Институт проблем передачи информации имени А.А. Харкевича РАН, г. Москва, Россия;

Юлдашев З.М. – д.т.н., проф., Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина), г. Санкт-Петербург, Россия.

# Вторые моменты очередей в системах массового обслуживания с групповыми пуассоновскими потоками

Б.Я. Лихтциндер  
Поволжский университет  
телекоммуникаций и информатики,  
Самара, Россия  
lixht@psuti.ru

В. И. Моисеев  
Пермский государственный  
национальный исследовательский  
университет,  
Пермь, Россия.  
vim@psu.ru

А.Ю. Привалов  
Самарский национальный  
исследовательский университет им.  
академика С.П. Королева  
Самара, Россия  
privalov1967@gmail.com

**Аннотация** — Рассматривается применение интервальных методов анализа очередей в системах массового обслуживания с групповыми пуассоновскими потоками. Получены соотношения для вторых моментов размеров очередей в системах массового обслуживания с групповыми пуассоновскими потоками. Показано, что они зависят от третьего центрального момента числа заявок, поступающих в течение интервалов времени обработки одной заявки.

**Ключевые слова** — групповые пуассоновские потоки, системы массового обслуживания, вторые моменты, коэффициент загрузки, очереди.

## 1. ВВЕДЕНИЕ

Одной из популярных моделей пакетного трафика, сочетающего в себе простоту анализа, свойственную классическим Пуассоновским моделям и возможность учёта пачечного характера современного пакетного трафика являются неординарные потоки Пуассона. Они являются альтернативой моделям, учитывающим фрактальные свойства потоков, которые, в виду весьма высокой сложности, нашли ограниченное применение на практике. Этапы развития указанных моделей представлены в обзоре [1], а самые популярные их виды рассмотрены в [2, 7-9].

К такого рода моделям относится и неординарный пуассоновский поток событий. В нем выполняются свойства стационарности и отсутствия последствия, но не выполняется свойство ординарности. Такой поток называют пуассоновским неординарным (групповым) потоком независимых событий [5]. Глубокое исследование статистических характеристик таких потоков в системах телекоммуникаций представляется весьма актуальной задачей.

Известно, что среднее значение числа заявок  $M(A(\tau)) = \lambda\tau M(B)$  указанного потока, поступающих в течение некоторого интервала  $\tau$ , пропорционально длительности этого интервала. Если интервал является временем обработки заявки в СМО, то  $M(A(\tau)) = \lambda\tau M(B) = \rho$ . Дисперсия  $D(A(\tau))$  также пропорциональна коэффициенту загрузки  $\rho$ :

$$D(A(\tau)) = \lambda\tau M(B^2) = \rho M(B)(1+v^2) = E_{B\rho}$$

Здесь,  $B$  – случайная величина, равная размеру пачки (в числе заявок), а  $v^2 = D(B)/(M(B))^2$  – квадрат коэффициента вариации чисел заявок в пачках.

Разработанный нами интервальный метод (см., например, [4,6]) будет использоваться и здесь, и позволит относительно просто получить вторые моменты очереди в одноканальной СМО (для группового пуассоновского потока результат обладает

научной новизной). Здесь в обобщенной нами формуле Хинчина-Поллачека из [4,6], определяющей среднее значение размера очереди, будет отсутствовать элемент, учитывающий корреляционную зависимость, и формула будет иметь вид:

$$M(Q) = D(A(\tau))/(2(1-\rho)) - \rho/2 = E_{B\rho}/(2(1-\rho)) - \rho/2 \quad (1)$$

Дисперсия размера пачки и коэффициент загрузки полностью определяют средний размер очереди одноканальной СМО. В частном случае ординарного пуассоновского потока:  $B=1$  (константа),  $v^2=0$ ,  $E_B=1$  получается известная формула:

$$M(Q) = \rho^2/(2(1-\rho)).$$

Практическое значение представленных в докладе формул для первых двух моментов очереди в СМО с групповым пуассоновским потоком состоит в том, что они позволяют гораздо точнее моделировать такими потоками реальный трафик, подбирая распределение размера пачки пакетов, прибывающих одновременно.

## 2. ВТОРОЙ НАЧАЛЬНЫЙ МОМЕНТ РАЗМЕРОВ ОЧЕРЕДЕЙ

Найдём второй начальный момент размера очереди  $M(q^2)$ , для чего воспользуемся уравнением баланса [4]:

$$Q_i = Q_{i-1} + A_i - \delta_i \quad (2)$$

где  $\delta_i=0$ , если  $Q_{i-1} = A_i = 0$  и  $\delta_i=1$  в противном случае,  $Q_i$  – значение очереди на  $i$ -м интервале времени обработки одной заявки,  $A_i$  – число заявок, поступивших, а  $\delta_i$  – число заявок, обработанных в течение указанного интервала времени.

Необходимо учитывать, что при заданных ограничениях

$$(\delta_i)^k = \delta_i, A_i\delta_i = A_i, Q_{i-1}\delta_i = Q_{i-1} \text{ и } M(A_i) = M(\delta_i) \quad (3)$$

Возведем обе части уравнения (2) в третью степень, и после некоторых несложных преобразований с учётом принятых ограничений получим:

$$(Q_i)^3 = (Q_{i-1})^3 + 3(Q_{i-1})^2 A_i - 3(Q_{i-1})^2 + 3Q_{i-1}((A_i)^2 - 2A_i + 1) + (A_i)^3 - 3(A_i)^2 + 3A_i - \delta_i.$$

Произведем усреднение обеих частей с учётом стационарного состояния системы при стремлении времени к бесконечности, когда  $M((Q_i)^3) = M((Q_{i-1})^3)$ , и после некоторых преобразований, получим:

$$3 M((Q_{i-1})^2) (1 - M(A_i)) = 3 M(Q_{i-1}) D(A_i) + 3 M(Q_{i-1})(1 - M(A_i))^2 + M((A_i)^3) - 3M(A_i)^2 + 2M(A_i).$$

Выразим отсюда  $\mathbf{M}((Q_{i-1})^2)$ :

$$\mathbf{M}((Q_{i-1})^2) = \mathbf{M}(Q_{i-1}) (1 - \mathbf{M}(A_i))^2 / (1 - \mathbf{M}(A_i)) + \\ + (\mathbf{M}((A_i)^3) - 3\mathbf{M}(A_i)^2 + 2\mathbf{M}(A_i)) / (3(1 - \mathbf{M}(A_i))).$$

В дальнейшем нам удобнее будет пользоваться не начальным третьим моментом  $\mathbf{M}((A_i)^3)$ , а центральным  $\mu_3(A_i) = \mathbf{M}(A_i - \mathbf{M}(A_i))^3$ , при этом несложно показать, что

$$\mathbf{M}((A_i)^3) = \mu_3(A_i) + 3\mathbf{D}(A_i)\mathbf{M}(A_i) + (\mathbf{M}(A_i))^3,$$

откуда можно получить, что

$$\mathbf{M}((Q_{i-1})^2) = \mathbf{M}(Q_{i-1}) (1 - \mathbf{M}(A_i))^2 / (1 - \mathbf{M}(A_i)) + \\ + ((\mathbf{M}(A_i))^3 + 3\mathbf{D}(A_i)\mathbf{M}(A_i) - 3(\mathbf{M}(A_i))^2 + 2\mathbf{M}(A_i)) / (3(1 - \mathbf{M}(A_i))) + \\ + \mu_3(A_i) / (3(1 - \mathbf{M}(A_i))).$$

При устремлении времени к бесконечности индекс  $i-1$  можно опустить, и воспользовавшись тем, что  $\mathbf{M}(A_i) = \rho$ , а также (1), после некоторых преобразований, получить:

$$\mathbf{M}(Q^2) = (\mathbf{D}(A(\tau)) - \rho(1 - \rho))(\mathbf{D}(A(\tau)) + (1 - \rho)^2) / (2(1 - \rho)^2) + \\ + (\rho^3 + 3\mathbf{D}(A(\tau))\rho - 3\mathbf{D}(A(\tau)) - 3\rho^2 + 2\rho) / (3(1 - \rho)) + \\ + \mu_3(A(\tau)) / (3(1 - \rho)).$$

Найти  $\mu_3(A(\tau))$  можно методом производящих функций, используя для производящей функции  $G_A(z)$  с.в.  $A(\tau)$  её очевидное выражение через  $G_B(z)$  – производящую функцию размера пачки:

$$G_A(z) = \exp(-\lambda) \sum ((G_B(z))^n \lambda^n / n! = \exp(\lambda(G_B(z) - 1)).$$

Найдя отсюда выражение трёх первых факториальных моментов  $A(\tau)$  через факториальные моменты  $B$ , и далее, по известным формулам, подставив их в выражение третьего центрального момента через факториальные, можно получить окончательный результат:

$$\mu_3(A(\tau)) = \lambda\tau(\mathbf{M}(B^3)).$$

Итак, второй начальный момент размера очереди в СМО с групповыми пуассоновскими потоками определяется соотношением:

$$\mathbf{M}(Q^2) = (E_B \rho - \rho(1 - \rho))(E_B \rho + (1 - \rho)^2) / (2(1 - \rho)^2) + \\ + (\rho^3 + 3E_B \rho^2 - 3E_B \rho - 3\rho^2 + 2\rho) / (3(1 - \rho)) + \\ + \mu_3(A(\tau)) / (3(1 - \rho)). \quad (4)$$

В частном случае, для простейшего пуассоновского потока, где  $E_B = 1$ ,  $\mu_3(A(\tau)) = \rho$  (т.к.  $\mathbf{M}(B^3) = \mathbf{M}(B) = 1$ ), выражение упростится:

$$\mathbf{M}(Q^2) = (1 - \rho/3 + \rho^2/3) \rho^2 / (2(1 - \rho)^2).$$

Оно показывает, что второй начальный момент очереди пуассоновского потока определяется исключительно значением коэффициента загрузки.

### 3. ДИСПЕРСИЯ РАЗМЕРОВ ОЧЕРЕДЕЙ

Дисперсию размеров очередей определим на основании известного соотношения  $\mathbf{D}(Q) = \mathbf{M}(Q^2) - (\mathbf{M}(Q))^2$ , подставив туда (4) и (2). После некоторых преобразований, получим

$$\mathbf{D}(Q) = (E_{B\rho} - \rho(1 - \rho))(E_{B\rho} + 2 - 3\rho + \rho^2) / (4(1 - \rho)^2) + \\ + (\rho^3 + 3E_B \rho^2 - 3E_{B\rho} - 3\rho^2 + 2\rho) / (3(1 - \rho)) + \\ + \mu_3(A(\tau)) / (3(1 - \rho)).$$

В частном случае, для простейшего пуассоновского потока

$$\mathbf{D}(Q) = (1 - \rho/3 - \rho^2/6) \rho^2 / (2(1 - \rho)^2).$$

Так же, как и второй начальный момент, дисперсия очереди простейшего потока полностью определяется коэффициентом загрузки.

### 4. ЗАКЛЮЧЕНИЕ

Представлены формулы для вторых моментов очереди в одноканальной СМО с групповым пуассоновским потоком на входе.

Дисперсия очередей групповых пуассоновских потоков зависит от третьего центрального момента, который характеризует симметричность закона распределения размеров пачек заявок. Два различных групповых потока, имеющие одинаковые зависимости средних значений очередей от коэффициента загрузки, имеют различные дисперсии очередей, разность которых пропорциональна разности их третьих моментов.

### ЛИТЕРАТУРА

- [1] Вишнеvский, В.М. Системы массового обслуживания с коррелированными входными потоками и их применение для моделирования телекоммуникационных сетей / В.М. Вишнеvский, А.Н. Дудин // Автоматика и телемеханика. – 2017. – Т. 8. – С. 3–59.
- [2] Neuts, M.F. Versatile Markovian point process // Journal of Applied Probability. – 1979. – Vol. 16(4). – P. 764-779. DOI: <https://doi.org/10.2307/3213143>.
- [3] Дудин, А.Н. Системы массового обслуживания с коррелированными потоками / А.Н. Дудин, В.И. Клименок – Минск: БГУ, 2000. – 175 с.
- [4] Лихтциндер, Б.Я. Трафик мультисервисных сетей доступа (интервальный анализ и проектирование) // Б.Я. Лихтциндер. – М.: Горячая линия - Телеком, 2018. – 290 с.
- [5] Likhhtsinder, B. Ya. Models of group Poisson flows in telcommunications traffic control / B. Ya. Likhhtsinder, Yu. O. Bakay // Вестник Самарского государственного технического университета. Серия: Технические науки. – 2020. – Т. 28, № 3. – С. 75-89.
- [6] Likhhtsinder, B. Ya. Queue Analysis for Video Traffic Using the Generalized Interval Method / B. Ya. Likhhtsinder, E.V. Kitaeva, A. Yu. Privalov // 2022 VIII International Conference on Information Technology and Nanotechnology (ITNT). – 2022. – P. 1-4.
- [7] Lakatos, L. Introduction to Queueing Systems with Telecommunication Applications / L. Lakatos, L. Szeidl, M. Telek. – Springer Science+Business Media, 2013. – 388 p. DOI: <https://doi.org/10.1007/978-1-4614-5317-8>.
- [8] Klimenok, V.I. Retrial BMAP/PH/N Queueing System with a Threshold-Dependent Inter-Retrial Time Distribution / V. I. Klimenok, A.N. Dudin, V. M. Vishnevsky and O.V. Semenova // Mathematics. – 2022. – Vol.10(2). – P. 269. <https://doi.org/10.3390/math10020269>
- [9] Vishnevsky, V. Analysis of a MAP/M/1/N Queue with Periodic and Non-Periodic Piecewise Constant Input Rate / V. Vishnevsky, K. Vytovtov, E. Barabanova, O. Semenova // Mathematics. – 2022. – Vol.10(10). – P. 1684. <https://doi.org/10.3390/math101016840>

# Эффективная Распределенная Обработка Больших Данных на Основе Наименьшего Информационного Пространства

П.В. Голубцов

Московский государственный университет имени М. В. Ломоносова  
Москва, Россия  
golubtsov@physics.msu.ru

**Аннотация**—Рассматривается алгебраическая формализация распределенной обработки больших данных. Определяется понятие информационного пространства для заданной процедуры обработки данных и доказывается существование наименьшего информационного пространства, обеспечивающего самую компактную форму накопления информации и позволяющего наиболее эффективно распараллелить обработку. Показано, что в терминах информационного пространства естественным образом выражаются понятия сложения информации и качества информации.

**Ключевые слова**— большие данные, параллельная обработка, алгебра информации, качество информации, MapReduce

## 1. ВВЕДЕНИЕ

Данные в современных исследованиях нередко имеют огромный объем, распределены между многочисленными сайтами и постоянно пополняются. В результате собрать все относящиеся к исследованию данные на одном компьютере, как правило, невозможно и непрактично, поскольку один компьютер не сможет обработать их в разумные сроки. Подходящий алгоритм анализа данных должен, параллельно работая в распределенной системе, извлекать из каждого набора исходных данных некоторую промежуточную компактную информацию, постепенно объединять ее и, наконец, использовать накопленную информацию для получения результата.

В предыдущих работах автора (напр., [1]) были рассмотрены конкретные типы задач обработки данных и исследованы возникающие в них специальные виды представления информации, содержащейся в данных. Было показано, что для эффективной обработки распределенных данных ключевую роль играет возможность введения специальной промежуточной формы представления информации, обладающей определенными алгебраическими свойствами. В рассмотренных задачах были введены соответствующие информационные пространства и исследованы их свойства.

Данная работа призвана подвести общий фундамент под эти исследования путем построения алгебраической формализации распределенной обработки данных. Определяется понятие информационного пространства для заданной процедуры обработки и, в частности, наименьшего информационного пространства, предоставляющего максимально компактную форму представления информации и, как следствие, позволяющего наиболее эффективно распараллелить обработку данных. При этом в терминах информационного пространства естественным образом выражаются бинарная операция сложения фрагментов

информации и упорядочение, отражающее понятие качества информации.

Следует отметить, что существует довольно много подходов к понятию информация, например, комбинаторный, вероятностный, алгоритмический [2], однако все они определяют меру количества информации в том или ином контексте. Напротив, наименьшее информационное пространство приводит к понятию именно информации, содержащейся в данных, как максимально компактное представление набора данных, обеспечивающее тот же результат обработки что и этот набор. В результате, информация, извлеченная из данных, полностью заменяет эти данные.

## 2. ПРОЦЕДУРА ОБРАБОТКИ И ИНФОРМАЦИОННЫЕ ПРОСТРАНСТВА

Пусть  $D$  – множество возможных значений входных данных, а  $R$  – множество значений результатов обработки. В задачах больших данных на вход процедуры обработки поступают наборы элементов из  $D$ , причем эти наборы могут быть распределены по многим компьютерам. Для математического представления множества всех таких наборов с операцией их слияния обычно используется свободный моноид  $D^*$  с операцией конкатенации. Однако, поскольку результат обработки как правило не должен зависеть от порядка поступления данных, удобно представлять пространство всевозможных наборов исходных данных свободным коммутативным моноидом  $D^+$  с множеством образующих  $D$ . Его элементами являются конечные мультимножества на множестве  $D$  (в которых элемент может повторяться несколько раз) с операцией сложения мультимножеств (при которой кратности одинаковых элементов складываются).

**Определение.** *Процедурой обработки* с наборами данных из множества данных  $D$  и результатами из множества  $R$  будем называть отображение  $p$  из свободного коммутативного моноида  $D^+$  в множество  $R$ , т.е.  $p: D^+ \rightarrow R$ .

**Определение.** *Информационное пространство* (ИП)  $(U, q, r)$  для процедуры  $p: D^+ \rightarrow R$  это коммутативный моноид  $U$ , сюръективный гомоморфизм (СГ) моноидов  $q: D^+ \rightarrow U$  и отображение  $r: U \rightarrow R$  такие, что  $r \circ q = p$ .

Фактически, гомоморфизм  $q$  сжимает исходные данные без потери информации, представляя различные наборы данных одним и тем же элементом. Его гомоморфность означает, что объединению наборов данных отвечает сумма соответствующих фрагментов информации, а его сюръективность обеспечивает отсутствие в  $U$  элементов, которые не отвечают никаким наборам данных. Эффект от использования ИП

определяется тем, насколько оно позволяет сжать данные.

### 3. НАИМЕНЬШЕЕ ИНФОРМАЦИОННОЕ ПРОСТРАНСТВО

**Определение.** Будем говорить, что ИП  $(U, q, r)$  меньше, чем  $(U', q', r')$  и обозначать это как  $(U, q, r) \ll (U', q', r')$ , если существует такое отображение  $h: U' \rightarrow U$ , что  $h \circ q' = q$ .

Поскольку  $q'$  – СГ, такое преобразование информационных пространств  $h$  единственно и, т.к.  $q$  – СГ, также является СГ. При этом  $r \circ h = r'$ , т.е.  $(U, h, r)$  можно рассматривать как ИП для процедуры  $r': U' \rightarrow R$ . Отношение  $\ll$  является предпорядком, причем если  $U' \ll U$  и  $U \ll U'$ , то эти ИП изоморфны. Наименьшее в смысле этого упорядочения ИП  $(U, q, r)$  обладает тем свойством, что любое ИП  $(U', q', r')$  для  $p$  факторизуется через него, т.е. существует (единственный) СГ  $h: U' \rightarrow U$  для которого  $h \circ q' = q$  и  $r' = r \circ h$ .

**Теорема (Существование).** Наименьшее ИП для процедуры  $p: D^+ \rightarrow R$  существует и единственно с точностью до изоморфизма.

Для исследования структуры наименьшего ИП дадим следующее

**Определение.** Пусть  $U$  – коммутативный моноид. Будем говорить, что элементы  $x$  и  $y$  из  $U$  неразличимы относительно  $r: U \rightarrow R$  и обозначать  $x \sim_r y$ , если

$$\forall z \in U \quad r(x + z) = r(y + z).$$

**Теорема (Конструкция).** ИП  $(D^+ / \sim_p, q, r)$  является наименьшим ИП для процедуры  $p: D^+ \rightarrow R$ . Здесь  $D^+ / \sim_p$  – фактормоноид по конгруэнции неразличимости на  $D^+$  относительно  $p$ , гомоморфизм  $q: D^+ \rightarrow D^+ / \sim_p$  – соответствующий канонический эпиморфизм,  $q(x) = [x]_{\sim_p}$  для  $x \in D^+$ , а отображение  $r: D^+ \rightarrow R$  определяется как  $r([x]_{\sim_p}) = p(x)$  для  $x \in D^+$ .

В практических задачах (см., напр. [1]) анализ процедуры обработки нередко позволяет предложить естественный вариант ИП. Следующее утверждение дает критерий проверки того, что ИП является наименьшим.

**Теорема (Критерий).** ИП  $(U, q, r)$  является наименьшим если все его элементы различимы относительно  $r: U \rightarrow R$ .

### 4. КАЧЕСТВО ИНФОРМАЦИИ

Алгебраическая структура ИП позволяет естественным образом определить упорядочение, характеризующее качество информации.

**Определение.** Для элементов  $x$  и  $y$  из ИП  $U$  будем говорить, что  $x$  содержит больше информации, чем  $y$  и обозначать  $x \geq y$  если

$$\exists z \in U \quad x = y + z.$$

Отношение  $\geq$  на ИП  $U$  является отношением предпорядка, согласованным с алгебраической структурой, т.е.,  $x' \geq x \wedge y' \geq y \Rightarrow x' + y' \geq x + y$  и  $x \geq 0$ . Более того, преобразование ИП  $h: U' \rightarrow U$  сохраняет упорядочение качества:  $x \geq y \Rightarrow h(x) \geq h(y)$ .

### 5. НАКОПЛЕНИЕ ИНФОРМАЦИИ В MAPREDUCE

Использование наименьшего ИП позволяет максимально эффективно распараллеливать процесс

накопления информации в рамках модели распределенного анализа данных MapReduce [3] и организовать эффективную обработку без необходимости передачи и накопления самих исходных данных. В контексте этой модели Map преобразует наборы исходных данных в элементы ИП путем применения отображения  $q$ , а Reduce складывает все эти фрагменты частичной информации в один элемент, представляющий все исходные данные, Рис. 1.

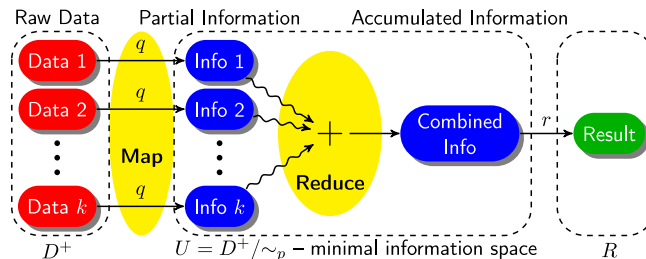


Рис. 1. Параллельная обработка распределенных данных с использованием наименьшего информационного пространства в модели MapReduce

При этом наименьшее информационное пространство определяет наиболее эффективную математическую структуру для представления информации, содержащейся в данных, и описывает «теоретический предел» компактности представления информации.

### 6. ЗАКЛЮЧЕНИЕ

Как показано в этой работе, проблема оптимизации распределенной обработки данных приводит к математическому представлению информации, содержащейся в данных, как элементу наименьшего ИП. При этом в терминах ИП естественным образом выражаются сложение и качество информации.

Понятие информации всегда было предметом преимущественно теоретического интереса. Сейчас проблематика больших данных требуют компактных, эффективных и хорошо организованных форм представления информации. Такие идеальные формы могут отражать самую суть информации, содержащейся в данных. Поэтому изучение таких форм и их свойств может приблизить нас к адекватному математическому описанию самого понятия информации.

### БЛАГОДАРНОСТИ

Работа выполнена при финансовой поддержке РФФИ, грант № 19-29-09044.

### ЛИТЕРАТУРА

- [1] Golubtsov, P. Scalability and Parallelization of Sequential Processing: Big Data Demands and Information Algebras / P. Golubtsov // Advances in Intelligent Systems and Computing, Springer. – 2020. – Vol. 1127. – P. 274–298.
- [2] Колмогоров, А.Н. Три подхода к определению понятия “количество информации” / А.Н. Колмогоров // Пробл. передачи информ. – 1965. – Том 1, № 1. – С. 3–11.
- [3] Dean, J. MapReduce: simplified data processing on large clusters / J. Dean, S. Ghemawat // Comm. of the ACM. – 2008. – Vol. 51(1). – P. 107–113.

# Численная идентификация граничных условий в модели реакции-диффузии

Д.В. Галушкина  
Ульяновский государственный  
университет  
Ульяновск, Россия  
smallcranberry@gmail.com

А.Н. Кувшинова  
Ульяновский государственный  
педагогический университет имени  
И.Н. Ульянова  
Ульяновск, Россия  
kuvanulspu@yandex.ru

Ю.В. Цыганова  
Ульяновский государственный  
университет  
Ульяновск, Россия  
tsyganovajv@gmail.com

**Аннотация**—В статье рассматривается задача численной идентификации граничных условий модели реакции-диффузии по данным зашумленных измерений значений искомой функции. Для решения поставленной задачи осуществляется переход от исходной непрерывной модели с уравнением в частных производных к дискретной линейной стохастической системе в пространстве состояний, в которой функции, входящие в граничные условия, представлены в виде неизвестного вектора входных воздействий. К полученной системе применяется рекуррентный алгоритм одновременного оценивания векторов состояния и входных воздействий Гиллейнса – Де-Мора. Приводятся результаты численного эксперимента, подтверждающие практическую применимость предложенного подхода.

**Ключевые слова**—модель реакции-диффузии, дискретные стохастические системы, алгоритмы рекуррентного оценивания

## 1. ВВЕДЕНИЕ

Задачи идентификации параметров математических моделей тепломассопереноса возникают при исследовании физических, химических, биологических и других природных и техногенных процессов. Такие задачи относятся к классу обратных задач, интерес к которым в последнее время заметно вырос. Среди различных типов обратных задач можно выделить граничные обратные задачи, связанные с определением поведения искомой функции на границе области.

Рассмотрим одномерную модель реакции-диффузии, описываемую уравнением (1) с начальным условием (2) и граничными условиями третьего рода (3):

$$\frac{\partial c}{\partial t} = \alpha \frac{\partial^2 c}{\partial x^2} - \beta c, \quad (1)$$

$$c(x, 0) = \varphi(x), \quad (2)$$

$$\begin{cases} \frac{\partial c(a, t)}{\partial x} = \lambda[c(a, t) - f(t)], \\ \frac{\partial c(b, t)}{\partial x} = -\lambda[c(b, t) - g(t)], \end{cases} \quad (3)$$

$$x \in [a, b], t \in [0, T],$$

где  $c(x, t)$  – искомая функция,  $x$  – пространственная координата,  $t$  – время,  $\alpha$  – коэффициент диффузии,  $\beta$  – коэффициент реакции,  $\varphi(x)$ ,  $f(x)$  и  $g(x)$  – заданные функции,  $a$  и  $b$  – границы рассматриваемой области (отрезка).

Рассмотрим задачу определения значений функций  $f(t)$  и  $g(t)$ , входящих в граничные условия (3), по результатам зашумленных измерений значений функции  $c(x, t)$  в отдельных точках рассматриваемого отрезка в последовательные моменты времени.

Одним из актуальных методов решения граничных обратных задач являются методы параметрической

идентификации, основанные на применении рекуррентных алгоритмов дискретной фильтрации [1], [2]. В работах [3] и [4] для идентификации граничных условий модели конвективно-диффузионного переноса было предложено использовать алгоритм Гиллейнса – Де-Мора [5], предназначенный для одновременного оценивания векторов состояния и неизвестных входных воздействий дискретной линейной стохастической системы. Применим данный подход для идентификации граничных условий модели (1)–(3).

## 2. ДИСКРЕТНАЯ ЛИНЕЙНАЯ СТОХАСТИЧЕСКАЯ МОДЕЛЬ

Для решения поставленной задачи перейдем от непрерывной модели (1)–(3) к дискретной линейной стохастической системе в пространстве состояний:

$$\begin{cases} c_k = F_{k-1}c_{k-1} + B_{k-1}u_{k-1}, \\ z_k = H_k c_k + \xi_k, \quad k = 1, \dots, K, \end{cases}$$

где  $c_k$  – вектор состояния,  $u_k$  – вектор входных воздействий,  $z_k$  – вектор измерений,  $\xi_k$  – шум в измерителе (нормально распределенная случайная последовательность с нулевым математическим ожиданием и ковариационной матрицей  $R_k > 0$ ). В данной системе первое уравнение называется уравнением объекта, а второе – уравнением измерений.

Зададим в области  $[a, b] \times [0, T]$  регулярную сетку  $\{(x_i, t_k) \mid i=0, 1, \dots, N, k=0, 1, \dots, K\}$  с пространственным шагом  $\Delta x$  и временным шагом  $\Delta t$ . Обозначим  $c_i^k = c(x_i, t_k)$ ,  $\varphi_i = \varphi(x_i)$ ,  $f^k = f(t_k)$ ,  $g^k = g(t_k)$  и заменим частные производные в уравнении (1) и граничных условиях (3) их конечно-разностными аналогами, тогда уравнение объекта может быть записано следующим образом:

$$\begin{bmatrix} c_0^k \\ c_1^k \\ c_2^k \\ \vdots \\ c_{N-2}^k \\ c_{N-1}^k \\ c_N^k \end{bmatrix} = \begin{bmatrix} a_3 a_1 & a_3 a_2 & a_3 a_1 & \dots & 0 & 0 & 0 \\ a_1 & a_2 & a_1 & \dots & 0 & 0 & 0 \\ 0 & a_1 & a_2 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & a_2 & a_1 & 0 \\ 0 & 0 & 0 & \dots & a_1 & a_2 & a_1 \\ 0 & 0 & 0 & \dots & a_3 a_1 & a_3 a_2 & a_3 a_1 \end{bmatrix} \begin{bmatrix} c_0^{k-1} \\ c_1^{k-1} \\ c_2^{k-1} \\ \vdots \\ c_{N-2}^{k-1} \\ c_{N-1}^{k-1} \\ c_N^{k-1} \end{bmatrix} + \begin{bmatrix} a_4 & 0 \\ 0 & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ 0 & 0 \\ 0 & a_4 \end{bmatrix} \begin{bmatrix} f^k \\ g^k \end{bmatrix}$$

где

$$a_1 = \frac{\alpha \Delta t}{\Delta x^2}, a_2 = 1 - \beta \Delta t - 2 \frac{\alpha \Delta t}{\Delta x^2}, a_3 = \frac{1}{1 + \lambda \Delta x}, a_4 = \frac{\lambda \Delta x}{1 + \lambda \Delta x}.$$

В полученном уравнении объекта функции  $f(t)$  и  $g(t)$  входят в неизвестный вектор входных воздействий.

К уравнению объекта добавим уравнение зашумленных измерений

$$z_k = H_k c_k + \xi_k, \quad k=1, \dots, K.$$

### 3. ПРИМЕР

Пусть требуется идентифицировать граничные условия на левом и правом концах отрезка следующей модели:

$$\begin{aligned} \frac{\partial c}{\partial t} &= \frac{\partial^2 c}{\partial x^2} - 2c, \\ c(x, 0) &= 0, \\ \begin{cases} \frac{\partial c(0, t)}{\partial x} = c(0, t) - t|\sin 4t|, \\ \frac{\partial c(1, t)}{\partial x} = -\left[c(1, t) - \frac{t}{4}\right], \end{cases} \\ x \in [0, 1], t \in [0, 2]. \end{aligned}$$

Процесс идентификации граничных условий будем моделировать в системе MATLAB. Зададим в области  $[0; 1] \times [0; 2]$  плоскости  $Oxt$  пространственно-временную сетку с 6 узлами по оси  $Ox$  и 201 узлом по оси  $Ot$  ( $\Delta x = 0.2$ ,  $\Delta t = 0.01$ ). Решение рассматриваемой задачи получим методом конечных разностей. Вектор состояния дискретной модели будет состоять из 6 узлов. Смоделируем зашумленные измерения с дисперсией шума  $R = 0.02^2$ . Матрицу измерений зададим в виде

$$H = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

На рис. 1 и 2 приведены графики решения задачи и смоделированных зашумленных измерений, а на рис. 3 и 4 – графики оценок левого и правого граничных условий соответственно.

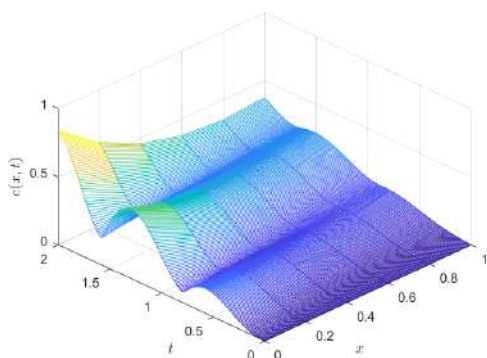


Рис. 1. График решения

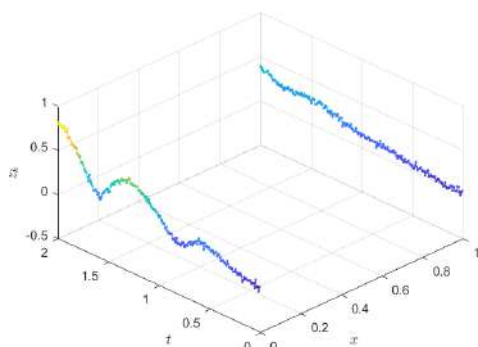


Рис. 2. График зашумленных измерений

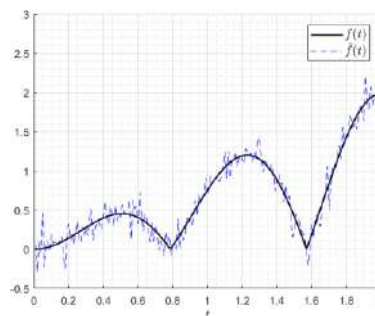


Рис. 3. Графики функции  $f(t)$  и ее оценки

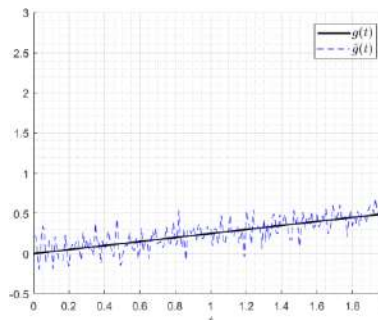


Рис. 4. Графики функции  $g(t)$  и ее оценки

### 4. ЗАКЛЮЧЕНИЕ

В работе рассмотрена задача идентификации граничных условий одномерного уравнения реакции-диффузии с граничными условиями третьего рода по данным зашумленных измерений. Для решения задачи предлагается использовать рекуррентный алгоритм Гиллейнса – Де-Мора для одновременного оценивания векторов состояния и неизвестных входных воздействий дискретной линейной стохастической системы в пространстве состояний. Результаты моделирования показывают работоспособность предложенного подхода.

### БЛАГОДАРНОСТИ

Исследование выполнено за счет гранта Российского научного фонда № 23-21-00361, <https://rscf.ru/project/23-21-00361/>.

### ЛИТЕРАТУРА

- [1] Пилипенко, Н.В. Применение фильтра Калмана в нестационарной теплотерии. Учебное пособие / Н.В. Пилипенко. — СПб.: Университет ИТМО, 2017. — 36 с.
- [2] Копытин, А. В. Применение расширенного фильтра Калмана для идентификации параметров распределенной динамической системы / А.В. Копытин, Е.А. Копытина, М.Г. Матвеев // Вестник ВГУ. Серия: Системный анализ и информационные технологии. — 2018. — Т. 3. — С. 44–50.
- [3] Цыганов, А. В. Динамическая идентификация граничных условий в модели конвективно-диффузионного переноса в условиях зашумленных измерений / А. В. Цыганов, Ю. В. Цыганова, А. Н. Кувшинова // Сборник трудов V международной конференции и молодежной школы «Информационные технологии и нанотехнологии» (ИТНТ-2019). — 2019. — Т. 3. — С. 169–177.
- [4] Кувшинова, А. Н. Динамическая идентификация смешанных граничных условий в модели конвективно-диффузионного переноса в условиях зашумленных измерений / А. Н. Кувшинова // Журнал Средневолжского Математического Общества. — 2019. — Т. 21, № 4. — С. 469–479. — DOI: 10.15507/2079-6900.21.201904.469-479
- [5] Gillijns, S. Unbiased minimum-variance input and state estimation for linear discrete-time systems / S. Gillijns, B. D. Moor // Automatica. — 2007. — Vol. 43. — P. 111–116.

# Идентификация параметров моделей дискретных стохастических систем с мультипликативными и аддитивными шумами

А.В. Цыганов  
Ульяновский государственный  
педагогический университет  
им. И.Н. Ульянова  
Ульяновск, Россия  
andrew.tsyganov@gmail.com

Ю.В. Цыганова  
Ульяновский государственный  
университет  
Ульяновск, Россия  
tsyganovajv@gmail.com

А.В. Голубков  
Ульяновский государственный  
педагогический университет  
им. И.Н. Ульянова  
Ульяновск, Россия  
kr8589@gmail.com

**Аннотация**—В работе рассмотрена задача идентификации параметров моделей дискретных стохастических систем с мультипликативными и аддитивными шумами. Для ее решения построен квадратичный критерий идентификации в форме отрицательной логарифмической функции правдоподобия на основе величин, вычисляемых по алгоритму дискретной линейной фильтрации с учетом мультипликативных шумов в уравнениях состояния и измерения. Для минимизации критерия идентификации применялись методы условной численной оптимизации. Результаты вычислительных экспериментов подтверждают работоспособность предложенного решения.

**Ключевые слова**—дискретные стохастические системы с аддитивными и мультипликативными шумами, параметрическая идентификация, квадратичный критерий идентификации, численные методы оптимизации

## 1. ВВЕДЕНИЕ

Дискретные стохастические системы с аддитивными и мультипликативными шумами рассматриваются при решении ряда практических задач, связанных с обработкой измерительной информации (например, задачи обработки изображений и сигналов, финансовой математики, задачи слежения и др.). Причины появления мультипликативных помех в системе могут иметь различную природу в зависимости от решаемой задачи, моделируемого объекта или процесса, например, это ошибки линеаризации, квантования, физические явления типа фединга и замирания в каналах связи, случайные нарушения в динамике системы или в датчиках, непосредственно сами ошибки моделирования.

Целью данной работы является разработка инструментального метода решения задачи идентификации параметров моделей дискретных стохастических системах с мультипликативными и аддитивными шумами на основе численной минимизации квадратичного критерия идентификации.

## 2. ПОСТАНОВКА ЗАДАЧИ

Рассмотрим инвариантную во времени дискретную линейную стохастическую систему:

$$\begin{cases} x_k = (F + \tilde{F}\xi_{k-1})x_{k-1} + Gw_{k-1}, \\ z_k = (H + \tilde{H}\zeta_k)x_k + v_k, \quad k = 1, 2, \dots, \end{cases} \quad (1)$$

в которой  $w_k \sim \mathcal{N}(0, Q)$  и  $v_k \sim \mathcal{N}(0, R)$  – аддитивные шумы,  $\xi_k \sim \mathcal{N}(0, \sigma_\xi^2)$  и  $\zeta_k \sim \mathcal{N}(0, \sigma_\zeta^2)$  – мультипликативные шумы.

Алгоритмы дискретной фильтрации калмановского типа для рассматриваемых систем известны (см,

например, [1,2,3]). Они позволяют вычислить линейные оптимальные оценки вектора состояния  $x_k$  по доступным измерениям  $z_i, i = 1, \dots, k$ .

Предположим, что матрицы, определяющие уравнения системы (1), зависят от неизвестного параметра  $\theta$ . Поставим задачу их идентификации по доступным измерениям  $z_k$ . Обозначим через  $\theta \in \mathcal{R}^p$  вектор неизвестных параметров. Тогда величина ошибки оценивания  $e_k = x_k - \hat{x}_k$  будет зависеть от значения параметра  $\theta$ , которое задается в уравнениях алгоритма дискретной фильтрации. Минимальное значение ошибки  $e_k$  можно получить при условии минимума по  $\theta$  квадратичного функционала

$$J_k^o(\theta) = \mathbb{E}\{e_k^T(\theta)e_k(\theta)\}. \quad (2)$$

Проблема заключается в том, что функционал (2) не является инструментальным, то есть он не реализуем на практике, поскольку ошибки  $e_k$  недоступны прямому наблюдению. Наиболее популярным подходом к решению данной проблемы являются методы МРЕ (Minimum Prediction Error) [4], основанные на минимизации критерия идентификации, зависящего от наблюдаемой невязки измерений  $v_k = z_k - H\hat{x}_k$ . К таким критериям относятся хорошо известные критерии наименьших квадратов и максимального правдоподобия [5]. Альтернативным подходом является метод ВФК (вспомогательного функционала качества) [6].

Таким образом, алгоритм численной минимизации исходного функционала (2) по параметру  $\theta$  заменяется на алгоритм численной минимизации выбранного инструментального критерия, которые является практически реализуемым.

## 3. МЕТОД ПАРАМЕТРИЧЕСКОЙ ИДЕНТИФИКАЦИИ

Для решения задачи идентификации параметров системы (1) построим инструментальный критерий в форме отрицательной логарифмической функции правдоподобия:

$$J_k(\theta) = \frac{Km}{2} \ln 2\pi + \frac{1}{2} \sum_{k=1}^K \left\{ \ln |B_k(\theta)| + \|v_k(\theta)\|_{B_k^{-1}(\theta)}^2 \right\}, \quad (3)$$

значения которого при заданном  $\theta$  будем вычислять с помощью дискретного фильтра калмановского типа [3].

Вычисление значения параметра  $\theta$  в точке минимума критерия идентификации (3) выполним с помощью численного метода условной минимизации.

#### 4. ЧИСЛЕННЫЙ ПРИМЕР

В качестве примера рассмотрим модель почти равномерного прямолинейного движения объекта с мультипликативными шумами в объекте и измерителе:

$$\begin{cases} x_k = \left( \begin{bmatrix} 1 & \theta \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \xi_{k-1} \right) x_{k-1} + \begin{bmatrix} \theta^2/2 \\ \theta \end{bmatrix} w_{k-1}, \\ z_k = \left( \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \zeta_k \right) x_k + v_k, \end{cases}$$

где  $x_k = [x, v_x]_k^T$ ,  $x$  – координата объекта  $v_x$  – скорость объекта,  $x_0 \sim N([0, 1]^T, 10I_2)$ ,  $w_k \sim N(0, 10^{-2})$ ,  $v_k \sim N(0, I_2)$ ,  $\xi_k \sim N(0, 10^{-4})$ ,  $\zeta_k \sim N(0, 10^{-4})$ ,  $\theta$  – параметр модели, подлежащий идентификации. Положим “истинное” значение параметра равным  $\theta^* = 0.1$ .

С целью подтверждения на практике работоспособности предложенного подхода к решению задачи параметрической идентификации в системе MATLAB проведена серия из 1000 численных экспериментов. В каждом эксперименте выполнена численная идентификация параметра  $\theta$  по результатам 100 измерений. Для численной минимизации критерия (3) выбран метод `fmincon` из библиотеки MATLAB Optimization Toolbox. Поиск решения осуществлялся на отрезке  $[0; 1]$ . Результаты численной идентификации параметра  $\theta$  приведены в таблице 1.

Таблица 1. РЕЗУЛЬТАТЫ ЧИСЛЕННОЙ ИДЕНТИФИКАЦИИ

Mean	RMSE	MAPE
0,0983	0,0112	8,8976

Здесь Mean – среднее значение оценок параметра  $\theta$ , RMSE – корень из среднеквадратической ошибки оценивания, MAPE – средняя абсолютная ошибка в процентах. Данные таблицы 1 показывают, что численная минимизация критерия (3) позволила получить несмещенную оценку модельного параметра  $\theta$  с приемлемой точностью (Mean  $\approx \theta^*$ , MAPE  $\approx 8,9\%$ ).

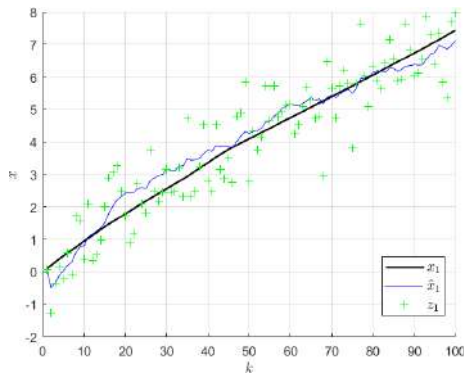


Рис. 1. Графики координаты  $x$ , ее оценки и измерений

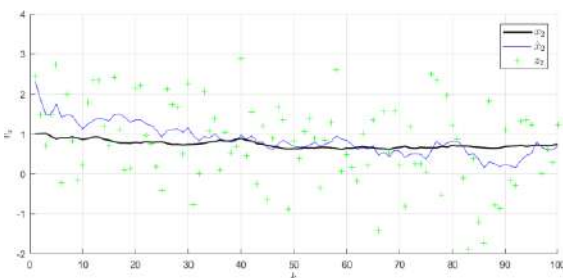


Рис. 2. Графики скорости  $v_x$ , ее оценки и измерений

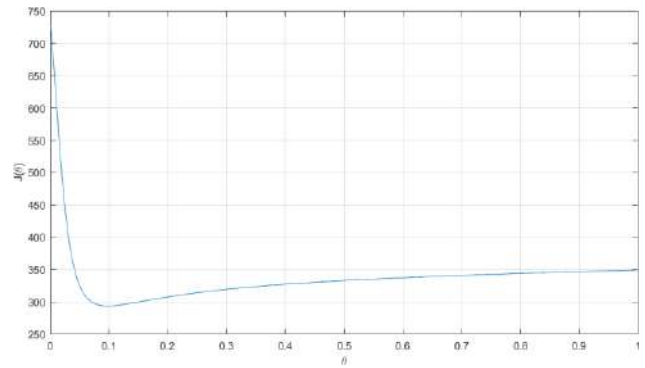


Рис. 3. График критерия идентификации

#### 5. ЗАКЛЮЧЕНИЕ

В работе предложен инструментальный метод идентификации параметров моделей дискретных стохастических систем с мультипликативными и аддитивными шумами. Построен квадратичный критерий идентификации (3) в форме отрицательной логарифмической функции правдоподобия на основе величин, вычисляемых по алгоритму дискретной линейной фильтрации с учетом мультипликативных шумов в уравнениях состояния и измерения. Минимизация критерия идентификации выполнена методом условной численной оптимизации `fmincon` системы MATLAB. На численном примере показана работоспособность предложенного решения.

#### БЛАГОДАРНОСТИ

Исследование выполнено за счет гранта Российского научного фонда № 22–21–00387, <https://rscf.ru/project/22-21-00387/>.

#### ЛИТЕРАТУРА

- [1] Yang, F. Robust Kalman filtering for discrete time-varying uncertain systems with multiplicative noises / F. Yang, Z. Wang, Y. Hung // IEEE Trans. Automat. Contr. – 2002. – Vol. 47(7). – P. 1179–1183. DOI: 10.1109/TAC.2002.800668.
- [2] Wu, Y. Kalman filtering with multiplicative and additive noises / Y. Wu, Q. Zhang, Z. Shen // In: Proc. of the 12th World Congress on Intelligent Control and Automation (WCICA). – 2016. – P. 483–487. DOI: 10.1109/WCICA.2016.7578352.
- [3] Tsyganov, A.V. UD-based Linear Filtering for Discrete-Time Systems with Multiplicative and Additive Noises / A.V. Tsyganov, J.V. Tsyganova, T.N. Kureneva // In: Proc. of the 19th European Control Conference (May 12–15, 2020. Saint Petersburg, Russia). – 2020. – P. 1389–1394. DOI: 10.23919/ECC51009.2020.9143804.
- [4] Astrom, K.-J. Maximum Likelihood and Prediction Error Methods / K.-J. Astrom // Automatica. – 1980. – Vol. 16(5). – P. 551–574.
- [5] Gibbs, B. P. Advanced Kalman filtering, least-squares and modeling: a practical handbook / B. P. Gibbs. – Hoboken, New Jersey : John Wiley & Sons, Inc., 2011. – 632 p.
- [6] Semushin, I. Adaptation in Stochastic Dynamic Systems—Survey and New Results IV: Seeking Minimum of API in Parameters of Data / I. Semushin, J. Tsyganova // International Journal of Communications, Network and System Sciences. – 2013. – Vol. 6(12). – P. 513–518. DOI: 10.4236/ijcns.2013.612055.

# Предсказание метеорологических величин с помощью гибридного метода обработки временных рядов

Е.А. Черных  
Московский государственный  
университет имени М.В.Ломоносова  
Москва, Россия  
chernykh.ea18@physics.msu.ru

Н.Е. Шапкина  
Московский государственный  
университет имени М.В.Ломоносова  
ИТПЭ РАН  
Москва, Россия  
neshapkina@mail.ru

П.В. Голубцов  
Московский государственный  
университет имени М.В.Ломоносова  
Москва, Россия  
golubtsov@physics.msu.ru

**Аннотация** — В данной работе представлен гибридный метод анализа временных рядов, основанный на авторегрессионной интегрированной модели скользящего среднего (ARIMA) и модели линейной регрессии, адаптированной к эффективной обработке больших данных, накапливаемых в режиме реального времени. Его практическое применение продемонстрировано на примере обработки реальных данных для предсказания временного ряда метеорологических величин.

**Ключевые слова** — временные ряды, большие данные, ARIMA, линейная регрессия

## 1. ВВЕДЕНИЕ

В силу большого объёма данных и их постоянного пополнения, анализ динамики метеорологических показателей в реальном времени является крайне ресурсоёмкой задачей.

В данной работе предложен гибридный метод обработки временных рядов, сочетающий в себе возможность параллельной обработки накапливаемой в течение длительного промежутка времени информации и построения достаточно точного прогноза в кратковременной перспективе без привлечения больших вычислительных мощностей.

## 2. ЭФФЕКТИВНАЯ ОБРАБОТКА ДАННЫХ В МОДЕЛИ ЛИНЕЙНОЙ РЕГРЕССИИ

Рассматриваются  $n$  пар наблюдений вида  $(t_i, y_i)$ ,  $i = 1, \dots, N$ ,  $t_i$  — временная метка,  $y_i$  — измерение прибора. В данной работе строится математическая модель линейной регрессии вида:

$$y = f(t) + \varepsilon(t), \quad (1)$$

где  $f(t)$  — функция регрессии,  $\varepsilon(t)$  — случайная величина. Функция регрессии представима в виде

$$f(t_i) = \beta_1 f_1(t_i) + \dots + \beta_m f_m(t_i) = \vec{F}(t_i) \cdot \vec{\beta} \quad (2)$$

где  $\vec{F}(t) = (f_1(t), \dots, f_m(t))$  — вектор строка из  $m$  функций,  $\vec{\beta} = (\beta_1, \dots, \beta_m)$  — вектор-столбец неизвестных коэффициентов. Коэффициенты оцениваются с помощью метода наименьших квадратов в сочетании с методом накопления «канонической информации» (КИ) [1]  $(T, v, V, n)$ , где

$$T = \sum_{i=1}^n \vec{F}^T(x_i) \vec{F}(x_i), \quad v = \sum_{i=1}^n \vec{F}^T(x_i) y_i, \quad (3)$$

$$V = \sum_{i=1}^n y_i^2. \quad (4)$$

На основании этих данных получается оценка коэффициентов регрессии  $\vec{\beta} = T^{-1}v$ , функции  $\hat{f}(x) = \vec{F}(x)T^{-1}v$ , которая аппроксимирует исходный ряд, и коридора погрешности этой функции  $D \hat{f}(x) = \frac{v - v^T T^{-1} v}{n-m} \vec{F}(x)T^{-1} \vec{F}^T(x)$ .

Такой подход позволяет разделить обработку данных на две фазы: выделение промежуточной информации и её последующая обработка [1]. Если рассмотреть два набора статистических данных размера  $n_1$  и  $n_2$ , то в силу аддитивности КИ  $(T, v, V, n) = (T_1 + T_2, v_1 + v_2, V_1 + V_2, n_1 + n_2)$ . Это позволяет извлекать КИ из предварительно разделённого ряда одновременно на нескольких устройствах и оперативно обновлять ее при поступлении новых измерений.

## 3. ИНТЕГРИРОВАННАЯ МОДЕЛЬ АВТОРЕГРЕССИИ СКОльзяЩЕГО СРЕДНЕГО

Другой распространённый способ анализа временных рядов — это интегрированная модель авторегрессии скользящего среднего (ARIMA). Описывающая стационарный стохастический процесса модель состоит из авторегрессионной модели порядка  $p$  и модели скользящего среднего порядка  $q$ :

$$y_t = \mu + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \dots - \theta_q \varepsilon_{t-q} \quad (5)$$

Модель ARIMA способна работать с нестационарными рядами. Она включает в себя операцию дифференцирования  $\Delta y_t = y_t - y_{t-1}$ , которая при её  $d$ -кратном повторении позволяет сделать ряд стационарным [2].

## 4. ПРЕИМУЩЕСТВА И НЕДОСТАТКИ МОДЕЛЕЙ

Каждая из описанных выше моделей обладает своими особенностями. Модель линейной регрессии с накоплением КИ предоставляет возможность параллельной обработки данных, позволяет добавлять новые измерения в режиме реального времени, нечувствительна к пропускам измерений во временных рядах, используется для выделения систематической составляющей ряда, учитывающей как дневную, так и годовую сезонность, по большому временному промежутку и отслеживания отклонений от общей тенденции. Однако она не позволяет делать прогноз для локальных отклонений от систематического поведения.

Модель ARIMA, напротив, способна делать точные предсказания на короткие временные промежутки, учитывая с наибольшими весами именно последние измерения во временных рядах. Однако, при добавлении новых измерений необходимо перестраивать модель или

вовсе подбирать новую с совершенно другими параметрами  $(p, d, q)$ . Кроме того, эта модель требовательна к данным, не допускает наличие пропущенных значений, а для обработки слишком большого количества измерений, например, чтобы отразить дневную и годовую сезонность, необходимы значительные вычислительные мощности.

Предлагается одновременно использовать преимущества каждой модели для эффективной обработки данных. Идея заключается в следующем: выполняется построение модели линейной регрессии с накоплением КИ для всех имеющихся значений временного ряда. Таким образом выделяется систематическая компонента ряда, которая включает в себя тренд, годовые изменения и суточный профиль. К ней добавляется прогноз модели ARIMA для фрагмента последних измерений и так формируется уточнённое предсказание.

## 5. МОДЕЛИРОВАНИЕ И ПРОГНОЗИРОВАНИЕ РЯДОВ

Исследовался временной ряд показателей атмосферного давления на территории заповедника во Вьетнаме (метеорологическая станция "AsiaFlux" [3]) в период с 2013 по 2021 годы с интервалом в 30 минут.

Рис. 1 и 2 демонстрируют предсказания, полученные тремя способами: моделью линейной регрессии,

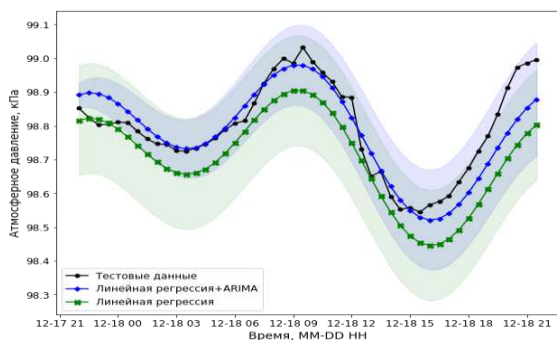


Рис.1. Сравнение предсказания гибридной модели и линейной регрессии на сутки вперёд для атмосферного давления.

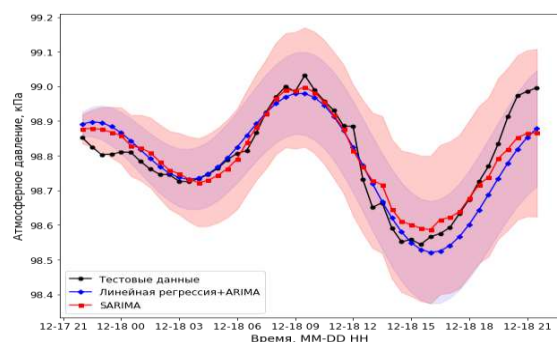


Рис.2. Сравнение предсказания гибридной модели и сезонной ARIMA на сутки вперёд для атмосферного давления.

сезонной ARIMA и предложенным гибридным методом.

В модели линейной регрессии для временного ряда использованы периодические функции синуса и косинуса вплоть до третьей гармоники, учитывающие сезонность, как годовую, так и суточную. Линия тренда

аппроксимирована полиномом второй степени. Для предсказания используются данные, накопленные за длительный промежуток времени в несколько лет. Параметры сезонной модели ARIMA подобраны исходя из анализа корреляционной и автокорреляционной функций исходного ряда. Предсказание строилось по данным за предшествующую неделю.

Для сравнения полученных предсказаний используются следующие метрики: средняя абсолютная ошибка (MAE), средняя квадратическая ошибка (MSE) и средняя абсолютная ошибка в процентах (MAPE) [4]. Случайным образом выбраны 10 недельных отрезков из всего ряда и построены предсказания на сутки вперёд для каждой модели (таблица I).

Таблица I. СРЕДНЕЕ ЗНАЧЕНИЕ МЕТРИК ДЛЯ РАЗНЫХ МОДЕЛЕЙ ВРЕМЕННЫХ РЯДОВ

	MSE	MAE	MAPE
Линейная регрессия	0,0315	0,1473	0,1981
Сезонная ARIMA	0,0093	0,0749	0,0758
Гибридная модель	0,0074	0,0436	0,0441

Минимальное значение среди всех метрик достигается гибридной моделью, что говорит о более высокой точности построения предсказания, чем у каждой из исходных моделей. Кроме того, гибридная модель не требует больших вычислительных и временных ресурсов.

## 6. ЗАКЛЮЧЕНИЕ

В работе было проведено сравнение трёх моделей прогнозирования метеорологических временных рядов – линейной регрессии с возможностью накопления канонической информации, сезонной модели ARIMA и гибридной. Для каждой из моделей помимо предсказания был получен его коридор погрешности. Предложенная смешанная модель представила достаточно точный прогноз с наименьшими затратами времени и ресурсов, поэтому её можно считать оптимальной для быстрой обработки метеорологических временных рядов.

## БЛАГОДАРНОСТИ

Авторы благодарят сотрудников ИПЭЭ РАН им. А.Н. Северцова Ю.А. Курбатову и В.К. Авилова за предоставленные данные метеорологических величин.

## ЛИТЕРАТУРА

- [1] Golubtsov, P. V. The concept of information in big data processing / P. V. Golubtsov // Automatic Documentation and Mathematical Linguistics. – 2018. – Vol. 52. – P. 38-43.
- [2] Бокс, Дж. Анализ временных рядов. Прогноз и управление / Дж. Бокс, Г. Дженкинс. – М.: Мир, 1994. – 407с.
- [3] Сайт метеорологической станции AsiaFlux [Электронный ресурс]. – Режим доступа: [http://asiaflux.net/index.php?page\\_id=86](http://asiaflux.net/index.php?page_id=86) (25.05.2022).
- [4] Chicco, D. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation / D. Chicco, M. J. Warrens, G. Jurman // PeerJ Computer Science. – 2021. – Vol. 7. – P. e623.

# Построение алгоритма аннотирования русскоязычных текстовых данных социальных сетей с использованием переносимого обучения

Д.С. Баканов

Самарский национальный исследовательский университет  
им. академика С.П. Королева  
Самара, Россия  
dima.bakanov.1999@mail.ru

А.В. Куприянов

Самарский национальный исследовательский университет  
им. академика С.П. Королева  
ИСОИ РАН  
Самара, Россия  
akupr@ssau.ru

**Аннотация**—В данной работе рассматриваются способы построения алгоритма аннотирования русскоязычных текстов из социальных сетей. В качестве аннотирования будем понимать оценку эмоционального окраса текста. Статья затрагивает как классические базовые методы статистического обучения, так и современные методы глубокого обучения, основанные на переносимом обучении и трансформерах. В заключении строится модель, которая совмещает модель трансформера и статистическую модель машинного обучения градиентного бустинга. Актуальность данной работы заключается в создании легковесной и независимой от тематики модели, которую можно использовать для анализа текстового содержимого постов в социальных сетях.

**Ключевые слова**— обработка естественного языка, трансформер, переносимое обучение, анализ социальных сетей, TF-IDF, статистическое обучение, оценка эмоционального окраса, BERT

## 1. ВВЕДЕНИЕ

В наши дни все большую роль играют социальные сети, которые становятся местом притяжения все большего количества людей с разными интересами. Поэтому социальные сети могут служить хорошим местом при проведении социальных экспериментов.

Эмоции – это быстрые и короткие реакции человеческих чувств, их недопонимание при проведении рекламных компаний, маркетинга может повлечь за собой большие финансовые потери [1]. Уникальность социальной сети заключается в том, что сами пользователи создают контент, которого становится все больше. Но с прогрессом информационных технологий и развитием технологий машинного обучения и больших данных можно автоматически анализировать такую информацию [2].

На данный момент существует большое количество моделей, которые предсказывают эмоциональный окрас для постов с отзывами о продуктах и услугах [3, 4]. В данной работе описывается полный цикл построения алгоритма аннотирования данных, который не зависит от тематики постов в социальных сетях, что можно использовать при анализе любого поста в социальных сетях, с использованием трансформера и градиентного бустинга.

## 2. ИСХОДНЫЕ ДАННЫЕ

В качестве обучающего набора данных были использованы следующие обучающие наборы: RuReviews [3], RuTweetCorp [4], Kaggle Russian News Dataset [5]. Всего набор данных насчитывает 358190

образцов. В данном наборе фигурирует три класса эмоционального окраса: негативный (-1), нейтральный (0) и положительный (1). Ниже представлено распределение образцов по классам (см. Рисунок 1).

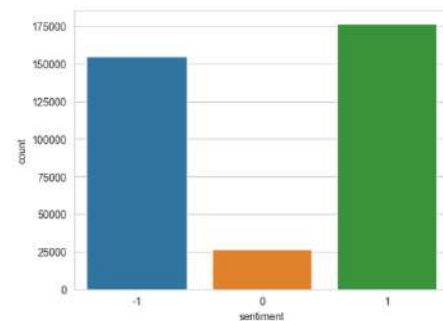


Рис. 1 Распределение образцов набора данных по классам

Из рисунка 1 можно видеть, что наблюдается сильный дисбаланс классов. Данную проблему стоит учесть при выборе метрики качества и моделей, которые устойчивы к дисбалансу классов.

## 3. РАЗВЕДЫВАТЕЛЬНЫЙ АНАЛИЗ

Перед проведением разведывательного анализа данные были предобработаны: из текста была удалена HTML-разметка, текст приведен к общему регистру, удалены знаки препинания и стоп-слова.

На рисунке 2 приведена визуализация данных из обучающего набора, которая была сделана при помощи применения метода латентного семантического анализа (индексирования) [6].

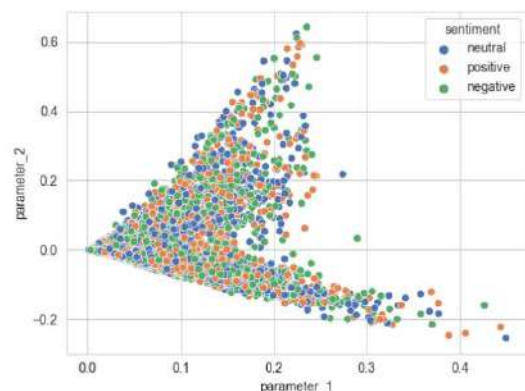


Рис. 2 Векторное представление набора данных на плоскости

Как можно видеть данные не являются линейно разделимыми и находятся вперемешку друг с другом.

#### 4. СТАТИСТИЧЕСКОЕ МАШИННОЕ ОБУЧЕНИЕ ПРИ АНАЛИЗЕ ЭМОЦИОНАЛЬНОГО ОКРАСА РУССКОЯЗЫЧНЫХ ТЕКСТОВ

В качестве базового решения задачи классификации можно выбрать статистические методы машинного обучения, которые хорошо себя показывают при решении задач в области обработки естественного языка [6].

Перед обучением данные были приведены в векторное представление:

- удалены пустые значения;
- разбиты на N-граммы;
- созданы векторы на основе метрики TF-IDF.

Чтобы учесть дисбаланс классов при оценке моделей для задачи многоклассовой классификации, была использована метрика F1-мера.

Была исследована зависимость метрики точности модели от выбора N-грамм (см. Таблицу I).

Таблица I. ЗАВИСИМОСТЬ F1-МЕРЫ ДЛЯ КАЖДОЙ МОДЕЛИ В ЗАВИСИМОСТИ ОТ ВЫБОРА N-ГРАММЫ

Модель машинного обучения	N-грамма		
	Униграмма	Биграмма	Триграмма
Наивный байесовский классификатор	0,67	0,59	0,48
Метод опорных векторов (SVM)	0,45	0,28	0,25
Линейный метод опорных векторов	0,67	0,57	0,43
Логистическая регрессия	0,67	0,55	0,4
Деревья решений	0,62	0,47	0,38
Градиентный бустинг (CatBoost)	0,63	0,48	0,37

Точность модели не превосходит 0,67 и с увеличением N-граммы уменьшается, что показывает значимость каждого слова при анализе эмоционального окраса.

#### 5. ТРАНСФОРМЕРЫ И ПЕРЕНОСИМОЕ ОБУЧЕНИЕ

Трансформеры – класс моделей глубокого обучения, которые с использованием механизма внутреннего внимания решают задачу преобразования последовательности в другую последовательность. Рекомендуемым методом обучения трансформеров является переносимое обучение (transfer learning), которое заключается в настройке уже обученной модели на своем наборе данных [7].

##### А. Выбор BERT-модели

В качестве модели для дообучения была выбрана легковесная модель DeepPavlov/rubert-base-cased. Данная модель отличается своей легковесностью и обучена на русскоязычных статьях Википедии [8].

##### Б. Обучение и оценка точности

Для точной настройки трансформера были использованы следующие параметры:

- оптимизатор: Adam;
- размер пакета: 32;
- learning rate:  $2e-5$ ;
- размер вектора: 128.

На рисунке 3 показан график потерь на обучающем и проверочном наборе для 10 эпох обучения.

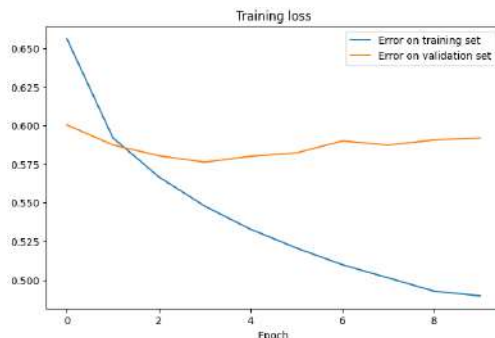


Рис. 3 Потери при обучении и проверке трансформера

На тестовых данных F1-мера трансформера оказалась равной 0,7.

##### В. Сочетание с моделями статистического обучения

После получения векторов трансформера к ним были применены статистические модели машинного обучения. Самую большую точность показал градиентный бустинг (CatBoost). Точность по F1-мере составила 0,76. Данная точность превышает точности предсказаний статистических моделей машинного обучения, а также модели трансформера.

#### 6. ЗАКЛЮЧЕНИЕ

В данной работе был рассмотрен метод построения алгоритма аннотирования текстовых данных социальных сетей. Полученный алгоритм обладает рядом особенностей:

- независимость оценки эмоционального окраса от тематики поста;
- использование легковесных моделей;
- точность составила по F1-мере 0,76, что превышает показатели точности тяжеловесных моделей при анализе эмоционального окраса русскоязычных текстов [3].

#### ЛИТЕРАТУРА

- [1] Канарская Л. Как работает эмоциональный контент в SMM (на примере популярных групп «ВКонтакте») [Электронный ресурс]. – Режим доступа: <https://texterra.ru/blog/kak-rabotaet-emotsionalnyy-kontent-v-smm-na-primere-populyarnykh-grupp-vkontakte.html> (дата обращения: 06.06.2022).
- [2] Рыцарев, И.А. Кластеризация медиаконтента из социальных сетей с использованием технологий Big Data / И.А. Рыцарев, Д.В. Кириш, А.В. Курпиров // Компьютерная оптика. – 2018. – Т. 42, № 5. – С. 921-927. .
- [3] Smetanin, S. "Sentiment Analysis of Product Reviews in Russian using Convolutional Neural Networks" / S. Smetanin, M. Komarov // 2019 IEEE 21st Conference on Business Informatics (CBI). – 2019. – P. 482-486.
- [4] Рубцова, Ю. Автоматическое построение и анализ корпуса коротких текстов (постов микроблогов) для задачи разработки и тренировки тонового классификатора / Ю. Рубцова // Инженерия знаний и технологии семантического веба. – 2012. – Т. 1. – С. 109-116.
- [5] Sentiment Analysis in Russian [Электронный ресурс]. — Режим доступа: <https://www.kaggle.com/c/sentiment-analysis-in-russian> (01.06.2022).
- [6] Маннинг, Д. Введение в информационный поиск / Д. Маннинг, Р. Прабхакар, Х. Шютце. – СПб.: ООО «Диалектика», 2020. – 528 с.
- [7] Vaswani, A. Attention is all you need / A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, I. Polosukhin // Computer Science. – 2017.
- [8] Kuratov, Y. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language / Y. Kuratov, M. Arkhipov // Computer Science, Linguistics. – 2019.

# Математическое моделирование вольт-амперной характеристики мемристора с учетом его неоднородности

Д.В. Продан

Сколковский институт науки и технологий  
Москва, Россия  
dmitrii.prodan@skoltech.ru

*Аннотация—Рассмотрена модель вольт-амперной характеристики мемристора, где неоднородности в устройстве учитываются в рамках суперстатистики, а динамика его внутреннего состояния описывается дифференциальным уравнением с дробной производной. Показано, что предложенная модель дает хорошее согласие с результатами экспериментальных измерений вольт-амперной характеристики реального мемристора.*

*Ключевые слова— мемристор, вольт-амперная характеристика, суперстатистика, дробная производная*

## 1. ВВЕДЕНИЕ

Мемристор – это пассивный двухполюсный элемент, сопротивление которого меняется в зависимости от интеграла по времени от напряжения, приложенного к полюсам устройства. Мемристор был теоретически предложен в 1971 году как естественное дополнение к известным элементарным двухполюсным устройствам – резистору, конденсатору и катушке индуктивности [1]. В настоящее время мемристоры находят применение в перспективных устройствах хранения и обработки информации [2-4].

В общем случае модель мемристора описывается системой уравнений [1]

$$i = G(t,x,v)v, \quad x' = f(t,x,v) \quad (1)$$

где  $i$  – проходящий через устройство ток,  $v$  – приложенное к нему напряжение,  $x$  – переменная его состояния,  $G$  – проводимость, функция  $f$  описывает динамику внутренней переменной, и  $t$  – время. Характерной особенностью решения системы уравнений (1) для мемристора является нелинейная вольт-амперная характеристика (ВАХ) с двусторонней петлей гистерезиса, проходящей через начало координат.

В последние годы предложены разнообразные модели, позволяющие учесть различные физические процессы, происходящие внутри мемристора. Как правило, такие модели хорошо описывают конкретный тип мемристора как, например, модель мемристора НР для устройства на основе оксида титана [5]. Широкое применение получила обобщенная модель Якопича [6], релевантная для мемристоров с различной реализацией. Недавно данная модель была расширена за счет учета неоднородностей мемристора в рамках подхода суперстатистики [8]. В настоящей работе предложена модель, где динамика внутренней переменной состояния мемристора описывается дифференциальным уравнением с дробной производной, что позволяет увеличить количество степеней свободы модели за счет добавления нового параметра – порядка производной – и точнее описать долговременные эффекты памяти.

## 2. МОДЕЛЬ ТОКА НА ГРАНИЦЕ ЭЛЕКТРОД-СРЕДА

В работе рассматривается модель, где уравнение для тока имеет вид

$$i = \gamma_1 x \sinh_q(\delta_1 v) + \gamma_2 (1-x) \sinh_q(\delta_2 v) \quad (2)$$

Уравнение (2) получается на основе подхода суперстатистики, где в гиперболическом синусе вместо обычной экспоненты используется ее  $q$ -деформированный аналог [9]. Здесь  $\gamma_1$ ,  $\gamma_2$ ,  $\delta_1$ ,  $\delta_2$ ,  $q$  – параметры модели. Во втором уравнении системы (1) вместо производной по времени используется дробная производная,

$$D_t^\alpha x = f(x,v,t) \quad (3)$$

где  $D_t^\alpha$  – это дробная производная порядка  $0 < \alpha < 1$  по Капуто, а правая часть уравнения (3) взята из обобщенной модели Якопича [6].

## 3. ВЫЧИСЛЕНИЕ ПАРАМЕТРОВ МОДЕЛИ

Для вычисления параметров модели разработан стохастический метод, основанный на многократной минимизации квадратов ошибки со стохастической вариацией начальных значений. Вариация необходима для поиска глобального минимума ошибки и осуществляется на двух уровнях. В первой процедуре минимизации используются заданные пользователем начальные значения параметров модели. Полученный вектор параметров и ошибка заносятся в соответствующие динамические массивы.

Далее производится несколько циклов оптимизации. В начале каждого цикла из массива полученных параметров взвешенным стохастическим методом выбирается один из векторов. В значения параметров вносится гауссовский шум со стандартным отклонением равным

$$\sigma = \sqrt{|p_i + 1|} \quad (4)$$

где  $p_i$  – компоненты вектора параметров, а добавление единицы используется для предотвращения "замерзания" параметров малых по модулю.

Внутри внешнего цикла выполняется внутренний цикл, на каждой итерации которого начальные параметры подвергаются воздействию равномерного шума с диапазоном в 4 раза меньшим, чем стандартное отклонение шума во внешнем цикле. По результатам каждой итерации, начиная со второй, новые значения принимаются согласно алгоритму, схожему с алгоритмом метода Метрополиса-Гастингса (Accept-

Reject Method) [7] где в экспоненциальной функции в числителе используется разность между старым и новым значением ошибки, а в знаменателе - условная переменная, которая умножается на коэффициент в интервале (0,1) на каждом шаге итерации. По завершении внутреннего цикла полученные итоговые вектор параметров и ошибка заносятся в массив параметров и ошибок и внешний цикл начинается заново.

Для решения метода наименьших квадратов в каждой итерации используется метод нелинейной многомерной оптимизации из библиотеки GNU Scientific Library для языков C/C++, а именно интерфейс `gsl_multifit_nlinear` [10].

Для оценки погрешности используется параметр NRMSE описываемый формулой:

$$NRMSE = \frac{1}{\langle y \rangle} \sqrt{\frac{\sum_{t=1}^T (y_t^* - y_t)^2}{T}} \quad (5)$$

где  $\langle y \rangle$  – среднее по всем измерениям,  $T$  – общее время измерения, а под знаком суммы стоит квадрат разности между измеренным и рассчитанным значениями.

#### 4. ЗАКЛЮЧЕНИЕ

Предложенная математическая модель была использована для расчета ВАХ реального мемристора (см. Рис. 1).

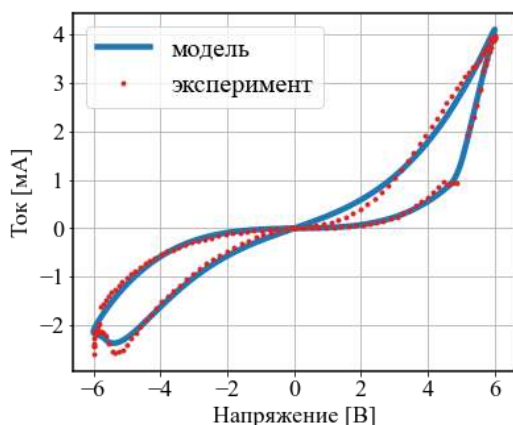


Рис. 1. Расчет ВАХ мемристора в рамках модели с q-деформированной экспонентой и дробно-дифференциальным уравнением состояния (синяя кривая) в сравнении с экспериментальными данными (красные точки)

Параметры модели были вычислены так, чтобы минимизировать среднеквадратичную ошибку между результатом расчета модели и данными эксперимента. Анализ среднеквадратичной ошибки показывает лучшее согласие разработанной модели с экспериментальными данными ВАХ по сравнению с более простыми моделями мемристора (в частности, обобщенной модели [6] и модели [8]), см. Табл. I. Таким образом, переход к дробной производной позволяет точнее описывать динамику состояния устройства и может

рассматриваться как обобщение предыдущих математических моделей мемристора.

Таблица I. ОЦЕНКА ПОГРЕШНОСТИ МОДЕЛЕЙ

Модель	NRMSE
Обобщённая [6,8]	0,533
С q-деформированной экспонентой [8]	0,467
С q-деформированной экспонентой и дробной производной	0,443

Разработанные электрические модели мемристоров позволяют описывать поведение электронных компонент в симуляторах исходных схем, которые в дальнейшем могут быть использованы при создании альтернативных электронных устройств на базе мемристоров. Примером таких устройств являются перспективные нейроморфные мемристивные матрицы на перекрёстных массивах электродов, имитирующих синаптическую передачу сигнала с заданными весовыми коэффициентами. Реализация такой архитектуры возможна с использованием программы SPICE [11].

#### ЛИТЕРАТУРА

- [1] Chua, L.O. Memristor—the missing circuit element. / L.O. Chua // IEEE Trans. Circuit Theory. – 1971. – Vol.18(5). – P. 507–519.
- [2] Huang, H. Artificial Neural Networks Based on Memristive Devices: From Device to System / H. Huang, Z. Wang, T. Wang, Y. Xiao, X. Guo // Advanced Intelligence Systems. – 2020. – Vol. 2(12). – P.2000149. DOI: 10.1002/aisy.202000149
- [3] Bao, H. Toward memristive in-memory computing: principles and applications. / H. Bao, H. Zhou, J. Li et al. // Front. Optoelectron. – 2022. – Vol. 15(23). DOI: 10.1007/s12200-022-00025-4
- [4] Peotta, S. Superconducting Memristors / S. Peotta, M. Ventra. // Phys. Rev. Appl. – 2014. – Vol. 2(3). – P. 034011. DOI: 10.1103/PhysRevApplied.2.034011
- [5] Strukov, D. The missing memristor found. / D. Strukov, G. Snider, D. Stewart, R. Williams // Nature. – 2008. – Vol. 453. – P. 80–83. DOI: 10.1038/nature06932
- [6] Yakopcic, C. A Memristor Device Model / C. Yakopcic, T. Taha, G. Subramanyam, R. Pino, S. Rogers // IEEE Electron Device Letters. – 2011. – Vol. 32(10). – P. 1436-1438. DOI: 10.1109/LED.2011.2163292
- [7] Metropolis, N. Equation of State Calculations by Fast Computing Machines / N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, E. Teller // J. Chem. Phys. – 1953. – Vol. 21. – P. 1087-1092. DOI: 10.1063/1.1699114
- [8] A superstatistics approach to memristor current-voltage modelling / R. Konlechner, A. Allagui, V. Antonov, D. Yudin // Physica A: Statistical Mechanics and its Applications. – 2023. – Vol. 614.
- [9] Umarov, S. On a q-Central Limit Theorem Consistent with Nonextensive Statistical Mechanics. /S. Umarov, C. Tsallis, S. Steinberg // Milan j. math. – 2008. – Vol.76. – P.307–328. DOI: 10.1007/s00032-008-0087-y
- [10] Библиотека численных методов GNU Scientific Library, нелинейный метод наименьших квадратов [Электронный ресурс]. – Режим доступа: <https://www.gnu.org/software/gsl/doc/html/nls.html>
- [11] Yakopcic, C. Memristor SPICE model and crossbar simulation based on devices with nanosecond switching time. / C. Yakopcic, T. Taha G. Subramanyam, R. Pino // Proceedings of the 2013 ICNN. – 2013. – P. 1-7. DOI: 10.1109/ICNN.2013.6706773.

# Квадратно-корневой алгоритм вычисления отношения правдоподобия в задаче обнаружения изменения и идентификации режима движения

А.В. Голубков  
Ульяновский государственный  
педагогический университет  
им. И.Н. Ульянова  
Ульяновск, Россия  
kr8589@gmail.com

Ю.В. Цыганова  
Ульяновский государственный  
университет  
Ульяновск, Россия  
tsyganovajv@gmail.com

А.В. Цыганов  
Ульяновский государственный  
педагогический университет  
им. И.Н. Ульянова  
Ульяновск, Россия  
andrew.tsyganov@gmail.com

**Аннотация**—В работе рассмотрена задача обнаружения изменения и идентификации режима движения объекта. Решение построено с помощью последовательного решающего правила. С целью повышения качества работы алгоритма предложены новые выражения для вычисления отношения правдоподобия на основе квадратно-корневой модификации фильтра Калмана.

**Ключевые слова**—последовательное решающее правило, фильтр Калмана, квадратно-корневой алгоритм

## 1. ВВЕДЕНИЕ

Задачи математического моделирования траекторий движущихся объектов, слежения за движущимися объектами, распознавания движущихся объектов, сопровождения целей являются актуальным предметом современных научных исследований в силу крайней важности современных практических приложений, в которых используются решения этих задач [1].

Одним из наиболее популярных методов оценивания параметров движения объектов на протяжении многих десятилетий является фильтр Калмана [2].

Предположим, что траекторию объекта можно разделить на отдельные достаточно длинные участки, на каждом из которых его движение может быть представлено линейной дискретной стохастической моделью, описывающей либо прямолинейное равномерное движение, прямолинейное движение с ускорением, остановку, либо круговое движение против/по часовой стрелке с заданным радиусом. Для описания такого движения предлагается использовать гибридную стохастическую модель [3].

Задача заключается в скорейшем обнаружении изменения режима движения на каждом таком участке траектории с целью вычисления оптимальных оценок параметров движения объекта в режиме реального времени. Один из подходов к решению данной задачи, основанном на применении последовательного решающего правила с ограниченным объемом банка фильтров Калмана, рассмотрен в [4]. Другой подход, предложенный в [5], позволяет получить решение на ограниченном наборе значений функции отношения правдоподобия за счет представления неизвестного момента изменения режима движения объекта случайной величиной с равномерным распределением на заданном отрезке времени. Оба подхода основаны на методах последовательного анализа, подробный обзор которых содержит работа Т.Л. Lai [6]. Следует отметить, что в настоящее время методы оценивания параметров

движения объектов активно развиваются и в направлении современного мультиагентного подхода [7].

В данной работе предлагается развитие полученных ранее результатов с целью повышения качества работы алгоритмов, основанное на применении численно устойчивой квадратно-корневой модификации фильтра Калмана.

## 2. КВАДРАТНО-КОРНЕВОЙ АЛГОРИТМ

Предположим, что момент возможного изменения режима движения и номер режима движения априорно неизвестны.

Рассмотрим  $M$  возможных режимов движения ( $q = 0, \dots, M-1$ ). Предположим, что начальное состояние системы соответствует номинальному режиму движения с номером 0. Необходимо по результатам измерений подтвердить или опровергнуть факт изменения режима движения объекта и идентифицировать его номер.

Решение поставленной задачи может быть получено с помощью последовательного решающего правила, которое определяет выбор одной из  $M$  гипотез, и может быть записано следующим образом:

- Если  $\forall q: \lambda_{qk} \leq B$ , тест завершают с выбором гипотезы  $H_0$ . Изменение режима движения не обнаружено.
- Если  $\exists! i: \lambda_{ik} \geq A$ , тест завершают с выбором гипотезы  $H_i$ . Обнаружено изменение режима движения на режим с номером  $i$ .
- Если  $\forall q: A > \lambda_{qk} > B$ , тест продолжают для следующего  $k$ .
- Если  $\exists i, n: \lambda_{nk} \geq A$  и  $\lambda_{ik} \geq A$ , тест завершают с выбором гипотезы  $H_q$ , где  $q = \max(i, n)$ . Обнаружено изменение режима движения на режим с номером  $q$ .

Предположим, что момент возможного изменения режима движения с номинального (под номером 0) на альтернативный (под номером  $q$ ,  $q = 1, \dots, M-1$ ) представляет собой дискретную случайную величину  $\theta$ , равномерно распределенную на отрезке  $[1, k]$ . Тогда отношение функций правдоподобия в последовательном решающем правиле определяется выражением:

$$\lambda_{qk} = \frac{1}{k} \sum_{j=1}^k \psi_j^q(k), \quad (1)$$

где значения функций  $\psi_j^q(k)$  вычисляются на основе оценок, получаемых фильтром Калмана.

С целью повышения качества работы алгоритма получены новые выражения для вычисления величин  $\psi_j^q(k)$ , основанные на применении численно устойчивого квадратно-корневого фильтра Калмана [8].

Пусть  $SR = \{F_{01}, F_{qj} | q = 1, \dots, M-1, j = 1, \dots, k\}$  – банк квадратно-корневых фильтров Калмана, необходимый для вычисления значений отношения функций правдоподобия. Тогда выражение для величин  $\psi_j^q(k)$  имеет вид:

$$\psi_j^q(k) = \begin{cases} \psi_j^q(k-1) \times \frac{(S_{R_{e,k}}^q)_j}{(S_{R_{e,k}}^0)_1} \exp \left[ \frac{\|(\bar{e}_k^{SR})_1^0\|^2 - \|(\bar{e}_k^{SR})_j^q\|^2}{2} \right], & 1, k < j, \\ \psi_j^q(k-1) \times \frac{(S_{R_{e,k}}^q)_j}{(S_{R_{e,k}}^0)_1} \exp \left[ \frac{\|(\bar{e}_k^{SR})_1^0\|^2 - \|(\bar{e}_k^{SR})_j^q\|^2}{2} \right], & k \geq j, \end{cases} \quad (2)$$

в котором невязки  $\bar{e}_k^{SR}$  и квадратный корень ковариационной матрицы невязок  $S_{R_{e,k}}$  получены по алгоритму квадратно-корневого ковариационного фильтра (SRCF).

### 3. ЧИСЛЕННЫЙ ПРИМЕР

В качестве примера рассмотрим движение объекта по траектории, состоящей из двух участков кругового движения: 1) движение при повороте вправо с радиусом 4 м (40 тактов дискретного времени), 2) движение при повороте влево с радиусом 5 м (60 тактов дискретного времени). В уравнениях движения объекта и измерений присутствуют гауссовы шумы с нулевыми математическими ожиданиями и ковариационными матрицами  $Q = 0.001I_2$  и  $R = 0.3I_2$ , соответственно. Траектория движения объекта и зашумленные измерения представлены на рис. 1.

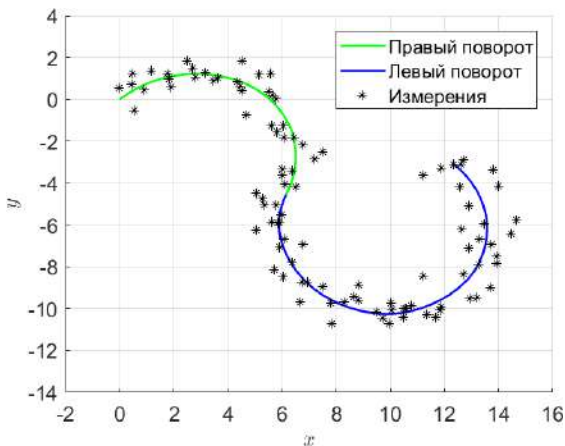


Рис. 1. Траектория движения и зашумленные измерения

На рис. 2 приведен график отношения функций правдоподобия. Менее чем за 10 тактов дискретного времени после 40 такта отношение функций правдоподобия пересекает верхний порог  $A$ , что соответствует принятию гипотезы об изменении режима движения с номинального (правый поворот) на альтернативный (левый поворот), то есть обнаружению факта изменения режима движения.

### 4. ЗАКЛЮЧЕНИЕ

В работе рассмотрена задача обнаружения изменения режима движения объекта на основе последовательного решающего правила и численно эффективной квадратно-корневой модификации фильтра Калмана. Получено новое выражение (2) для вычисления отношения правдоподобия на основе величин, вычисляемых квадратно-корневым ковариационным фильтром. Результаты численных экспериментов подтверждают работоспособность предложенного подхода.

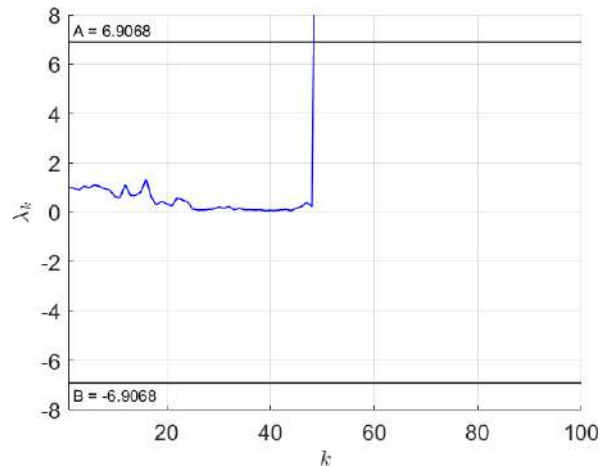


Рис. 2. Отношение правдоподобия

### БЛАГОДАРНОСТИ

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 20-31-90132 Аспиранты.

### ЛИТЕРАТУРА

- [1] Коновалов, А. А. Основы траекторной обработки радиолокационной информации / А. А. Коновалов. – СПб. : Изд-во СПбГУ ЛЭТИ, 2013. – 164 с.
- [2] Gibbs, V. P. Advanced Kalman filtering, least-squares and modeling: a practical handbook / V. P. Gibbs. – Hoboken, New Jersey : John Wiley & Sons, Inc., 2011. – 632 p.
- [3] Семушин, И.В. Моделирование и оценивание траектории движущегося объекта / И.В. Семушин, А.В. Цыганов, Ю.В. Цыганова, А.В. Голубков, С.Д. Винокуров // Вестник Южно-Уральского государственного университета. Серия: «Математическое моделирование и программирование». – 2017. – Т. 10, № 3. – С. 108–119. DOI: 10.14529/mmp170309
- [4] Голубков, А.В. Решение задачи обнаружения изменения режима движения объекта с ограниченным объемом банка фильтров Калмана / А.В. Голубков // Автоматизация процессов управления. – 2020. – Т.1, №59. –С. 14–23. – DOI: 10.35752/1991-2927-2020-1-5-14-23
- [5] Цыганова, Ю.В. Метод обнаружения факта нарушения и его диагностики в линейных стохастических системах в процессе фильтрации / Ю.В. Цыганова // Вестник Самарского государственного аэрокосмического университета Серия: Управление, вычислительная техника и информатика. – 2009. – Т. 2, №18. – С. 163–171.
- [6] Lai, T. L. Sequential Analysis: Some Classical Problems and New Challenges / T. L. Lai // Statistica Sinica. – 2001. – Vol. 11. – P. 303–408.
- [7] Amelina, N. Consensus-based distributed algorithm for multisensor-multitarget tracking under unknown-but-bounded disturbances / N. Amelina, V. Erofeeva, O. Granichin et. al. // IFAC-PapersOnLine. – 2020. – Vol. 53(2). – P. 3589–3595.
- [8] Kailath, T. Linear estimation / T. Kailath, A. H. Sayed, B. Hassibi. – New Jersey: Prentice Hall, 2000. – 856 p.

# Организация высокопроизводительных вычислений для исследования живучести энергетических инфраструктур

А.В. Еделев  
Институт систем энергетики им.  
Л.А. Мелентьева СО РАН  
Иркутск, Россия  
flower@isem.irk.ru

С.А. Горский  
Институт динамики систем и  
теории управления им. В.М.  
Матросова СО РАН  
Иркутск, Россия  
gorsky@icc.ru

А.Г. Феоктистов  
Институт динамики систем и  
теории управления им. В.М.  
Матросова СО РАН  
Иркутск, Россия  
agf@icc.ru

И.В. Бычков  
Институт динамики систем и  
теории управления им. В.М.  
Матросова СО РАН  
Иркутск, Россия  
idstu@icc.ru

**Аннотация**— Целью исследования является разработка предметно-ориентированного подхода к организации высокопроизводительной вычислительной среды для исследования живучести энергетических комплексов, которые относятся к ключевым критическим инфраструктурам. В качестве ядра среды выступает инструментальный комплекс Orlando Tools, используемый для разработки и применения распределенного пакета прикладных программ для исследования живучести, интеграции разнородных высокопроизводительных вычислительных ресурсов в единую среду и управления вычислениями в этой среде. Результаты исследования продемонстрированы на примере решения практических задач оценки живучести энергетических инфраструктур.

**Ключевые слова**— критические инфраструктуры, системы энергетики, исследование живучести, анализ данных, высокопроизводительные вычисления

## 1. ВВЕДЕНИЕ

К критическим инфраструктурам относятся те системы, нарушение функционирования которых отрицательно влияет на государство, экономику и благосостояние общества. Энергетическая инфраструктура является одной из ключевых, так как она обеспечивает функционирование зависящих от нее прочих критических инфраструктур. Под живучестью энергетической инфраструктуры понимается ее свойство противостоять крупным возмущениям, не допуская их каскадного развития с массовым нарушением режима энергоснабжения потребителей, и восстанавливать исходное состояние системы или близкое к нему [1]. В работе [2] вводятся сводные показатели для количественной оценки живучести на основе формы и размеров её кривой. Пусть  $t$  – число периодов сценария возмущения, вектор  $x^t = (x_1^t, x_2^t, \dots, x_m^t)$  описывает состояние энергетической инфраструктуры в период  $t \in \overline{1, t}$ , а показатель  $h(X, t): x^t \rightarrow \mathbf{R}$  отображает состояние энергетической инфраструктуры  $x^t$  из множества  $X = \{x^1, x^2, \dots, x^t\}$  в скалярное значение. Затем сводная метрика  $w(X, [a, b]): (h(i_1), h(i_2), \dots, h(i_k)) \rightarrow \mathbf{R}$  отображает сегмент кривой устойчивости, который характеризует одно из измерений устойчивости, в

скаляр,  $a = i_1, b = i_k, i_1 < i_2 < \dots < i_k, i_1, i_2, \dots, i_k \in \overline{1, t}$ .

Например, участок кривой живучести, расположенный сразу после возмущения, отражает «пассивную» реакцию энергетической инфраструктуры на возникновение экстремальных условий в виде падения производительности. Этот показатель характеризует уязвимость как одно из измерений живучести энергетической инфраструктуры (рис. 1).

Целью анализа глобальной уязвимости является определение зависимости значения сводной метрики, показывающей наибольшее падение производительности энергетической инфраструктуры по отношению к числу отказавших элементов. Моделируются несколько серий сценариев возмущения с увеличивающимся в конечном итоге числом отказавших элементов энергетической инфраструктуры. Так как процессы моделирования возмущений для различных сценариев могут выполняться независимо друг от друга, то ускорение обработки всего множества сценариев возмущений достигается за счет использования параллельных и распределенных вычислений.

## 2. СРЕДА МОДЕЛИРОВАНИЯ ЖИВУЧЕСТИ ЭНЕРГЕТИЧЕСКИХ СИСТЕМ

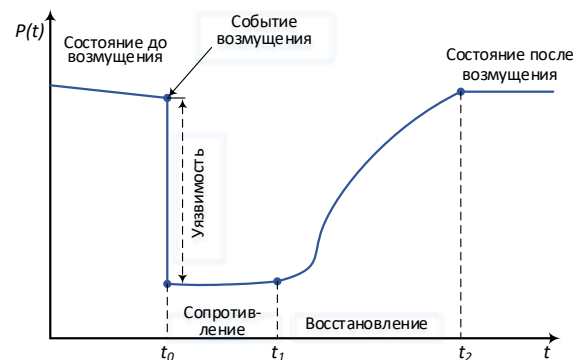


Рис. 1. Кривая живучести



Рис. 2. Среда для исследования живучести энергетических инфраструктур

Анализ известных литературных источников, в которых описывается использование высокопроизводительных вычислений для исследования живучести энергетической инфраструктуры, позволяет сделать следующие выводы. В рамках рассмотренных работ, как правило, решается специфическая задача из области исследования живучести конкретной системы энергетики, например, оценка возможности каскадного развития аварий в электроэнергетике [3]. Нет единого подхода к организации параллельных или распределенных вычислений. Выбор вычислительной среды во многом обуславливается размерностями задач и особенностями алгоритмов их решения. При этом зачастую осуществляется адаптация прикладного программного обеспечения к возможностям имеющихся программно-аппаратных средств (см., например, [4]).

В докладе представлена среда (рис. 2), ориентированная на исследование различных аспектов живучести как свойств энергетической инфраструктуры. В настоящее время с помощью данной среды реализованы различные виды анализа уязвимости, которая отражает «пассивную» реакцию энергетической инфраструктуры в виде падения производительности при возникновении экстремальных условий. В качестве ядра среды выступает инструментальный комплекс Orlando Tools [5]. Данный комплекс используется для разработки распределенных пакетов прикладных программ для исследования живучести энергетических инфраструктур, интеграции ресурсов разных вычислительных кластеров в единую среду и управления вычислениями в этой среде. Общим для различных пакетов исследования живучести является использование технологии In-Memory Data Grid (IMDG). IMDG-кластеры достигают высокой скорости обработки и большой степени масштабируемости за счет хранения данных полностью в оперативной памяти и распределения данных между несколькими серверами.

В качестве иллюстративного примера мы рассматриваем эксперимент по оценке масштабируемости вычислений при различных способах выделения узлов для вычислений и записи в базу данных Apache Ignite. Мы использовали 2, 4, 6 и 8 узлов НРС-кластера для сравнения следующих трех способов распределения узлов: W1 – выделение только одного узла для записи в базу данных Apache Ignite и распределение остальных узлов для вычислений; W2 – выделение равного числа узлов для записи в базу данных и узлов для вычислений; W3 – одновременное использование узлов для записи в базу данных и вычислений. Каждый узел имеет следующие характеристики: 2x16 cores CPU AMD Opteron 6276, 2.3

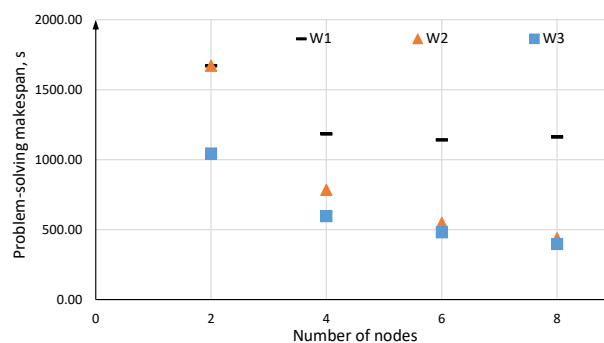


Рис. 3. Время решения задачи

GHz, 16 MB L3 cache, 4 FLOP/cycle, 64 GB RAM DDR3-1600. Результаты расчетов представлены на рис. 3. В качестве отправной точки рассматриваются 2 узла. Не наблюдается масштабируемости вычислений в отношении сокращения времени решения задач за счет увеличения числа узлов при использовании W1. Очевидно, что увеличение числа узлов с базой данных Apache Ignite сокращает время решения задач. Время решения задачи улучшается, когда число узлов с базой данных Apache Ignite равно числу вычислительных узлов согласно W2. Наилучшие результаты достигаются при использовании W3. Именно этот способ распределения ресурсов используется в Orlando Tools.

### 3. ЗАКЛЮЧЕНИЕ

В докладе предложен предметно-ориентированный подход к организации высокопроизводительных вычислений для исследования живучести энергетических инфраструктур с использованием технологии IMDG. Разработка и применение распределенных пакетов прикладных программ для исследования живучести осуществляется в Orlando Tools.

### БЛАГОДАРНОСТИ

Работа выполнена при финансовой поддержке Министерства науки и высшего образования Российской Федерации, проект № FWEW-2021-0005 «Технологии разработки и анализа предметно-ориентированных интеллектуальных систем группового управления в недетерминированных распределенных средах» с использованием ресурсов Иркутского суперкомпьютерного центра СО РАН» [6].

### ЛИТЕРАТУРА

- [1] Воропай, Н.И. Надежность систем энергетики. (Сборник рекомендуемых терминов) / Н.И. Воропай. – М. : ИАЦ “Энергия,” 2007. – 194 с.
- [2] Poulin, C.R., Kane M.B. Infrastructure resilience curves: Performance measures and summary metrics. Reliability Engineering & System Safety. 2021. – Vol. 216. – P. 107926.
- [3] Dobson, I. Complex systems analysis of series of blackouts: Cascading failure, critical points, and self-organization / I. Dobson, B.A. Carreras, V.E. Lynch, D.E. Newman // Chaos: An Interdisciplinary Journal of Nonlinear Science. – 2007. – Vol. 17(2). – P. 026103.
- [4] Gorton, I. A high-performance hybrid computing approach to massive contingency analysis in the power grid / I. Gorton, Z. Huang, Y. Chen, B. Kalahar, S. Jin, // 2009 Fifth IEEE International Conference on e-Science. – 2009. – P. 277-283.
- [5] Gorsky, S. Orlando Tools: Supporting High-performance Computing in Distributed Environments / S. Gorsky, R. Kostromin, A. Feoktistov, I. Bychkov // Proceedings of the 6th International Conference on Information Technology and Nanotechnology (ITNT 2020). – 2020. – P. 1-6.
- [6] ЦКП Иркутский суперкомпьютерный центр СО РАН [Электронный ресурс]. – Режим доступа: – <http://hpc.icc.ru> (02.11.2022).

# Метод анализа нестационарных сигналов на основе декомпозиции данных и вейвлет-преобразования

Б.С. Мандрикова

Институт космических  
исследований и распространения  
радиоволн ДВО РАН

Камчатский край, Паратунка, Россия  
555bs5@mail.ru

О.И. Есиков

Санкт-Петербургский  
государственный  
электротехнический университет  
«ЛЭТИ» имени В. И. Ульянова  
(Ленина)

Санкт-Петербург, Россия  
oiesikov@stud.etu.ru

**Аннотация** — Предложен новый автоматизированный метод анализа нестационарных сигналов сложной структуры. Метод включает операции анализа сингулярного спектра, дискретного вейвлет-преобразования и пороговых функций. Представлен алгоритм численной реализации метода. Показано применение метода к данным нейтронных мониторов (вторичные космические лучи). Эмпирически доказано, что совмещение декомпозиции данных с вейвлет-преобразованием позволяет детектировать аномалии в сигнале космических лучей. Результат важен для прогноза космической погоды.

**Ключевые слова** — нестационарный сигнал, аномалии, вейвлет-преобразование, анализ сингулярного спектра, нейтронные мониторы.

## 1. ВВЕДЕНИЕ

На протяжении всего периода своего существования человечество непрерывно совершало технический прогресс. В настоящий момент для решения практических задач, многие из которых уже давно стали повседневными, люди активно используют различные технические объекты и средства связи. Работоспособность объектов программно-аппаратной инфраструктуры, а также поддержка бесперебойной связи во многом зависит от состояния космической погоды [1]. Негативные воздействия космической погоды подвергают опасности громадное количество различных объектов наземной и космической инфраструктуры [2]: системы телесвязи, радиосвязи, спутниковой связи, нефтепроводы и газопроводы, линии электропередач, спутники, системы позиционирования GPS, ГЛОНАСС, Galileo, а также могут повлечь сбой в работе электроники [2, 3, 4].

Значимым фактором космической погоды является интенсивность космических лучей (КЛ). Сигнал КЛ может включать регулярные (периодические) и аномальные (непериодические) вариации [5]. Последние, как правило, наблюдаются в периоды магнитосферных возмущений. Поскольку сигнал КЛ является нестационарным, содержит шумы аппаратного и природного происхождения, детектирование аномальных вариаций является особо актуальной и сложной задачей. На данный момент ещё не разработано математического аппарата, который позволял бы с удовлетворительной точностью и эффективностью определять наступление таких событий [6].

В докладе представлен новый метод детектирования аномальных вариаций в данных КЛ на основе

применения операций анализа сингулярного спектра, дискретного вейвлет-преобразования и пороговых функций. Анализ сингулярного спектра не зависит от типа значимых компонент ряда, позволяет исследовать нестационарные временные ряды и выполнить очистку сигнала от шумовых составляющих [7]. Вейвлет-преобразование позволяет провести детальный частотно-временной анализ нестационарного сигнала [8]. Для снижения риска наступления ложной тревоги предложено применение адаптивных пороговых функций.

## 2. ОПИСАНИЕ МЕТОДА

Предлагаемый метод анализа нестационарных сигналов на основе декомпозиции данных и вейвлет-преобразования включает следующие операции:

1. Преобразование исходного одномерного ряда в траекторную матрицу:

$$X_i = (f_{i-1}, \dots, f_{i+L-2})^T, 1 \leq i \leq N - L + 1,$$

где  $f_i$  – элемент исходного ряда,  $L$  – длина окна,  $N$  – длина исходного ряда.

2. Сингулярное разложение траекторной матрицы:

$$X = \sum_i \sqrt{\lambda_i} U_i V_i^T,$$

где  $\sqrt{\lambda_i}$  – сингулярное число,  $U_i$  – левый сингулярный вектор траекторной матрицы,  $V_i$  – правый сингулярный вектор траекторной матрицы.

3. Группировка множества индексов на  $m$  непересекающихся подмножеств. Результирующая матрица, соответствующая группе:

$$X_i = X_{i_1} + \dots + X_{i_p},$$

где  $\{i_1, \dots, i_p\}$  – индексы группы.

4. С помощью применения диагонального усреднения [9] каждая результирующая матрица сгруппированного разложения преобразуется в новый ряд  $F^{(s)}$  длины  $N$ . Восстановленный ряд определяется следующим образом:

$$F = \sum_i F_i^{(s)}, \quad (1)$$

5. Применение непрерывного вейвлет-преобразования

$$WF(u, s) = \int_{-\infty}^{+\infty} F \frac{1}{\sqrt{s}} \Psi^* \left( \frac{t-u}{s} \right) dt, \quad (2)$$

где  $\Psi$  – вейвлет,  $u$  – сдвиг во времени,  $s$  – масштаб,  $s \neq 0$ ,  $s, u \in R$ .

6. Применение пороговой функции:

$$P_{T_s^l}[WF(u, s)] = \begin{cases} WF(u, s), & |WF(u, s)| \geq T_s^l \\ 0, & |WF(u, s)| < T_s^l \end{cases}, \quad (3)$$

где  $T_s^l = q \times \sigma_s^l$ ,  $\sigma_s^l$  – среднеквадратическое отклонение коэффициентов, рассчитанное в скользящем окне длины  $l$ ,  $q$  – пороговый коэффициент.

7. Оценка интенсивности аномалий в момент времени  $t = u$ :

$$E(u) = \sum_s P_{T_s^l}[WF(u, s)]. \quad (4)$$

### 3. РЕЗУЛЬТАТЫ ПРИМЕНЕНИЯ МЕТОДА

На Рис. 1 представлен результат применения метода к данным нейтронного монитора (НМ) ст. Оулу за период с 8 по 29 марта 2022 г. На Рис. 1а изображены исходные данные НМ [10], на Рис. 1б представлен результат операции (1) при  $I = \{i_1, \dots, i_2\}$ , на Рис. 1в и 1г показаны результаты применения операций (2) и (3) соответственно. Красными вертикальными линиями отмечены моменты регистрации геомагнитных бурь по данным [11].

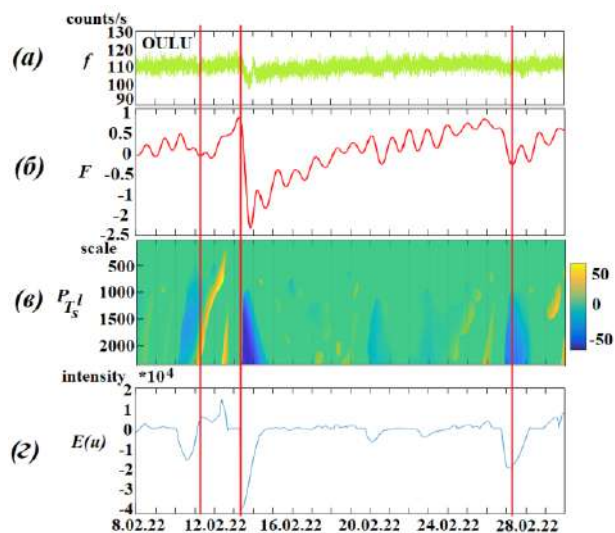


Рис. 1. Результат применения метода: (а) данные НМ ст. Оулу, (б) результаты применения операции (1) при  $I = \{i_1, i_2, i_3\}$ , (в) результаты применения операций (2) и (3), (г) результаты применения операции (4)

Результаты показывают, что предлагаемый метод позволяет детектировать аномалии разной частотно-

временной структуры в данных КЛ. По исходным данным нейтронного монитора визуально определить момент наступления геомагнитных бурь 11 и 27 марта не представляется возможным. Применение метода позволяет не только детектировать момент начала геомагнитной бури (Рис. 1в), но и оценить ее интенсивность и продолжительность (Рис. 1г). Результат демонстрирует эффективность метода для анализа вторичных космических лучей, а также подтверждает важность учета интенсивности КЛ в прогнозе космической погоды.

### БЛАГОДАРНОСТИ

Работа выполнена в рамках ГЗ по теме “Физические процессы в системе ближнего космоса и геосфер при солнечных и литосферных воздействиях” (2021–2023 гг.), регистрационный номер АААА-А21-121011290003-0.

### ЛИТЕРАТУРА

- [1] Владимирский, Б.М. Космическая погода и наша жизнь / Владимирский, Б.М., Темурьян Н.А., Мартынюк В.С. – Век 2, 2004. – 224 с.
- [2] Кузнецов, В.Д. Космическая погода и риски космической деятельности / В.Д. Кузнецов // Космическая техника и технологии. – 2014. – Т. 3, №6. – С. 3-13.
- [3] Авакян, С.В. Влияние магнитных бурь на аварийность систем электроэнергетики, автоматики и связи / С. В. Авакян, Н. А. Воронин, К. А. Дубаренко // Материаловедение. Энергетика. – 2012. – Т. 3-2, №154. – С. 253-266.
- [4] Severe Space Weather Events — Understanding Societal and Economic Impacts / A Workshop Report. Washington DC. The National Academies Press, – 2009.
- [5] Топтыгин, И.Н. Космические лучи в межпланетных магнитных полях / И. Н. Топтыгин. – М.: Наука, 1983. – 304 с.
- [6] Акасофу, С.И. Солнечно-земная физика / С.И. Акасофу, С. Чепмен. – М.: Мир, 1972. – 384 с.
- [7] Кашкин, В. Б. Применение сингулярного спектрального анализа для выделения слабо выраженных трендов / В. Б. Кашкин, Т. В. Рублева // Известия Томского политехнического университета. Инжиниринг георесурсов. – 2007. – Т. 311, №. 5. – С. 116-119.
- [8] Каплинский, А.Е. Анализ временных рядов наблюдений характеристик байкальского аэрозоля с помощью вейвлет-преобразования / А.Е. Каплинский, О.Г. Хуторова // Ползуновский вестник. – 2010. – Т. 1. – С. 160-164.
- [9] Голяндина, Н. Э. Варианты метода «Гусеница»-SSA для анализа многомерных временных рядов / Н. Э. Голяндина, В. В. Некруткин, Д. Степанов // Труды II Международной конференции «Идентификация систем и задачи управления» SICPRO. – 2003. – Т. 3. – С. 2139-2168.
- [10] Real Time DB of NM. [Electronic resource]. — Access mode: www.nmdb.eu (01.11.2022).
- [11] Forecast of space weather according to the data of Federov IAG. [Electronic resource]. — Access mode: http://ipg.geospace.ru (01.11.2022).

# Построение и идентификация параметров дискретной стохастической модели годового хода температуры воздуха

М.А. Шугурова

Ульяновский государственный педагогический университет им.

И.Н. Ульянова

Ульяновск, Россия

m.a.shugurova@gmail.com

А.В. Цыганов

Ульяновский государственный педагогический университет им.

И.Н. Ульянова

Ульяновск, Россия

andrew.tsyganov@gmail.com

**Аннотация**—В работе рассматривается задача построения математической модели годового хода температуры воздуха в классе линейных дискретных стохастических систем в пространстве состояний. Идентификация параметров построенной модели выполняется с использованием методов рекуррентной дискретной фильтрации. Приводятся результаты численных экспериментов по идентификации неизвестных параметров на основе данных атмосферных реанализов.

**Ключевые слова**—годовой ход температуры воздуха, линейные дискретные стохастические системы, параметрическая идентификация, фильтр Калмана, атмосферные реанализы

## 1. ВВЕДЕНИЕ

Температуру воздуха принято считать одной из основных характеристик климата и погоды. Важной характеристикой изменения температуры воздуха является годовой ход температуры воздуха – изменение температуры воздуха в течение года. Математическое и компьютерное моделирование годового хода температуры воздуха находит применение в ЖКХ, строительстве, сельском хозяйстве. Например, различные модели годового хода температуры воздуха могут использоваться при расчетах воздушно-теплого режима помещений зданий и оценке годового энергопотребления [1].

Целью данной работы является построение математической модели годового хода температуры воздуха в классе линейных дискретных стохастических систем в пространстве состояний и идентификация параметров построенной модели с использованием методов рекуррентной дискретной фильтрации.

## 2. МОДЕЛИРОВАНИЕ ГОДОВОГО ХОДА ТЕМПЕРАТУРЫ ВОЗДУХА И ИДЕНТИФИКАЦИЯ ПАРАМЕТРОВ МОДЕЛИ

Важным источником данных для изучения климата являются атмосферные реанализы – динамически разглаженные и согласованные данные определенного набора архивных наблюдений, полученные при помощи гидродинамической модели с фиксированной конфигурацией [2]. Данные многолетних наблюдений для разных местностей показывают, что годовой ход температуры воздуха носит ярко выраженный периодический характер, в котором могут быть выделены две составляющие: детерминированная (тренд), в основе которой лежат движение по орбите и наклон оси вращения Земли, и стохастическая, обусловленная множеством случайных факторов. В качестве примера рассмотрим данные реанализа NCEP [3] среднесуточной температуры воздуха на высоте 2 м в

узле гауссовой сетки T62 с координатами (21, 18) за 2018–2020 гг., представленные на рис. 1.

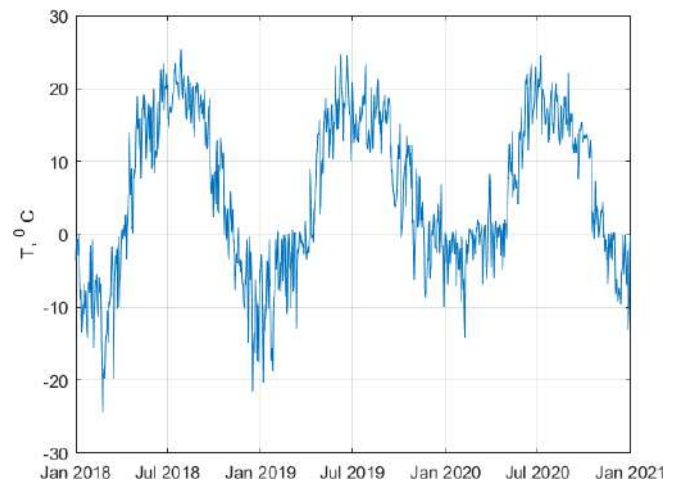


Рис. 1. Данные реанализа температуры воздуха

Для построения модели будем использовать подход, предложенный в [4] для моделирования данных суточной термометрии здорового человека. Разделим рассматриваемый процесс на следующие аддитивные составляющие: 1)  $\bar{\theta}_t$  – математическое ожидание отклонения температуры от среднегодового уровня  $\theta^*$ , 2)  $\tilde{\theta}_t \triangleq \{\tilde{\theta}_t(\omega)\}$  – стохастический процесс с нулевым средним ( $\omega$  – произвольная точка выборочного пространства  $\Omega$ ), 3)  $\hat{\theta}_t \triangleq \bar{\theta}_t + \tilde{\theta}_t$ , для которого  $d\hat{\theta}_t = d\bar{\theta}_t$ . Тогда  $\theta_t \triangleq \bar{\theta}_t + \tilde{\theta}_t$  – суммарный процесс, моделирующий рассматриваемые данные.

Предположим, что тренд  $\bar{\theta}_t$  удовлетворяет уравнению гармонического осциллятора:

$$\begin{cases} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix}_t = \begin{bmatrix} 0 & 1 \\ -\omega_n^2 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, t \in [0; \infty), \\ z_t = [1 \quad 0]x_t, \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}_0 = \begin{bmatrix} A \sin \varphi \\ A \omega_n \cos \varphi \end{bmatrix}, \end{cases} \quad (1)$$

где

$$\bar{\theta}_t \triangleq x_{1t} = A \sin(\omega_n t + \varphi), \quad \bar{\omega}_t \triangleq x_{2t}, \quad A = \sqrt{\bar{\theta}_0^2 + \left(\frac{\bar{\omega}_0}{\omega_n}\right)^2},$$

$$\sin \varphi = \bar{\theta}_0/A, \quad \cos \varphi = \left(\frac{\bar{\omega}_0}{\omega_n}\right)/A, \quad \operatorname{tg} \varphi = \bar{\theta}_0/\left(\frac{\bar{\omega}_0}{\omega_n}\right).$$

Пусть  $\tilde{\theta}_t$  – экспоненциально коррелированный по времени случайный процесс с автокорреляцией  $\Psi_{\tilde{\theta}\tilde{\theta}}(\tau) \triangleq \sigma^2 e^{-|\tau|/T}$ , где  $T$  – интервал корреляции. Обозначим  $\lambda = 1/T$ ,  $\eta = \sigma\sqrt{2\lambda}$ . Добавляя к (1) третью переменную  $\tilde{\theta}_t$ , получим следующую модель:

$$\begin{cases} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix}_t = \begin{bmatrix} 0 & 1 & 0 \\ -\omega_n^2 & 0 & 0 \\ 0 & 0 & -\lambda \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}_t + \begin{bmatrix} 0 \\ 0 \\ \lambda \end{bmatrix} u_t + \begin{bmatrix} 0 \\ 0 \\ \eta \end{bmatrix} w_t, \\ z_t = [1 \ 0 \ 1]x_t + v_t, \quad \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}_0 = \begin{bmatrix} A \sin \varphi \\ A \omega_n \cos \varphi \\ 0 \end{bmatrix}. \end{cases} \quad (2)$$

Дискретизируя модель (2), получаем:

$$\begin{cases} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}_{k+1} = \begin{bmatrix} c & s & 0 \\ -\omega_n s & c & 0 \\ 0 & 0 & d \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}_k + \begin{bmatrix} 0 \\ 0 \\ a \end{bmatrix} u_k + \begin{bmatrix} 0 \\ 0 \\ b \end{bmatrix} w_k, \\ z_k = [1 \ 0 \ 1]x_k + v_k, \quad \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}_0 = \begin{bmatrix} A \sin \varphi \\ A \omega_n \cos \varphi \\ 0 \end{bmatrix}, \\ k = 1, 2, \dots, \end{cases} \quad (3)$$

где

$$d \triangleq e^{-\lambda\tau}, \quad a \triangleq 1 - d, \quad b \triangleq \sigma\sqrt{1-d^2}, \quad c \triangleq \cos \omega_n\tau, \quad s \triangleq \sin \omega_n\tau.$$

Здесь  $x_k$  – вектор состояния,  $u_k$  – управляющее воздействие,  $z_k$  – измерения,  $\tau$  – период дискретизации,  $w_k$  и  $v_k$  – независимые нормально распределенные случайные последовательности с нулевыми математическими ожиданиями и ковариационными матрицами  $Q = 1$  и  $R > 0$  соответственно).

Обозначим через  $\theta = [A, \varphi, \lambda, \sigma]^T$  вектор неизвестных параметров модели (3). Для идентификации неизвестных параметров используем критерий идентификации в форме отрицательной логарифмической функции правдоподобия

$$J_k(\theta) = \frac{Km}{2} \ln 2\pi + \frac{1}{2} \sum_{k=1}^K \left\{ \ln |\Sigma_k(\theta)| + \|v_k(\theta)\|_{\Sigma_k^{-1}(\theta)}^2 \right\}, \quad (4)$$

значения которого при заданном  $\theta$  вычисляют с помощью дискретного фильтра Калмана. Критерий (4) находит широкое применение для решения задач параметрической идентификации дискретных стохастических систем в пространстве состояний [5].

### 3. ЧИСЛЕННЫЙ ЭКСПЕРИМЕНТ

Идентифицируем параметры модели (3) по данным реанализа, представленным на рис. 1. Положим  $\tau = 1$  день,  $\omega_n = 2\pi/365,25$  день<sup>-1</sup>. В качестве управляющего воздействия  $u_k$  возьмем среднее значение температуры воздуха за рассматриваемый период равно 5,54 °С. Для обработки данных реанализов и идентификации неизвестных параметров модели (3) использовалась программа [6]. Дополнительно в программе были реализованы: критерий идентификации (4) и процедура его минимизации на основе функции fmincon системы MATLAB.

В результате идентификации получены следующие значения:  $A = 13,27$ ,  $\varphi = 180,29$ ,  $\lambda = 0,26$ ,  $\sigma = 4,52$ . На рис. 2 приведен пример компьютерного моделирования годового хода температуры воздуха при помощи модели (3) с использованием идентифицированных значений параметров.

Для проверки адекватности построенной модели была проведена серия из 1000 экспериментов по моделированию годового хода температуры воздуха в выбранном диапазоне дат. Результаты каждого эксперимента сравнивались с данными реанализа при помощи теста Колмогорова-Смирнова для двух выборок с уровнем значимости  $\alpha = 0,001$ . Общее количество

принятых гипотез о принадлежности выборок одному распределению составило 91%.

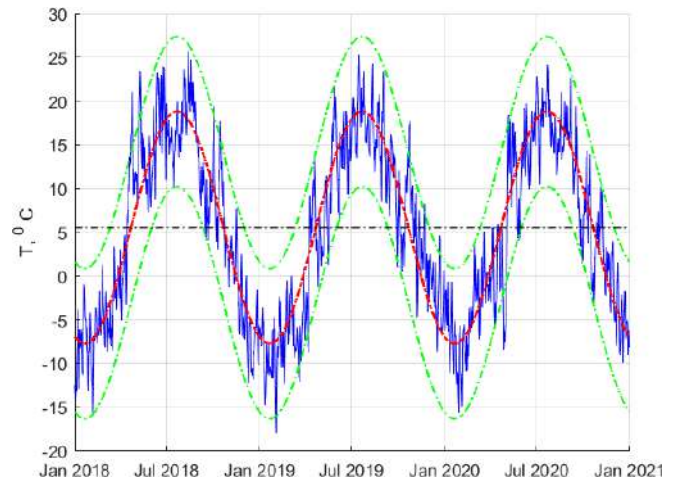


Рис. 2. Результаты компьютерного моделирования

### 4. ЗАКЛЮЧЕНИЕ

В работе рассмотрен подход к моделированию годового хода температуры воздуха в классе линейных дискретных стохастических систем в пространстве состояний. Идентификация параметров построенной модели выполнялась с использованием методов рекуррентной дискретной фильтрации на основе данных атмосферных реанализов. Полученные результаты демонстрируют работоспособность предложенного подхода.

### ЛИТЕРАТУРА

- [1] Самарин, О.Д. Вероятностно-статистическое моделирование годового хода температуры наружного воздуха и ее значений в теплый период / О.Д. Самарин // Вестник МГСУ. — 2018. — Т. 13, № 3. — С. 378–384.
- [2] Гавриков, А. Атмосферные реанализы [Электронный ресурс]. – Режим доступа: [https://ocean.ru/phocadownload/pl\\_univer/pl\\_univer\\_2019\\_01.pdf](https://ocean.ru/phocadownload/pl_univer/pl_univer_2019_01.pdf) (21.12.2022).
- [3] NOAA Physical Sciences Laboratory NCEP/DOE Reanalysis II [Electronic resource]. — Access mode: <https://psl.noaa.gov/data/gridded/data.ncep.reanalysis2.html> (06.11.2022)
- [4] Semushin, I.V. Identification of a Simple Homeostasis Stochastic Model Based on Active Principle of Adaptation / I.V. Semushin, J.V. Tsyganova, A.G. Skovikov // Proceedings of International Conference Applied Stochastic Models and Data Analysis ASMDA 2013 & DEMOGRAPHICS 2013. — 2013. — P. 775–783.
- [5] Gibbs, B. P. Advanced Kalman filtering, least-squares and modeling: a practical handbook / B. P. Gibbs. — Hoboken, New Jersey : John Wiley & Sons, Inc., 2011. — 632 p.
- [6] Шугурова, М.А. Программа для чтения и обработки данных атмосферных реанализов / М.А. Шугурова, Д. В. Галушкина // Ученые записки УлГУ. Сер. Математика и информационные технологии. УлГУ. Электрон. журн. — 2021. — Т. 1. — С. 137–143.

# Метод обнаружения аномалий в природных данных на основе нейронных сетей и вейвлет-фильтрации

О.В. Мандрикова  
Институт космических исследований и распространения радиоволн ДВО РАН  
Паратунка, Россия  
oksanam1@mail.ru

Ю.А. Полозов  
Институт космических исследований и распространения радиоволн ДВО РАН  
Паратунка, Россия  
up\_agent@mail.ru

**Аннотация** — Предложен автоматизированный метод анализа природных данных и обнаружения аномалий, основанный на совмещении операций вейвлет-фильтрации с нейронной сетью NARX. Построен алгоритм вейвлет-фильтрации и способ оценки порогов, основанный на стохастическом подходе. Показано, что применение вейвлет-фильтрации подавляет шум, упрощает структуру данных и, как следствие, позволяет получить более точную нейросетевую модель NARX. На примере ионосферных данных показана эффективность метода для обнаружения ионосферных аномалий в периоды магнитных бурь.

**Ключевые слова**— анализ данных, вейвлеты, нейронные сети, ионосфера

## 1. ВВЕДЕНИЕ

Задача обработки и анализа природных данных важна для изучения процессов и явлений разной природы и актуальна в различных сферах человеческой деятельности (физика, биология, медицина, экономика и др.). Особо актуальны методы, направленные на диагностику состояний объектов и обнаружение аномалий [1], [2]. В настоящее время для таких задач активно развиваются гибридные подходы и методы [1], [2], которые позволяют повысить качество процедуры анализа данных. В работе предложен метод анализа природных данных, основанный на совместном применении операций вейвлет-фильтрации и нейронных сетей (НС) NARX [3]. Вейвлет-преобразование является гибким инструментом и широко используется в задачах обработки и анализа данных. Обширная библиотека вейвлет-фильтров и широкий набор конструкций разложения обеспечивают возможность адаптации этого инструмента для данных разной структуры [4]. Сети NARX выполняют аппроксимацию временных рядов на основе «моделей нелинейной авторегрессии с экзогенными входами» [3]. К очевидным преимуществам регрессионных моделей относятся их математическая обоснованность, формализованная методика идентификации модели и её проверки на адекватность. В данной работе полученные после вейвлет-фильтрации временные ряды подаются на вход сети NARX. Процедура вейвлет-фильтрации подавляет шум, упрощает структуру данных и повышает эффективность нейронной сети NARX. В работе предложен алгоритм вейвлет-фильтрации и способ оценки порогов. Приведена схема решения задачи обнаружения аномалий.

В качестве экспериментальных данных используются временные ряды критической частоты ионосферного слоя F2 (foF2). Ионосферные временные ряды имеют регулярный ход, а также аномалии разной формы и временной протяженности, которые наблюдаются в

периоды возмущений в околоземном пространстве. Применяемые традиционные методы анализа ионосферных данных не достаточно эффективны для обнаружения ионосферных аномалий [5]. В работе демонстрируется эффективность предлагаемого метода для обнаружения ионосферных аномалий разной частотно-временной структуры. Работа является продолжением исследования [6].

## 2. ОПИСАНИЕ МЕТОДА

В работе использовались данные критической частоты ионосферы foF2, имеющие дискретизацию один час. Регистрация данных выполняется на станции «Паратунка» (Камчаткий край, Россия, координаты станции: 53.0 N, 158.7 E) с 1969 года.

Операции подавления шума включают применение конструкции кратномасштабного анализа [4] и пороговой функции. Используя работу [7], применены стохастические пороги. Алгоритм подавления шума :

1. Вейвлет-разложение сигнала  $f_0(t)$  на компоненты:

$$f_0(t) = \sum_{j=-1}^{-m} g_j(t) + f_{-m}(t),$$

где  $f_{-m}(t) = \sum_k c_{-m,k} \phi_{-m,k}(t)$  – сглаженная компонента,  $c_{-m,k} = \langle f_0, \phi_{-m,k} \rangle$ ,  $\phi_{-m,k}(t) = 2^{-m/2} \phi(2^{-m}t - k)$  – скейлинг-функция,  $g_j(t) = \sum_k d_{j,k} \Psi_{j,k}(t)$  – детализирующие компоненты,  $d_{j,k} = \langle f_0, \Psi_{j,k} \rangle$ ,  $\Psi_{j,k}(t) = 2^{j/2} \psi(2^j t - k)$  – вейвлет,  $j$  – уровень разложения, для исходного сигнала предполагается уровень разложения  $j = 0$ .

2. Применение пороговой функции к коэффициентам компонент  $g_j(t)$ :

$$T(d_{j,k}) = \begin{cases} 0, & \text{если } |d_{j,k}| \leq T_j \\ d_{j,k}, & \text{если } |d_{j,k}| > T_j \end{cases}, \quad (1)$$

где  $T_j = t_{1-\frac{\alpha}{2}, N-1} \hat{\sigma}_j$ ,  $t_{\alpha, N} - \alpha$ -квантили распределения Стьюдента,  $\hat{\sigma}_j$  – выборочное стандартное отклонение, уровни разложения  $j = -1, -m$ .

3. Вейвлет-восстановление сигнала:

$$\tilde{f}_0(t) = \sum_{j,k} T(d_{j,k}) \Psi_{j,k}(t) + f_{-m}(t).$$

Пороги  $T_j$  в (1) определяются как  $T_j = t_{1-\frac{\alpha}{2}, N-1} \hat{\sigma}_j$ , где  $t_{\alpha, N} - \alpha$ -квантили распределения Стьюдента.

Для построения нейросетевой модели использовались сети архитектуры NARX с обратными связями. В сети использовались блоки линий задержки

входа и выхода  $l_f = l_f$ , значения которых подаются на нейроны скрытого слоя, что позволяет регулировать глубину ретроспективного анализа. Значение выхода нейронной сети имеет вид:

$$\hat{f}_0(t+1) = F[\tilde{f}_0(t), \tilde{f}_0(t-1), \dots, \tilde{f}_0(t-l_x), \hat{f}_0(t), \hat{f}_0(t-1), \dots, \hat{f}_0(t-l_y)].$$

где  $F(\cdot)$  - функция отображения нейронной сети.

### 3. РЕЗУЛЬТАТЫ ПРИМЕНЕНИЯ МЕТОДА ДЛЯ ДАННЫХ ИОНОСФЕРЫ

На рис. 1 показаны результаты обработки данных в период магнитной бури 5-6 августа 2019 г. Начало бури указано красной вертикальной линией. Для анализа уровня геомагнитной активности на рис. 1е показаны

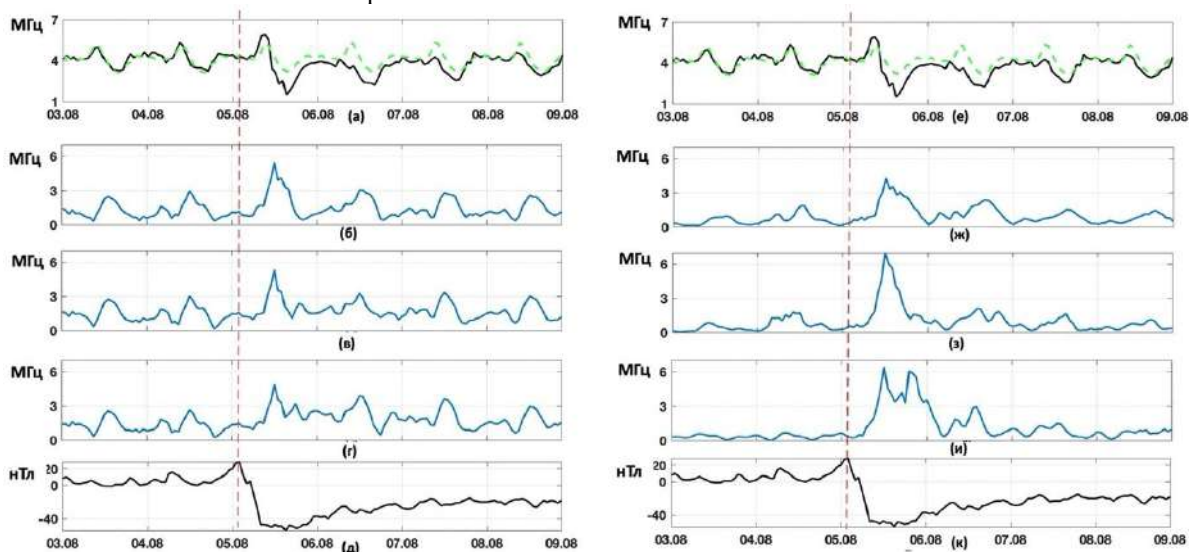


Рисунок 1. Результаты обработки данных за период 03 – 08 августа 2019 года. (а),(е) – исходные данные foF2 (черн.) и медиана foF2 (зел.); (б) - (г) – ошибки НС с задержками 2, 3 и 5, соответственно, полученные без применения вейвлет-фильтрации; (ж) - (з) – ошибки с задержками 2, 3 и 5, соответственно, полученные с применением вейвлет-фильтрации; (д), (к) – DST. Красный пунктир – начало магнитной бури.

### 4. ЗАКЛЮЧЕНИЕ

Применение метода показало его эффективность в задаче моделирования и анализа ионосферных данных. Предлагаемая процедура вейвлет-фильтрации позволяет повысить качество работы нейронных сетей NARX и дает возможность получить адекватную модель для зашумленных и нестационарных данных.

На примере магнитной бури 5-6 августа 2019 г., подтверждена возможность метода для обнаружения ионосферных аномалий по данным foF2 во время магнитосферных возмущений. Метод может быть применен для мониторинга состояния ионосферы при выполнении прогноза космической погоды.

### БЛАГОДАРНОСТИ

Работа выполнена в рамках ГЗ «Физические процессы в системе ближнего космоса и геосфер при солнечных и литосферных воздействиях», регистрационный номер: АААА-А21-121011290003-0. В работе использовалось оборудование Центра коллективного пользования «Северо-восточный гелиогеофизический центр» СКР\_558279.

### ЛИТЕРАТУРА

[1] Wu, X. The development of a hybrid wavelet-ARIMA-LSTM model for precipitation amounts and drought analysis / X. Wu, J. Zhou, H. Yu, D. Liu, K. Xie, Y. Chen, J. Hu, H. Sun, F. Xing // Atmosphere. – 2021. – Vol. 12(1). – P. 74. DOI: 10.3390/atmos12010074.

значения DST индекса геомагнитной активности. Оцененные значения скользящей медианы (рис. 1а) показывают длительные изменения во временном ходе данных foF2 во время бури. Анализ ошибок НС подтверждает рост ошибок в период события, что свидетельствует о возникновении аномалий в данных. Сравнение результатов НС без вейвлет-фильтрации (рис. 5б-г) и с вейвлет-фильтрацией (рис. 5ж-и) показывает значительное уменьшение ошибок НС на основе предложенного в работе подхода. Наилучшие результаты показывает НС с линиями задержки  $l_f = l_f = 5$  (рис. 1з), которая имеет наименьшие ошибки и четко детектирует аномальный период в ионосферных данных.

[2] Sebastian, D.E. Multi-scale association between vegetation growth and climate in India: A wavelet analysis approach / D.E. Sebastian, S. Ganguly, J. Krishnaswamy, K. Duffy, R. Nemani, S. Ghosh // Remote Sensing. – 2019. – Vol. 11(22). – P. 2703. DOI: 10.3390/rs11222703.

[3] Haykin, S. S. Neural networks: a comprehensive foundation / S. S. Haykin, 2nd ed-e. – Upper Saddle River, N.J: Prentice Hall, 1999. – 842 p.

[4] Mallat, S. G. A wavelet tour of signal processing / S. G. Mallat. – San Diego: Academic Press, 1999. – 620 p.

[5] Danilov, A.D. Ionospheric F-region response to geomagnetic disturbances / A.D. Danilov // Advances in Space Research. – 2013. – Vol. 52(3). – P. 343–366. DOI: /10.1016/j.asr.2013.04.019.

[6] Mandrikova, O. Hybrid Model for time series of complex structure with ARIMA components / O. Mandrikova, N. Fetisova, Y. Polozov // Mathematics. – 2021. – Vol. 10(9). – P. 1122. DOI: 10.3390/math9101122.

[7] Mandrikova, O. Hybrid method for detecting anomalies in cosmic ray variations using neural networks autoencoder / O. Mandrikova, B. Mandrikova // Symmetry. – 2022. – Vol. 14(4). – P. 744. DOI: 10.3390/math9070737.

# Применение метода активных контуров в задачах цефалометрии

Ю.Ж. Пчелкина  
Самарский национальный  
исследовательский  
университет им.  
академика С.П. Королева  
Самара, Россия  
musina@yandex.ru

Р.А. Парингер  
Самарский национальный  
исследовательский  
университет им.  
академика С.П. Королева  
Самара, Россия  
rusparinger@gmail.com

А.В. Куприянов  
Самарский национальный  
исследовательский  
университет им.  
академика С.П. Королева  
Самара, Россия  
akupr@ssau.ru

П.Е. Савельева  
Московский областной  
научно-  
исследовательский  
клинический институт  
им. М.Ф. Владимирского  
Москва, Россия  
gezulya76@yandex.ru

**Аннотация** — Алгоритмы активных контуров были применены для автоматизации процесса определения цефалометрических признаков профиля лица по фотоснимкам. Исследовались такие признаки как тип профиля и гармоничность профиля. Произведен анализ результатов автоматической разметки данных.

**Ключевые слова** — обработка изображений, активные контуры, сегментация изображений, ортодонтия, цефалометрия

## 1. ВВЕДЕНИЕ

При проведении цефалометрического анализа изображений врачами ортодонтами производится антропометрическое и фотометрическое исследование головы пациента. При заполнении описательной части фотопротокола [1] кроме прочего производят изучение формы и эстетики лица, аномалий челюстно-лицевой области. На сегодняшний день определение ключевых цефалометрических точек и расчет антропометрических характеристик производится врачом вручную. Автоматизация этого процесса позволит сократить время осмотра и диагностики.

## 2. ОПИСАНИЕ МЕТОДА

Для определения профиля лица измеряется угол между прямыми, проходящими характерные точки: наиболее выступающая точка лба, точка, соединяющая кожную перегородку носа с верхней губой, наиболее выступающая точка подбородка.

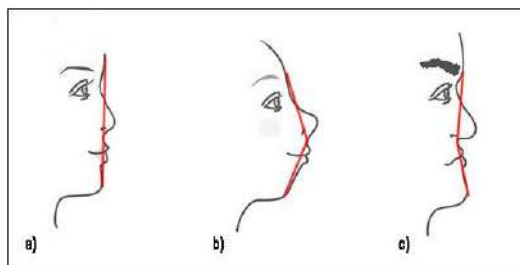


Рис. 1. Типы профиля лица: а) прямой профиль; б) выпуклый профиль; в) вогнутый профиль

При антропометрическом анализе выделяют три типа профиля: прямой профиль – угол между прямыми равен 180 градусов; выпуклый профиль – угол между прямыми меньше 180 градусов; вогнутый профиль – угол между прямыми больше 180 градусов.

Эстетическая линия профиля лица по Ricketts – это линия, соединяющая кончик носа и наиболее

выступающую точку подбородка. Профиль считается гармонично развитым, если данная линия не пересекает кайму губ, при этом верхняя губа отстает от этой линии на 2-3 мм, нижняя губа отстает на 1-2 мм.

## 3. МЕТОД АКТИВНЫХ КОНТУРОВ

Решение задачи автоматизации определения ключевых цефалометрических характеристик основано на распознавании и анализе изображений. На сегодняшний день существует много различных способов сегментации изображений. Выбор алгоритма зависит от поставленной задачи и, как следствие, от необходимой степени точности выделения контура объекта. В случае определения ключевых цефалометрических признаков корректность и наибольшая точность выделения контура изображения играет значительную роль. Метод активных контуров является вариационным методом нахождения границ на заданном изображении [2]-[6]. Изначальный контур инициализируется как некоторая простая линия, состоящая из  $n$  точек. Контур деформируется итеративно до тех пор, пока его форма не будет достаточно близкой к форме изучаемого объекта. Для каждой точки  $v_i$  изменяемого контура значение энергии находится из уравнения:

$$E_i = \alpha \cdot E_{int}(v_i) + \beta \cdot E_{ext}(v_i)$$

$E_{int}(v_i)$  – это функция внутренней энергии, заданная как сумма функций сглаживающей энергии контура и распирающей энергии контура.  $E_{ext}(v_i)$  – это функция внешней энергии, являющаяся суммой функций энергии изображений и энергии градиента.

Весовые коэффициенты определяют вклад соответствующей энергии в общее уравнение критерия. При увеличении коэффициента сглаживающей энергии контура кривая будет сокращаться быстрее, при росте коэффициента распирающей энергии кривая будет более гладкой, коэффициент функции энергии градиента отвечает за изменение яркости изображения.

В большинстве случаев применения метода активных контуров изначально контур задается пользователем вручную. Одна из задач, решаемых в данной работе – это автоматическое выделение начального контура. Для автоматизации инициализации начального контура использовались результаты обнаружения лиц при помощи уже обученных моделей детектирования лиц на изображениях [7]-[10]. Так, например, для обнаружения лиц, изображенных в анфас или в профиль, можно воспользоваться обученной моделью каскадных

признаков Хаара. Тогда в качестве координат центра окружности изначального контура можно использовать координаты центра найденной прямоугольной области.

Ключевыми цефалометрическими точками профиля являются наиболее углубленные или наиболее выступающие точки профиля. Поэтому было предложено исследовать функцию контура профиля на экстремумы. Для этого после применения метода активных контуров была произведена развертка обнаруженного замкнутого контура профиля (путем перехода из полярных координат в декартовы).

При обратном переходе в полярные координаты Таким образом найдены координаты всех интересующих ключевых точек профиля. По ключевым точкам (наиболее выступающая точка лба, точка, соединяющая кожную перегородку носа с верхней губой, наиболее выступающая точка подбородка) были заданы уравнения прямых линий, их соединяющих, найдены углы между этими прямыми, определен тип профиля.

Через точку кончика носа и наиболее выступающую точку подбородка проведена эстетическая линия профиля лица по Ricketts, оценено положение относительно этой прямой точек верхней и нижней красной каймы губ, определена гармоничность профиля.

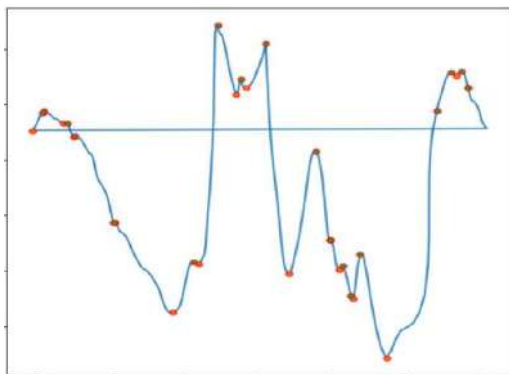


Рис. 2. Развертка контура, экстремумы

#### 4. РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЙ

Для экспериментальных исследований использовался набор из 120 изображений. Для анализа результатов автоматической разметки изображения были сведены к одному размеру и подвергнуты предобработке согласно вышеописанному методу. Все изображения были размечены двумя способами: медицинским специалистом вручную и автоматически. Были обозначены ключевые цефалометрические точки профиля, определен тип профиля, определена гармоничность профиля. В качестве метрики точности использовалось среднее расстояние между точками, размеченными вручную и точками, полученными автоматически при различных весовых коэффициентах соответствующих энергий. В качестве оценки ошибочной вероятности использовалось отношение неверно детектированных типов профиля к общему количеству изображений в выборке.

Таблица I. АНАЛИЗ АТОМАТИЧЕСКОГО ДЕТЕКТИРОВАНИЯ

коэффициенты соответствующих функций	при	метрики качества	вероятность ошибочной автоматической
--------------------------------------	-----	------------------	--------------------------------------

энергии				классификации	
сглаживающая энергия контура	распирающая энергия контура	энергия градиента		тип профиля	гармоничность профиля
0,01	0,01	0,01	29,76	0,64	0,12
0,02	0,01	0,01	25,27	0,11	0,97
0,05	0,01	0,01	37,41	0,08	0,86
0,01	0,01	0,01	26,83	0,48	0,09
0,001	0,001	0,005	74,35	0,98	0,23
0,001	0,001	0,001	03,66	0,02	0,03

Лучшие результаты автоматической разметки данных получены при значениях весовых коэффициентов: 0,001 для сглаживающей энергии, 0,001 для распирающей энергии и 0,001 для энергии градиента. В этом случае отличие результатов детектирования ключевых точек автоматически и вручную не повлияло на результаты определения типа и гармоничности профиля.

#### 5. ЗАКЛЮЧЕНИЕ

Анализ результатов автоматической разметки данных показал, что предложенный способ выделения контура профиля на основе метода активных контуров не обеспечивает точного совпадения цефалометрических точек, выделенных автоматически и вручную. Однако, при правильной установке весовых коэффициентов вероятность ошибки неверной классификации типа профиля и гармоничности профиля может быть снижена до 0,03.

#### БЛАГОДАРНОСТИ

Работа выполнена в рамках государственного задания по теме FSSS-2023-0006.

#### ЛИТЕРАТУРА

- [1] Токаревич, И.В. Общая ортодонтия: учебное пособие для студ. выш. учеб. заведений / И.В. Токаревич. – Минск: БГМУ, 2015. – 219с.
- [2] Álvarez, L. Morphological snakes / L. Álvarez, L. Baumela, P. Henríquez, P. Márquez-Neila // IEEE Computer Society Conference on Computer Vision and Pattern Recognition. – 2010. – P. 2197–2202.
- [3] Kass, M. Snakes: Active Contour Models / M. Kass, A. Witkin, D. Terzopoulos // International Journal of Computer Vision. – 1987. – Vol. 1. – P. 321-331.
- [4] Yu, C. Out-of-distribution detection for reliable face recognition / C. Yu, X. Zhu, Z. Lei, SZ. Li // IEEE Signal Process Lett. – 2020. – Vol. 27. – P. 710-714.
- [5] Hassan, A. An Efficient Detection Framework for Linear Skin Lesions with Pigmentary Disorders / A. Hassan, I. Tawfik // 2020 30th International Conference on Computer Theory and Applications (ICCTA). – 2020. – P. 128-133.
- [6] Hajer, W. Preprocessing Latent-Fingerprint Images For Improving Segmentation Using Morphological Snakes / W. Hajer, R.H. Lamia, M. Ahmed, E. Najoua // 2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP). – 2020. – P. 1-6.
- [7] Jones, M. Robust Real-Time Face Detection / M. Jones, P. Viola // International Journal of Computer Vision. – 2004. – Vol. 57(2). – P. 137-154.
- [8] Rudinskaya, E. Face detection accuracy study based on race and gender factor using haar cascades / E. Rudinskaya, R. Paringer // CEUR Workshop Proceedings. – 2020. – Vol. 2667. – P. 238–242.
- [9] Hoang, V.T. Monitoring Employees Entering and Leaving the Office with Deep Learning Algorithms / V.T. Hoang, K.T. Minh, N.D. Hieu // Artificial Intelligence in Data and Big Data Processing. – 2022. – Vol. 124. – P. 641.
- [10] King, D. Dlib-ml: A Machine Learning Toolkit / D.E. King // Journal of Machine Learning Research. – 2009. – Vol. 10. – P. 1755–1758.

# Технология автоматизированного интеллектуального отбора информативных признаков для задачи классификации областей натуральных гиперспектральных изображений

М.И. Хотилин

Самарский национальный  
исследовательский университет им.  
академика С.П. Королева  
Самара, Россия  
khotilin.mi@ssau.ru

**Аннотация**—В данной статье описывается процесс создания автоматизированной технология отбора информативных признаков гиперспектрального изображения для осуществления процесса классификации. Описаны методы и алгоритмы поиска признаков принадлежности к определенным классам, их применения. Указаны дальнейшие пути и перспективы развития технологии и ее реализации.

**Ключевые слова**— гиперспектральные изображения, дискриминантный анализ, метод опорных векторов, нейронные сети, классификация, отбор признаков, Python, снижение размерности, информативные признаки

## 1. ВВЕДЕНИЕ

Гиперспектральные изображения представляю собой трёхмерный массив данных, включающий в себя пространственную информацию об объекте, дополненную спектральной информацией по каждой пространственной координате [1]. Анализ гиперспектральных изображений и их областей является одной из популярных тематик в области обработки изображений и компьютерного зрения. Автоматизация процесса анализа, и процесса поиска информативных признаков гиперспектральных изображений является актуальной задачей в настоящее время. В рамках данной работы рассматривается процесс построения технологии, использующей метод поиска информативных признаков и сверточные нейронные сети, для задачи классификации гиперспектральных изображений.

## 2. ПОСТАНОВКА ЗАДАЧИ

Выполнение обработки RGB изображений классическими методами, их классификация занимает сравнительно небольшое время и может выполняться практически на любом устройстве, в том числе и носимом. Ввиду ряда особенностей, обработка гиперспектральных изображений, будет требовать значительных вычислительных ресурсов. Например, для классификации одной области гиперспектрального изображения размером  $10 \times 10$  пикселей, количество необходимых совокупных яркостных и текстурных признаков может составлять десятки тысяч.

Актуальность данной работы заключается в создании технологии, основанной на методе поиска информативных признаков изображения, а также использовании нейронных сетей, позволяющей автоматизировать процесс поиска и значительно снизить

временные и аппаратные ресурсы, используемые в процессе анализа гиперспектральных изображений.

Весь процесс данной работы можно разделить на последовательно выполняемые этапы. На первом этапе выбирается набор рассматриваемых изображений. Далее посредством использования модуля предобработки происходит отображение каждого из рассматриваемых изображений на набор значений своих текстурных и яркостных признаков.

Следующим этапом необходимо провести обработку полученного массива данных с целью снижения его размерности, поскольку данные, которые он содержит, являются значительными по объему и могут содержать значения, которые не несут значимой информации, важной при классификации. В связи с этим, необходимо произвести сокращение размерности и поиск признаков, являющихся информативными. Также можно убрать из рассмотрения константные признаки и признаки не имеющие значения для отдельных изображений.

Далее, используя метод снижения размерности, описанный автором ранее в [5], основывающийся на совместном использовании пороговой фильтрации, линейного дискриминантного анализа, произведем дальнейшее снижение размерности. Добавляя метку класса к каждому из экземпляров отображения, получаем набор для проведения дальнейшей классификации. После этого переходим к поиску информативных признаков, посредством применения метода последовательного добавления признаков.

Используя полученные на предыдущих шагах данные и информативные признаки, можно произвести обучение сверточной нейронной сети, позволяющей по набору входных параметров, например изображения и количества признаков, в качестве вывода предоставлять необходимое количество искомым информативных признаков.

## 3. ПРАКТИЧЕСКОЕ ИССЛЕДОВАНИЕ И ПРОГРАММНАЯ РЕАЛИЗАЦИЯ

В качестве исходных данных рассматривался набор изображений HSI Dataset v1.3, состоящий из изображений листьев растений различных классов. Размер каждого из изображений  $512 \times 512$  пикселей с количеством спектральных каналов 237. Было выбрано 4 класса для рассмотрения: листья яблоны, картофеля, травы, клубники. Для исследования описанного выше алгоритма, посредством Python был реализован модуль

предобработки. Данный модуль позволяет исследовать изображение, найти ряд его признаков, таких как текстурные, гистограммные, морфологические и ряд других. Общее количество признаков для каждого изображения составило 62152539.

Далее было проведено 4 серии экспериментов с различными наборами данных и классификаторами. В качестве исходных данных рассматривались:

- исходный полный набор признаков;
- фильтрованный набор, из которого были удалены «выбросы» - изображений, которые ни один из классификаторов не смог верно классифицировать;
- признаки, отобранные посредством корреляционного анализа;
- «трансформированные» признаки – первоначальный набор признаков, трансформированный посредством применения метода главных компонент.

Каждая новая серия экспериментов позволяла значительно улучшать результаты классификации. Например, при использовании корреляционного анализа, при удалении константных, N/A признаков, и варьировании коэффициента пороговой корреляции удалось значительно снизить размерность рассматриваемых данных. Рассматривались пороговые коэффициенты корреляции от 0.5 до 0.99 с шагом в 0.01. При этом количество признаков составило от 28 (при коэффициенте в 0.5) до 2364 (при 0.99).

В качестве вспомогательных использовались алгоритмы классификации: линейный дискриминантный анализ, метод опорных векторов, логистическая регрессия, метод случайного леса, а также многослойный перцептрон (MLPClassifier) с различными решающими функциями, ядрами и алгоритмами оптимизации. Результаты классификации с помощью разных алгоритмов представлены на рисунке 1.

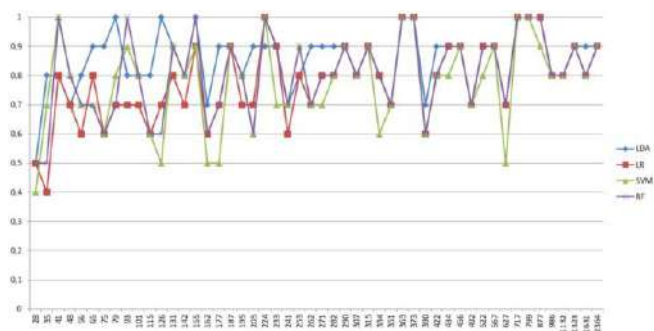


Рис. 1. Зависимость точности классификации рассматриваемых алгоритмов от размера выборки (порогового коэффициента корреляции)

В качестве наиболее эффективного, по точности классификации и затраченным ресурсам, был выбран LDA (линейный дискриминантный анализ).

Следующим шагом является поиск информативных признаков из набора, оставшегося в результате снижения размерности. Для этого использовался метод, описанный

ранее автором в [5]. В итоге получаем паттерн обработки гиперспектральных изображений, обладающий меньшими ресурсными требованиями, по сравнению с классическими методами.

Увеличивая объем выборки, и обучая на полученных данных сверточную нейронную сеть, получаем технологию автоматизированного интеллектуального отбора информативных признаков для гиперспектрального изображения, поданного на вход данной сети. В качестве нейронной сети была выбрана сверточная CFR-сеть, использующая алгоритм минимизации контрфактических сожалений Монте-Карло (MCCFR). В настоящее время ведется работа над обучением и тонкой настройкой нейронной сети, позволяющей выполнять описанные в статье вычисления.

#### 4. ЗАКЛЮЧЕНИЕ

Поиск признаков, определяющих однозначно принадлежность изображений к классу, является одной из значимых задач обработки изображений. В случае рассмотрения гиперспектральных изображений существующие методы обработки, определения признаков изображений и алгоритмы классификации являются ресурсозатратными и, для оперативного решения, требуется оптимизация используемых ресурсов.

Использование сверточных нейронных сетей позволит в значительной степени сократить требуемые для вычислений аппаратные и программные ресурсы, что потенциально позволяет использовать описанные в статье алгоритм и общую технологию на ряде устройств, в том числе и на носимых мобильных устройствах и беспилотных летательных аппаратах.

#### ЛИТЕРАТУРА

- [1] Zimichev, E.A Spectral-spatial classification with k-means++ participational clustering / E.A. Zimichev, N.L. Kazanskiy, P.G. Serafimovich // Computer Optics. – 2014. – Vol. 38. – № 2. – P. 281-286. DOI: 10.18287/0134-2452-2014-38-2-281-286
- [2] Kazanskiy, N.L. Simulation of hyperspectrometer on spectral linear variable filters / N.L. Kazanskiy, S.I. Kharitonov, S.N. Khonina, S.G. Volotovskiy, Yu.S. Strelkov // Computer Optics. – 2014. – Vol. 38.(2). – P. 256-270. DOI: :10.18287/0134-2452-2014-38-2-256-270
- [3] Khotilin, M. Classification of objects of natural hyperspectral images / M. Khotilin, N. Kravtsova, I. Rytsarev, A. Kupriyanov // 2020 International Conference on Information Technology and Nanotechnology (ITNT). – 2020. - P. 1-3. DOI: 10.1109/ITNT49337.2020.9253254.
- [4] Goncharova, E.F. Greedy algorithms of feature selection for multiclass image classification / E.F. Goncharova, A.V. Gaidel // CEUR Workshop Proceedings (IPERS-ITNT 2018 - Proceedings of the International Conference on Information Technology and Nanotechnology - Session: Image Processing and Earth Remote Sensing). – 2018. - Vol. 2210. - P. 38-46. DOI: 10.18287/1613-0073-2018-2210-38-46
- [5] Khotilin M. The technology of constructing an informative feature of a natural hyperspectral image area for the classification problem / M. Khotilin // 2021 International Conference on Information Technology and Nanotechnology (ITNT). – 2021. - P. 1-4. DOI: 10.1109/ITNT52450.2021.9649178.
- [6] Khotilin M. The technology of informative features searching method applied for the problem of classifying areas of natural hyperspectral images/ M. Khotilin // 2022 International Conference on Information Technology and Nanotechnology (ITNT). – 2022. DOI: 10.1109/ITNT55410.2022.9848638

# Быстрая одноклассовая SVM классификация для большой обучающей совокупности

М.Ю. Курбаков

Тульский государственный университет,  
Лаборатория когнитивных технологий и симуляционных систем,  
Тула, Россия  
muwsik@mail.ru

В. В. Сулимова

Тульский государственный университет,  
Лаборатория когнитивных технологий и симуляционных систем,  
Тула, Россия  
vsulimova@yandex.ru

**Аннотация**—В основу данной работы положен популярный метод одноклассовой классификации OCSVM. Мы предлагаем усовершенствованный вариант данного метода, целью создания которого является обеспечение возможности работы с большими обучающими совокупностями, что является проблематичным для OCSVM из-за высокой трудоемкости обучения. Основная идея предлагаемого подхода заключается в применении OCSVM к независимым случайным подвыборкам из исходной обучающей совокупности с последующим объединением результатов в единое решение, совпадающее по виду с решающим правилом OCSVM. Экспериментальное исследование показало, что предложенный подход позволяет существенно ускорить решение задачи, получая при этом точное (или близкое к точному) решение.

**Ключевые слова** — одноклассовая классификация, OCSVM, большие задачи, повышение производительности.

## 1. ВВЕДЕНИЕ

Одноклассовая классификация является частным случаем многоклассовой классификации, когда обучающая совокупность состоит только из объектов одного класса (как правило, целевого). При этом требуется на основе анализа этой обучающей совокупности построить решающее правило, позволяющее определить, относится ли новый объект к присутствующему на обучении классу или нет [1].

Одним из наиболее популярных методов решения задачи одноклассовой классификации является метод OCSVM [2], который активно используется для решения различных прикладных задач, в частности, для классификации текстов [3], обнаружения мошеннических транзакций по кредитным картам [4], обнаружения вторжений [5], в системах видеонаблюдений [6], медицинских системах [7] и т.д.

В случае небольшого размера обучающей совокупности задача построения оптимального решающего правила OCSVM может быть достаточно быстро решена при помощи классических методов оптимизации. Однако в случае большого количества объектов, что характерно для многих современных задач анализа данных, построение решающего правила оказывается очень трудоемким по времени и памяти, что существенно затрудняет, а в ряде случаев и делает невозможным его непосредственное применение на практике.

Несмотря на массовые исследования в данной области, до сих пор не удалось найти решение, которое бы полностью удовлетворяло практические нужды. В частности, работы, направленные на повышение

скорости обучения, обычно не решают проблему нехватки оперативной памяти и обеспечения быстрого доступа к объектам [8], методы, ориентированные на более экономичное использование памяти, часто оказываются существенно менее точными или имеют низкую скорость сходимости [9]. Применение технологий параллельных и распределенных вычислений [10] смягчает проблему большого количества данных, но не решает ее, поскольку большинство методов имеют итерационную природу с зависимостями по данным, что существенно ограничивает возможности повышения производительности данным путем.

## 2. ПРЕДЛАГАЕМЫЙ ПОДХОД

В данной работе мы предлагаем добиться повышения производительности решения задачи одноклассовой классификации по методу OCSVM путем замены одной большой исходной задачи на серию существенно более мелких задач, исходными данными для каждой из которых является случайная выборка объектов из полной обучающей совокупности, сформированная независимо от остальных. Результаты обучения по частным подвыборкам предлагается объединять в единое решающее правило согласно принципу усреднения, предложенному нами ранее для двухклассовых задач SVM [11], в результате чего итоговое решение имеет ту же структуру, что и традиционное решение задачи по методу OCSVM. В результате такой подход имеет существенное преимущество перед бэггингом, который, предполагает построение ансамбля классификаторов для частных подвыборок, а итоговое решение получает путем голосования или какой-либо другой схемы построения ансамблей, что требует хранения всех частных классификаторов и отдельного применения каждого из них на этапе распознавания.

Следует отметить, что предложенный подход снимает теоретическое ограничение на размер обучающей совокупности, поскольку, как и бэггинг [12], фактически, не требует одновременной загрузки всех объектов в память. Однако, даже при использовании выборочных методов, на практике в большинстве случаев по-прежнему в память загружается полная обучающая совокупность (например, в рамках пакета `scikit-learn python`). Это обусловлено тем фактом, что традиционный формат хранения данных `libsvm` не позволяет вычислить позицию начала объекта с заданным номером, что сильно снижает эффективность произвольного доступа к объектам в файле при его традиционном чтении.

Для снятия практического ограничения на объем обучающей совокупности при решении задачи одноклассовой классификации мы предлагаем использовать принцип оптимальной работы с данными [13], разработанный нами ранее для двухклассового распознавания, но который может быть почти без изменений применен и для одноклассовой ситуации. Его основная идея заключается в осуществлении предварительной разметки данных с сохранением областей файла, содержащих диапазоны объектов обучающей совокупности, и последующем использовании этой разметки для быстрого формирования подвыборок за счет предварительной примерной локализации объектов, а также за счет привлечения предоставляемого операционной системой механизма отображения файлов в память, который позволяет работать с данными на диске как с обычными данными в памяти.

### 3. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

В таблице 1 приведено время обучения и AUC (на тестовой выборке) для `libsvm` и для предложенного метода (для различного размера случайных подвыборок SRS и числа случайных подвыборок NRS для двух модельных наборов данных с разным числом объектов и выбросов (аномальных объектов). Модельные данные генерировались по следующей схеме. Положительный класс (нормальные объекты) генерировались в соответствии с нормальным распределением, выбросы – равномерно вокруг положительного класса. Для параметров `libsvm` (как отдельно, так и для обучения по подвыборкам в составе предложенного подхода) были установлены значения:  $nu = 0.001$ ,  $\gamma = 0.01$ .

Таблица 1. РЕЗУЛЬТАТЫ РАБОТЫ МЕТОДОВ

Метод	Параметры		Набор данных (объектов/выбросов)			
			100000 / 100		500000 / 500	
	NRS	SRS	время	AUC	время	AUC
Libsvm	-	-	0,2314	0,9750	6,2159	0,9920
Предпо- женный подход	10	300	0,0063	0,9966	0,0199	0,9975
	10	1000	0,0047	0,9968	0,0103	0,9996
	100	300	0,0196	0,9973	0,0296	0,9975
	100	1000	0,0362	0,9999	0,0468	0,9999

Как видно из таблицы 1, предложенный подход при различных значениях параметров позволяет получить решение, превосходящее по качеству решение, полученное при помощи популярной библиотеки `libsvm`. При этом время, затраченное на построение решающего правила, оказывается на 1-2 порядка.

Существенное повышение производительности (даже в условиях последовательной реализации предложенного метода) обусловлено нелинейной вычислительной сложностью решения задачи SVM относительно числа объектов, в результате чего оказывается вычислительно более выгодно решать серию небольших задач, чем одну более крупную. Повышение качества обусловлено снижением чувствительности к наличию выбросов в обучающей совокупности за счет усреднения частных решающих правил, построенных по небольшим подвыборкам.

### ЗАКЛЮЧЕНИЕ

В работе предложен подход к решению больших одноклассовых задач SVM, основанный на обучении на подвыборках и последующем усреднении результатов с получением решения, совпадающего по структуре с традиционным решением задачи SVM. Эксперименты на модельных данных показали существенное преимущество предложенного подхода по сравнению с наиболее популярной библиотекой для решения задач SVM - `libsvm`.

### БЛАГОДАРНОСТИ

Работа выполнена при финансовой поддержке Министерства науки и высшего образования РФ в рамках государственного задания FEWG-2021-0012.

### ЛИТЕРАТУРА

- [1] Perera, P. One-Class Classification: A Survey / P. Perera, P. Oza, V. Patel // Computer Vision and Pattern Recognition. – 2021. – P. 19. doi:10.48550/arXiv.2101.03064.
- [2] Scholkopf, B. Estimating the support of a high-dimensional distribution / B. Scholkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, R. C. Williamson // Neural computation. – 2001. – Vol. 13(7). – P. 1443-1471. doi: 10.1162/089976601750264965.
- [3] Shravan, K. One-Class Text Document Classification with OCSVM and LSI / K. Shravan, V. Ravi // Artificial Intelligence and Evolutionary Computations in Engineering Systems. – 2017. – Vol. 517. doi:10.1007/978-981-10-3174-8\_50.
- [4] Wu, T. Locally Interpretable One-Class Anomaly Detection for Credit Card Fraud Detection / T. Wu, Y. Wang // 2021 International Conference on Technologies and Applications of Artificial Intelligence. – 2021. – P. 25-30. doi:10.48550/arXiv.2108.02501.
- [5] Krishnaveni, S. Anomaly-Based Intrusion Detection System Using Support Vector Machine / S. Krishnaveni, P. Vigneshwar, S. Kishore, B. Jothi, S. Sivamohan // Artificial Intelligence and Evolutionary Computations in Engineering Sys. – 2020. – Vol. 1056. doi:10.1007/978-981-15-0199-9\_62.
- [6] Seredin, O. S. A Skeleton Features-Based Fall Detection Using Microsoft Kinect v2 with One Class-Classifer Outlier Removal / O. S. Seredin, A. V. Kopylov, S. C. Huang, D. S. Rodionov // ISPRS-International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. – 2019. – Vol. 4212. – P. 189-195.
- [7] Xiao-Kang, W. KDE-OCSVM model using Kullback-Leibler divergence to detect anomalies in medical claims / W. Xiao-Kang, H. Wen-Hui, Z. Hong-Yu, W. Jian-Qiang, G. Mark, T. Zhang-Peng, S. Kai-Wen // Expert Systems with Applications. – 2022. – Vol. 200. doi:10.1016/j.eswa.2022.117056.
- [8] Zeyi, W. ThunderSVM: A Fast SVM Library on GPUs and CPUs / W. Zeyi, S. Jiashuai, L. Qinbin, H. Bingsheng, C. Jian // Journal of Machine Learning Research. – 2018. – Vol.19(21). – P. 1-5.
- [9] Lee, Y.-J. RSVM: Reduced Support Vector Machines / Y.-J. Lee, O. L. Mangasarian // In Proceedings of the SIAM International Conference on Data Mining. – 2021. – Vol. 1. – P. 325-361.
- [10] Stolpe, M. Distributed Support Vector Machines: An Overview / M. Stolpe, K. Bhaduri, K. Das // Solving Large Scale Learning Tasks. Challenges and Algorithms. – 2016. – Vol. 9580. – P. 109-138. doi:10.1007/978-3-319-41706-6\_5.
- [11] Makarova, A. Mean Decision Rule Method for Constructing Nonlinear Boundaries in Large Binary SVM Problems / A. Makarova, M. Kurbakov, V. Sulimova // Inf. Technology and Nanotechnology. – 2020. – P. 1-6. doi:10.1109/ITNT49337.2020.9253181.
- [12] Shieh, A. D. Ensembles of One Class Support Vector Machines / A. D. Shieh, D. F. Kamm // Multiple Classifier Systems. – 2009. – Vol. 5519. – P. 181-190. doi:10.1007/978-3-642-02326-2\_19.
- [13] Курбаков, М.Ю. Оптимизация загрузки данных в формате `libsvm` при решении двухклассовой задачи SVM методом усреднения решающих правил в условиях большой обучающей совокупности / М.Ю. Курбаков, А.И. Макарова, В.В. Сулимова // Информационные технологии и нанотехнологии. – 2019. – Т. 4. – С. 53-60.

# О логической классификации целочисленных данных

Е. В. Дюкова  
Федеральный исследовательский  
центр «Информатика и управление»  
Российской академии наук  
Москва, Россия  
edjukova@mail.ru

Г. О. Масляков  
Федеральный исследовательский  
центр «Информатика и управление»  
Российской академии наук  
Москва, Россия  
gleb-mas@mail.ru

А. П. Дюкова  
Федеральный исследовательский  
центр «Информатика и управление»  
Российской академии наук  
Москва, Россия  
anastasia.d.95@gmail.com

**Аннотация**— Рассматриваются основные подходы к задаче классификации на основе прецедентов, базируются на применении аппарата дискретной математики (логических методов анализа данных). Предлагается общая схема описания логических классификаторов с использованием терминологии процедур корректного голосования.

**Ключевые слова**—классификация на основе прецедентов, логический классификатор, отношение частичного порядка, представительный элементарный классификатор, сильная логическая закономерность, ДСМ-метод

## 1. ВВЕДЕНИЕ

Задача классификации на основе прецедентов рассматривается в следующей постановке.

Исследуется некоторое множество объектов  $M$ . Известно, что  $M$  представимо в виде объединения непересекающихся подмножеств  $K_1, \dots, K_l$ , называемых классами. Объекты из  $M$  описываются признаками  $x_1, \dots, x_n$ , каждый из которых является некоторой наблюдаемой или измеряемой характеристикой этих объектов и имеет ограниченное число допустимых значений. Значения признаков кодируются целыми числами. Имеется конечный набор  $S_1, \dots, S_m$  объектов из множества  $M$ , о которых известно, каким классам они принадлежат. Это прецеденты или обучающие объекты. Прецедент  $S_i$ ,  $i \in \{1, \dots, m\}$ , задаётся в виде набора  $(a_{i1}, \dots, a_{in})$ , где  $a_{ij}$  – значение признака  $x_j$ . Требуется по предъявленному набору значений признаков  $(a_1, \dots, a_n)$ , описывающему некоторый объект  $S$  из  $M$ , о котором, вообще говоря, неизвестно, какому классу он принадлежит, определить (распознать) этот класс.

Фундаментальную роль в создании отечественных методов логической классификации сыграли работы члена-корреспондента РАН С.В. Яблонского, в которых введено хорошо известное в дискретной математике понятие теста, и работы академика РАН Ю.И. Журавлева, опубликованные в 70-х и 80-х годах прошлого века. Понятие теста, первоначально применяемое в задачах контроля управляющих систем, явилось основным для конструирования одной из первых моделей классификаторов, именуемых далее процедурами корректного голосования (PCV). Основы проблематики были заложены также в статьях российских ученых М.М. Бонгарда (1967 г.) и М.Н. Вайнцвайга (1973 г.).

В дальнейшем направление PCV развивалось в работах отечественных и зарубежных авторов и существенное развитие получило в статьях [2–6].

Зарубежные исследования в области логической классификации представлены методами Logical Analysis of Data (LAD) и Formal Concept Analysis (FCA).

Основополагающие идеи LAD и FCA принадлежат соответственно П. Хаммеру (1986 г.) и Р. Вилле (1981 г.).

В России методы LAD предложены практически параллельно с зарубежными авторами и развиты в ряде работ Ю.И. Журавлёва, В.В. Рязанова (см., например, [7, 9]). Методы FCA представлены в работах В.К. Финна, С.О. Кузнецова, М.И. Забежайло, Д.И. Игнатова и Д.В. Виноградова ([1, 8, 10–12]). В [10] предложен так называемый метод автоматического порождения гипотез (или ДСМ-метод), который позднее в 1990-х годах был адаптирован В.К. Финном и его учениками для задач машинного обучения. ДСМ-классификатор можно отнести к FCA.

Все три названных направления PCV, LAD и FCA имеют много общего. С другой стороны, каждый из подходов использует свою терминологию и демонстрирует некоторую оригинальность. В настоящей работе предлагается общее описание подходов с использованием понятий PCV.

## 2. ОПИСАНИЕ ПРОЦЕДУР PCV, LAD и FCA

Введем основные понятия, используемые при синтезе процедур PCV.

Пусть  $H = \{x_{j_1}, \dots, x_{j_r}\}$  – набор из  $r$  различных признаков,  $\sigma = (\sigma_1, \dots, \sigma_r)$  – набор, в котором  $\sigma_i$  – допустимое значение признака  $x_{j_i}$ ,  $i = 1, 2, \dots, r$ . Пара  $(\sigma, H)$  называется элементарным классификатором (ЭК) ранга  $r$  [6].

Близость объекта  $S = (a_1, \dots, a_n)$  из  $M$  и ЭК  $(\sigma, H)$ ,  $\sigma = (\sigma_1, \dots, \sigma_r)$ ,  $H = \{x_{j_1}, \dots, x_{j_r}\}$ , оценивается величиной  $B(S, \sigma, H)$ , равной 1, если  $a_{j_t} = \sigma_t$  при  $t = 1, 2, \dots, r$ , и равной 0 в противном случае. Если  $B(S, \sigma, H) = 1$ , то говорят, что объект  $S$  содержит ЭК  $(\sigma, H)$ .

Множество прецедентов класса  $K$  обозначается через  $R(K)$ . ЭК  $(\sigma, H)$  называется *корректным для класса  $K$* , если для любой пары прецедентов  $S \in K$  и  $S' \notin K$  не выполнено  $B(S, \sigma, H) = B(S', \sigma, H) = 1$ . Корректный ЭК  $(\sigma, H)$  класса  $K$  называется *тупиковым*, если любой ЭК  $(\sigma', H')$  такой, что  $\sigma' \subset \sigma$ ,  $H' \subset H$ , не является корректным для  $K$ . ЭК  $(\sigma, H)$  – *(тупиковый) представительный для класса  $K$* , если  $(\sigma, H)$  – (тупиковый) корректный ЭК для  $K$  и хотя бы один объект из  $R(K)$  содержит  $(\sigma, H)$ .

При синтезе процедур LAD и в ДСМ-методе используются соответственно понятия «логическая закономерность» [ЛЗ] [7] и «ДСМ-гипотеза» [10, 12].

Представительный для класса  $K$  ЭК называется *сильной логической закономерностью*, если он содержится в наибольшем числе прецедентов класса  $K$ .

Положим  $R_K(\sigma, H) = \{S \in R(K) : B(S, \sigma, H) = 1\}$ ,  $|R_K(\sigma, H)|$  – мощность множества  $R_K(\sigma, H)$ .

Представительный для класса  $K$  ЭК  $(\sigma, H)$  порождает положительную ДСМ-гипотезу для  $K$ , если для любого ЭК  $(\sigma', H')$  такого, что  $\sigma \subset \sigma'$ ,  $H \subset H'$ , найдётся объект  $S \in R_K(\sigma, H)$ , не содержащий  $(\sigma', H')$ .

Классифицирующий алгоритм  $A$  на этапе обучения строит для каждого класса  $K$  некоторое множество  $P^A(K)$  представительных ЭК. В PCV в качестве элементов множества  $P^A(K)$  часто рассматриваются тупиковые представительные ЭК. В LAD строятся сильные логические закономерности, а в ДСМ-методе ЭК, порождающие положительные ДСМ-гипотезы.

В PCV и LAD каждый элемент множества  $P^A(K)$  «голосует» за отнесение объекта  $S$  классу  $K$ . Для оценки принадлежности объекта  $S$  классу  $K$  суммируются соответственно величины  $|R_K(\sigma, H)| \times B(S, \sigma, H)$  и  $B(S, \sigma, H)$ ,  $(\sigma, H) \in P^A(K)$ .

ДСМ-классификатор действует более строго. Объект  $S$  относится к классу  $K$ , если  $S$  содержит хотя бы один ЭК из  $P^A(K)$  и не содержит ни одного ЭК из  $P^A(K')$  при  $K' \neq K$ . В противном случае происходит отказ от классификации.

Предлагаемое описание общей схемы обучения алгоритмов логической классификации основано на приводимых ниже утверждениях 1 – 3.

Пусть  $\mathcal{P}(K)$  – множество всех представительных ЭК класса  $K$ , на котором задан некоторый частичный (предпорядок) порядок  $\leq$ . ЭК  $(\sigma, H) \in \mathcal{P}(K)$  называется *максимальным* относительно частичного (предпорядка) порядка  $\leq$ , если не существует ЭК  $(\sigma', H') \in \mathcal{P}(K)$  такого, что  $(\sigma, H) < (\sigma', H')$ .

Зададим на множестве  $\mathcal{P}(K)$  отношение частичного порядка  $\leq_1$ . Будем считать, что ЭК  $(\sigma_2, H_2) \in \mathcal{P}(K)$  следует за  $(\sigma_1, H_1) \in \mathcal{P}(K)$ , если  $H_2 \subseteq H_1$  и  $\sigma_2 \subseteq \sigma_1$ . Тогда справедливо

**Утверждение 1.** ЭК  $(\sigma, H)$  является тупиковым представителем для класса  $K$  тогда и только тогда, когда  $(\sigma, H)$  – максимальный относительно частичного порядка  $\leq_1$  элемент множества  $\mathcal{P}(K)$ .

Зададим на множестве  $\mathcal{P}(K)$  отношение частичного предпорядка  $\leq_2$ . Будем считать, что ЭК  $(\sigma_2, H_2) \in \mathcal{P}(K)$  следует за  $(\sigma_1, H_1) \in \mathcal{P}(K)$ , если  $|R_K(\sigma_1, H_1)| \leq |R_K(\sigma_2, H_2)|$ . Тогда справедливо

**Утверждение 2.** ЭК  $(\sigma, H)$  является сильной ЛЗ класса  $K$  тогда и только тогда, когда  $(\sigma, H)$  – максимальный относительно частичного предпорядка  $\leq_2$  элемент множества  $\mathcal{P}(K)$ .

Зададим на множестве  $\mathcal{P}(K)$  отношение частичного порядка  $\leq_3$ . Будем считать, что ЭК  $(\sigma_2, H_2) \in \mathcal{P}(K)$  следует за  $(\sigma_1, H_1) \in \mathcal{P}(K)$ , если  $R_K(\sigma_1, H_1) \subseteq R_K(\sigma_2, H_2)$  и  $H_1 \subseteq H_2$ . Тогда справедливо

**Утверждение 3.** ЭК  $(\sigma, H)$  порождает положительную ДСМ-гипотезу для класса  $K$  тогда и

только тогда, когда  $(\sigma, H)$  – максимальный относительно частичного порядка  $\leq_3$  элемент множества  $\mathcal{P}(K)$ .

Утверждения 1 – 3 имеют силу и в случае частично упорядоченных целочисленных данных [5].

### 3. ЗАКЛЮЧЕНИЕ

В работе описана общая схема синтеза алгоритмов логической классификации. В рамках данной схемы в единой терминологии описаны основные алгоритмы классификации направлений PCV, LAD и FCA. Приведены утверждения показывающие, что каждое из рассмотренных направлений логической классификации ориентировано на задание своего порядка на множестве  $\mathcal{P}(K)$  и поиске максимальных относительно заданного порядка элементов.

### ЛИТЕРАТУРА

- [1] Виноградов, Д. В. О представлении объектов битовыми строками для ВКФ-метода / Д. В. Виноградов // Научная и техническая информация, Сер. 2. – 2018. – Т.5. – С. 1–4.
- [2] Дюкова, Е. В. Об асимптотически оптимальном алгоритме построения тупиковых тестов / Е. В. Дюкова // Докл. АН СССР, 1977. – Т. 233, №4. – С. 527–530.
- [3] Дюкова, Е. В. Дискретный анализ признаков описаний в задачах распознавания большой размерности / Е. В. Дюкова, Ю. И. Журавлёв // Ж. вычисл. матем. и матем. физ. – 2000. – Т. 40, №8. – С. 1264–1278.
- [4] Дюкова, Е. В. Об алгебраическом синтезе корректирующих процедур распознавания на базе элементарных алгоритмов / Е. В. Дюкова, Ю. И. Журавлёв, К. В. Рудаков // Ж. вычисл. матем. и матем. физ. – 1996. – Т. 36, №8. – С. 217–225.
- [5] Дюкова, Е. В. О логическом анализе данных с частичными порядками в задаче классификации по прецедентам / Е. В. Дюкова, Г. О. Масляков, П. А. Прокофьев // Ж. вычисл. матем. и матем. физ. – 2019. – Т. 59, №9. – С. 1605–1616.
- [6] Дюкова, Е. В. Поиск информативных фрагментов описаний объектов в дискретных процедурах распознавания / Е. В. Дюкова, Н. В. Песков // Ж. вычисл. матем. и матем. физ. – 2002. – Т. 42, №5. – С. 741–753.
- [7] Журавлёв, Ю. И. Распознавание. Математические методы. Программная система. Практические применения / Ю. И. Журавлёв, В. В. Рязанов, О. В. Сенько // М.: ФАЗИС. – 2006. – Т. 176. – 159 с.
- [8] Забежайло, М. И. О некоторых оценках сложности вычислений в ДСМ-рассуждениях / М. И. Забежайло // Искусственный интеллект и принятие решений. – 2015. – Т. 1. – С. 3–17.
- [9] Ковшов, Н. В. Алгоритмы поиска логических закономерностей в задачах распознавания / Н. В. Ковшов, В. Л. Моисеев, В. В. Рязанов // Ж. вычисл. матем. и матем. физ. – 2008. – Т. 48, №2. – С. 329–344.
- [10] Финн, В. К. О возможности формализации правдоподобных рассуждений средствами многозначных логик / К. В. Финн // Всесоюз. симп. по логике и методологии науки. – Киев: Наукова думка, 1976. – С. 82–83.
- [11] Gnatyshak, D. V. Triadic Formal Concept Analysis and triclustering: searching for optimal patterns / D. V. Gnatyshak, D. I. Ignatov, S. O. // Kuznetsov Mach Learn. – 2015. – Vol. 101. – P. 271–302.
- [12] Kuznetsov, S. O. Mathematical aspects of concept analysis / S. O. Kuznetsov // Journal of Mathematical Science. – 1996. – Vol. 80(2). – P. 1654–1690

# Анализ влияния различных аспектов личности студента на академическую успеваемость

Н.В. Пустовалова  
Новосибирский государственный технический университет  
Новосибирск, Россия  
NVPustovalova@gmail.com

Т.В. Авдеевко  
Новосибирский государственный технический университет  
Новосибирск, Россия  
tavdeenko@mail.ru

**Аннотация**—В данной работе представлены результаты исследования датасета, сконструированного авторами для программной реализации компонентов персонализированной образовательной среды университета. Указанный набор данных получен авторами в результате тестирования психометрических характеристик студентов. В тестировании приняли участие 191 человек, учащиеся со 2-го по 4-ый курс Новосибирского государственного технического университета (НГТУ): 123 мужчины и 68 женщин в возрасте от 18 до 23 лет. После подготовки данных были построены регрессионные модели, в результате чего выявлено, что наиболее значимые предикторы — это «добросовестность» и «система торможения поведения». Эти же переменные оказались значимы при анализе совместного попарного влияния с категориальными предикторами «модальность», «стиль реагирования на изменения», «пол». Также была построена логистическая регрессия. Для этого студенты были разделены на две категории успеваемости.

**Ключевые слова**— персонализация, модель обучаемого, психометрические характеристики, образовательный контент, регрессионный анализ

## 1. ВВЕДЕНИЕ

Один из ведущих трендов современного образования – персонализация [1, 2]. В [3] персонализация определяется как процесс, формирующий функциональность, интерфейс, информационное содержание или отличительные особенности системы, с целью повышения ее важности для индивида. Это одна из стратегий, позволяющих реализовать методологические подходы современной педагогики в контексте образовательного процесса [9]. Персонализированная образовательная среда (PLE – personal learning environment) один из инструментов воплощения данной стратегии. При этом такая среда может реализовывать и другие образовательные стратегии, сочетая их для усиления эффекта. В составе ее архитектуры часто присутствует модель обучаемого [4]. Модель обучаемого с концептуальной точки зрения описывает те навыки и умения, которые уже есть у индивида, а также те, которые должны быть в результате обучения. Современная точка зрения также предполагает, что модель обучаемого содержит информацию о личности студента [5]. Эти сведения позволяют рекомендовать предметные области и дисциплины с учетом особенностей личности и интересов, выбирать образовательные технологии и контент, соответствующий целям обучения. Персонализация, совместно с использованием специализированного контента, являются важнейшими стратегиями реализации PLE в университете. В предыдущих работах [7] мы исследовали вопрос, какие персональные параметры стоит принимать во внимание при рекомендации различных форм образовательного контента при функционировании PLE. Мы самостоятельно опередили набор характеристик,

описывающих личность обучаемого, включив в него три группы характеристик, значимых для создания PLE (когнитивные, личностные и мотивационные). Цель данной работы – исследовать имеющийся датасет более подробно для поиска более сложных зависимостей между персональными характеристиками студентов и их академической успеваемостью. В том числе, проанализировать совместное влияние характеристик личности на академический результат.

## 2. ПОСТАНОВКА ЗАДАЧИ И ОПИСАНИЕ ИССЛЕДОВАНИЯ

В качестве входных данных для «модели обучаемого» используются результаты нескольких психометрических тестов. Для оценки когнитивных особенностей тест структуры интеллекта Амтхауэра, прогрессивные матрицы Равена, тест на определение модальности. В датасет были добавлены результаты теста для оценки стиля реагирования на изменения. Для оценки личностных характеристик — результаты тестов «Большая пятерка», тест Бачард на определение эмоционального интеллекта. Для оценки мотивационных характеристик используются русскоязычная адаптация опросника Грэй-Уилсона, а также тесты Элерса на стремление к достижениям и избеганию неудач. Итоговые баллы за каждый тест или за каждый из субтестов представляют в модели отдельную независимую переменную. В качестве зависимой переменной в исследовании выступает академическая успеваемость, которая рассчитывается как балл по шкале ECTS без учета пересдач, полученный за сдачу экзаменов, зачетов и курсовых. Кроме того, отдельно были рассчитаны средний балл по математическим предметам, предметам профессионального цикла и гуманитарным дисциплинам.

После необходимых проверок [14], используя язык программирования R, были построены регрессионные модели. Статистически незначимые предикторы пошагово исключались с контролем значений F-statistic и R-squared всей модели. При построении моделей сначала в качестве зависимой переменной (переменной отклика) была определена переменная AVG, которая содержит средний балл ECTS за все учебные дисциплины и виды деятельности без учета пересдач. Затем были построены модели, где в качестве переменной отклика по очереди выступают AVM, AVW, AVH (средний балл по математическим, профессиональным и гуманитарным дисциплинам соответственно). В таблице I указаны коэффициенты для полученных регрессионных моделей.

Таблица I. Коэффициенты Для Зависимых и Независимых Переменных

Y	$\beta_0$	iq3	iq4	iq6	riq E	Co ns	Ne uro	BIS	Suc Beh
AVG	75,555	1,92	-	1,808	-	2,764	-	2,208	-
AVM	72,45	2,776	-	1,3	-	2,897	-	2,025	-

Y	$\beta_0$	iq3	iq4	iq6	riq E	Cons	Neuro	BIS	Suc Beh
AVW	79,953	-	1,743	2,464	-	-	-2,42	2,568	2,022
AVH	75,904	-	2,034	-	1,803	2,285	-	2,273	-

Предположительно некоторые переменные могут оказывать влияние не по отдельности, а в комплексе. Оценив влияние отдельных предикторов на академическую успеваемость, мы проанализировали совместное влияние на нее пары предикторов, один из которых категориальный (таблица II). Кроме того, была построена логистическая регрессия. В этом случае мы поделили студентов по формальному признаку (сдача сессии в срок и без оценки «неудовлетворительно») на успевающих и неуспевающих.

### 3. ЗАКЛЮЧЕНИЕ

Для подтверждения результатов, полученных ранее, мы дополнили имевшуюся ранее выборку, включив в нее новые наблюдения и дополнительные переменные регрессоры. Полученные в результате многомерного регрессионного анализа модели, говорят о наличии зависимостей между индивидуальными особенностями студентов (когнитивных, персональных и мотивационных) и академической успеваемостью. Предиктор Cons – добросовестность (опросник «Большая 5») отражает такую важную черту, как уверенность в себе и стремление к достижению результата. Так же определено важны когнитивные способности индивида, в частности те, которые отражают владение языком, индукцию, дедукцию. Во всех моделях присутствует предиктор BIS (behavioral inhibition system – система торможения поведения), связанный с отрицательной мотивацией, а именно боязнью наказания и провала. При этом есть различия в тех факторах, которые влияют на успеваемость по категориям дисциплин. Было выявлено совместное влияние некоторых предикторов на успеваемость. Хотя стоит отметить, что это совместное влияние было выявлено с переменными Cons и BIS. Которые значимы и в основных моделях регрессионного анализа.

Таблица II. ПАРЫ ПРЕДИКТОРОВ, ОКАЗЫВАЮЩИХ СОВМЕСТНОЕ ВЛИЯНИЕ НА УСПЕВАЕМОСТЬ

Пара переменных	Значимые связи
BIS * Modality (AVG)	Modality_Visual
BIS * Modality (AVM)	Modality_Visual
BIS * Modality (AVH)	Modality_Visual
Cons * Sex (AVG)	Sex_Male
BIS * ReactSt1 (AVG)	ReactSt1_Mixed
BIS * ReactSt1 (AVM)	ReactSt1_Mixed
BIS * ReactSt1 (AVW)	ReactSt1_Mixed

Выявленные связи мы планируем дополнительно включить в онтологию «Модель обучаемого», которую разрабатываем для совершенствования существующей информационной системы НГТУ. Таким образом она

будет отражать более сложные виды влияния психометрических характеристик студентов на академическую успеваемость.

### БЛАГОДАРНОСТИ

Работа выполнена при финансовой поддержке Министерства Науки и Высшего Образования в рамках Госзадания (проект № FSUN-2020-0009)

### ЛИТЕРАТУРА

- [1] Bhutoria, A. Personalized education and artificial intelligence in United States, China, and India: A systematic Review using a Human-In-The-Loop model / A. Bhutoria // Computers and Education: Artificial Intelligence. – 2022. – P. 100068.
- [2] Liu, D. Y.-T. Data-driven personalization of student learning support in higher education / D. Y.-T. Liu, K. Bartimote-Aufflick, A. Pardo, A.J. Bridgeman // Learning analytics: Fundamentals, applications, and trends. Vol. 94. – P. 143-169.
- [3] Blom, J. Personalization: a taxonomy / J. Blom // CHI'00 extended abstracts on Human factors in computing systems. – 2000. – P. 313-314.
- [4] Pustovalova, N.V. University's Educational Environment Personalization Based on the Ontological Models / N.V. Pustovalova, T.V. Avdeenko, A.V. Pustovalova // 2022 IEEE 23rd International Conference of Young Professionals in Electron Devices and Materials (EDM). – 2022. – P. 289-294.
- [5] Abyaa, A. Learner modelling: systematic review of the literature from the last 5 years / A. Abyaa, M. Khalidi, S. Bennani // Educational Technology Research and Development. – 2019. – Vol. 67(5). – P. 1105-1143.
- [6] Taraghi, B. Personal learning environment-a conceptual study / B. Taraghi, M. Ebner, G. Till, H. Mühlburger // Int. J. Emerg. Technol. Learn. – 2010. – Vol. 5(S11). – P. 25-30.
- [7] Pustovalova, N. Multivariate analysis of the influence of students' characteristics on academic performance / N. Pustovalova, T. Avdeenko // 2022 VIII International Conference on Information Technology and Nanotechnology (ITNT). – 2022. – P. 1-6.
- [8] Sanchez-Puchol, F. Towards an unified information systems reference model for higher education institutions / F. Sanchez-Puchol, J.A. Pastor-Collado, B. Borrell // Procedia computer science. – 2017. – Vol. 121. – P. 542-553.
- [9] Buder, J. Learning with personalized recommender systems: A psychological view / J. Buder, C. Schwind // Computers in Human Behavior. – 2012. – Vol. 28(1). – P. 207-216.
- [10] Tarus, J. K. A hybrid knowledge-based recommender system for e-learning based on ontology and sequential pattern mining / J. K. Tarus, Z. Niu, A. Yousif // Future Generation Computer Systems. – 2017. – Vol. 72. – P. 37-48.
- [11] Zhao, L. T. A recommendation system for effective learning strategies: An integrated approach using context-dependent DEA / L.-T. Zhao, D.-S. Wang, F.-Y. Liang, J. Chen // Expert Systems with Applications. – 2023. – Vol. 211. – P. 118535.
- [12] Klačnja-Milićević, A. E-Learning personalization based on hybrid recommendation strategy and learning style identification / A. Klačnja-Milićević, B. Vesin, M. Ivanović, Z. Budimac // Computers & education. – 2011. – Vol. 56(3). – P. 885-899.
- [13] Shishehchi, S. Ontological approach in knowledge based recommender system to develop the quality of e-learning system / S. Shishehchi, S. Banihashem, N.A.M. Zin, S. A. M. Noah // Australian Journal of Basic and Applied Sciences. – 2012. – Vol. 6(2). – P. 115-123.
- [14] Zuur, A. F. protocol for data exploration to avoid common statistical problems / A.F. Zuur, E.N. Ieno, C.S. Elphick // Methods in ecology and evolution. – 2010. – Vol. 1(1). – P. 3-14.
- [15] Akinwande, M. O. Variance inflation factor: as a condition for the inclusion of suppressor variable(s) in regression analysis / M.O. Akinwande, H.G. Dikko, A. Samson // Open Journal of Statistics. – 2015. – Vol. 5(7). – P. 754.

# Выявление проблемных вопросов по социально-направленным тематикам на основе данных открытых источников

О.К. Головнин  
Самарский университет  
Самара, Россия  
golovnin@ssau.ru

А.В. Кривошеев  
Самарский государственный  
технический университет  
Самара, Россия  
arkas19@gmail.com

И.Н. Дубинина  
Самарский государственный  
технический университет  
Самара, Россия  
vartaric@yandex.ru

П.В. Ситников  
ООО «Открытый код»  
Самара, Россия  
sitnikov@o-code.ru

А.В. Иващенко  
Самарский государственный  
медицинский университет  
Самара, Россия  
anton.ivashenko@gmail.com

**Аннотация**—В работе предложен подход к выявлению проблемных вопросов по данным открытых источников для социально-направленных тематик. Подход предполагает использование текстов и метаданных публикаций в открытых тематических группах и на публичных страницах пользователей в социальных сетях. Выполняется многоэтапная обработка данных по публикациям: сбор данных, очищение от спама, фильтрация по проблемной области, определение тональности, классификация по темам, кластеризация. Программная реализация предложенного подхода выполнена на основе Цифровой платформы интегрального мониторинга. Экспериментальная апробация проведена на данных социальной сети ВК. Применение подхода позволяет выявить остросоциальные проблемные вопросы административного региона в оперативном режиме.

**Ключевые слова**—большие данные, социальная сеть, классификация текстов, Text Mining.

## 1. ВВЕДЕНИЕ

Социально-экономическое развитие регионов России затрудняется без оперативного реагирования на возникающие остросоциальные вопросы [1]. С развитием различных платформ, обеспечивающих общение граждан посредством сети Интернет, в частности, социальных сетей и форумов, появились методы, модели и технологии, позволяющие извлекать информацию о качестве жизни и благополучии населения, а также по социальному самочувствию [2, 3]. В работе предложен подход к выявлению проблемных вопросов по данным открытых источников для социально-направленных тематик, использующий методы на основе искусственного интеллекта для обработки данных. Подход обеспечивает анализ обсуждаемых вопросов не только по тематикам, но и по роли в социально-экономическом развитии региона, выявляя не только проблемы, но и достижения.

## 2. МЕТОДОЛОГИЯ ИНТЕГРАЛЬНОГО МОНИТОРИНГА

Выявление проблемных вопросов по социально-направленным тематикам выполняется на основе анализа публикаций в открытых тематических группах и на публичных страницах пользователей в социальных сетях.

На *первом этапе* осуществляется сбор исходных данных, связанных с заданным анализируемым административным регионом. Исходные данные для исследования состоят из следующих атрибутов: текст

публикации, время и дата поста, местоположение пользователя, пол и возраст пользователя. Информация собирается за указанный период из тематических групп социальной направленности административного региона с вопросами по оказанию помощи, начислению единовременных выплат и т.п.

На *втором этапе* собранные данные очищаются от спама, для чего используется классификатор, относящий пост к классу «спам» или «не спам»:

1) Грубая оценка поста по заданным правилам (например, низкая доля русских букв в тексте, завышенная доля спецсимволов, короткие/длинные тексты, наличие стоп-слов и их сочетаний и т.п.);

2) Выполняется обработка анализируемых текстов через ряд функций. В качестве исходных данных для таких функций выступают следующие значения: общее количество символов в тексте, количество символов кириллицы, латиницы и цифр, число хештегов, смайлов, спецсимволов. Функции ранжируют данные числовые значения и выдают вероятность того, что текст является спамом. Все оценки суммируются; на выходе у каждого текста формируется итоговая оценка;

3) Тексты с высокой оценкой на спам получают метку «спам»;

4) Производится очистка оставшихся текстов от спецсимволов, смайлов, хештегов, HTML-кода, ссылок и иных нетекстовых включений;

5) Нейронная сеть «sentence-transformers/stsb-xlm-g-multilingual» переводит очищенные тексты в вектора («эмбединги»), размерностью 768 чисел;

6) Полученные вектора классифицируются заранее обученной глубинной нейронной сетью на классы: «спам» и «не спам».

На *третьем этапе* после завершения очистки набора данных от спама происходит поиск таких сообщений, которые содержат вхождения ключевых слов. Список ключевых слов (маркеров) составляется заранее на этапе подготовки к исследованию. Список включает в себя основные словоформы и ключевые фразы, относящиеся к теме исследования. Составление списка ключевых слов (маркеров) требует участия специалиста в заданной предметной области. Таким образом, публикация из социальной сети относится к социально-направленной теме, если выполняются следующие условия: публикация не определена классификатором спама как «спам»; публикация содержит в себе один или более

ключевых слов (маркеров) из заданной предметной области.

На *четвертом этапе* проводится анализ тональности каждого сообщения из полученной отфильтрованной выборки с помощью модели на основе BERT. Классификация текста производится в соответствии с пятью классами: 1 – негативный; 2 – позитивный; 3 – нейтральный; 4 – речь; 5 – неизвестно.

На *пятом этапе* осуществляется определение общих тем и направлений публикаций. На основе текстов публикаций и обращений определяются кластеры, объединенные общей тематикой, и этим кластерам присваиваются общие названия, обозначающие суть проблемы или обращений. Кластеризация тем осуществляется следующим образом:

1) Между публикациями пользователей на основе эмбедингов считается косинусная мера близости между векторами A и B:

$$\text{cosine\_measure} = 1 - \frac{A \cdot B}{\|A\| \|B\|} = 1 - \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}};$$

2) Если косинусная мера менее 0.25, то публикации пользователей определяются как схожие по смыслу или значению и объединяются в один кластер;

3) После объединения в кластеры производится итеративный перебор пар кластеров, причем, если у меньшего кластера более половины публикаций принадлежит большему кластеру, то меньший кластер объединяется с большим кластером;

4) В качестве действующих кластеров выбираются такие, чей размер составляет не менее  $l = \max(3, L^{0.25})$ , где L – количество постов после проверок на спам;

5) Внутри каждого кластера подсчитывается количество биграмм (словосочетаний) с учетом синтаксиса в предложениях; N=3 наиболее частых биграмм становится названием кластера.

6) Для каждого кластера производится оценка уровня тональности  $T = \max(\text{count}(T_n)) / n \cdot 1.5$ , где n соответствует номеру категории тональности текста.

Такой подход позволяет определить кластеры, их названия и тональности, с которыми далее осуществляют работу лица, принимающие решения. Таким образом, негативные кластеры считаются проблемными вопросами региона, а позитивные кластеры – наиболее значимыми достижениями региона, нейтральные – не содержат остросоциальные проблемы.

### 3. ПРОГРАММНАЯ РЕАЛИЗАЦИЯ И РЕЗУЛЬТАТЫ

Программная реализация предложенного подхода выполнена на языке Python в среде Цифровой платформы интегрального мониторинга [4, 5]. Исследование эффективности проведено на основе данных социальной сети VK за период в 3 мес. Сбор данных осуществлялся с помощью парсеров платформы через API. Так, за 3 мес. в анализируемом регионе выявлено около 700 тыс. постов, из них 480 тыс. постов классифицировано как спам. Из оставшихся 220 тыс. постов только 50 тыс. постов являются социально значимыми, из этих 50 тыс. постов относятся к

позитивным 4 тыс. постов, а к негативным – 18 тыс. постов. В результате анализа позитивных постов выделено 9 кластеров, из них 7 отмечено аналитиком как целевые – содержащие актуальную информацию. В отрицательных постах выявлено 45 кластеров, из них 21 отмечены аналитиком как целевые. В таблице 1 представлена матрица несоответствий. Точность подхода составляет 0,42, полнота – 0,88, значение F-меры в исследовании 0,57.

Таблица 1. МАТРИЦА НЕСООТВЕТСТВИЙ

Категория социально-значимого обращения		Экспертная оценка	
		Положительная оценка	Отрицательная оценка
Оценка метода	Положительная оценка	21 тыс.	29 тыс.
	Отрицательная оценка	3 тыс.	167 тыс.

Таким образом, не смотря на наличие неактуальных кластеров в результатах, всё же существенно снижается нагрузка на аналитика-оператора на анализ ключевых проблем региона, поскольку просмотр нескольких десятков тематик значительно менее трудоемок, чем поиск и просмотр исходных данных даже отдельно взятых публичных групп. В результате анализа определено, что нецелевые кластеры в негативных постах возникают в результате того, что пользователи социальных сетей обсуждают не реальные события или проблемы в регионе, а высказывают свое мнение о ситуации абстрактно или без конкретных фактов.

### 4. ЗАКЛЮЧЕНИЕ

Таким образом, в работе предложен подход к выявлению проблемных вопросов по данным открытых источников для социально-направленных тематик. Подход программно реализован на основе Цифровой платформы интегрального мониторинга, применение которой позволяет снизить нагрузку на аналитика-оператора, выполняющего анализ ключевых проблем региона, за счет существенного сокращения количества просматриваемой информации.

### ЛИТЕРАТУРА

- [1] Аганбегян, А.Г. Анализ и прогнозирование социально-экономического развития регионов (методические заметки) / А.Г. Аганбегян // Среднерусский вестник общественных наук. – 2019. – Т. 14, № 4. – С. 15-28.
- [2] Овчар, Н.А. Технологии исследования социального самочувствия горожан на основе анализа web-контента / Н.А. Овчар, А.С. Воробьев, Д.С. Парыгин, Н.П. Садовникова // Системный анализ в науке и образовании. – 2019. – Т. 1. – С. 83-92.
- [3] Щекотин, Е.В. Цифровые следы как новый источник данных о качестве жизни и благополучии: обзор современных тенденций / Е.В. Щекотин // Вестник ТомГУ. – 2021. – Т. 467. – С. 170-181.
- [4] Сурнин, О.Л. Применение цифровой платформы интегрального мониторинга как средства бизнес-аналитики социально-экономического развития региона / О.Л. Сурнин, П.В. Ситников, А.В. Ивашенко, О.К. Головин [и др.] // Информ. технологии в управлении: сб. материалов. – СПб.: СПбГЭТУ ЛЭТИ, 2022. – С. 158-161.
- [5] Ситников, П.В. Анализ социально-экономического развития региона на базе цифровой платформы интегрального мониторинга / П.В. Ситников, Е.А. Додонова, И.Н. Дубинина [и др.] // Информ. технологии и нанотехнол.: тр. конф. – Самара: Смп. ун-т, 2022. – С. 052032.

# Использование свойств вейвлет-преобразования в задачах поиска закономерностей

Е.А.Нелюбина

Калининградский государственный  
технический университет  
Калининград, Россия  
e-mail: e.nelubina@gmail.com

В.В.Рязанов

Федеральный исследовательский  
центр "Информатика и управление"  
Российской академии наук  
Москва, Россия  
e-mail: rvvccas@mail.ru

А.П.Виноградов

Федеральный исследовательский  
центр "Информатика и управление"  
Российской академии наук  
Москва, Россия  
e-mail: vngrccas@mail.ru

**Аннотация**—Представлен подход к проблеме поиска закономерностей в данных, основанный на использовании обобщенных прецедентов и адаптации элементов вейвлет-преобразования в повышенных размерностях, приведены примеры решения прикладных задач с этих позиций.

**Ключевые слова**— закономерность, базовый кластер, обобщенный прецедент, вейвлет, преобразование Хафа

## 1. ВВЕДЕНИЕ

В настоящее время в области анализа изображений успешно используется множество подходов и методов, которые эффективны в малых размерностях [1], [2], но их применение в случае многомерных данных затруднено [3], [4], [5]. Далее центральным объектом будет закономерность, описываемая малым числом параметров [6]. Известно, что понятие закономерности является весьма сложным. Мы не касаемся здесь общих вопросов и исследуем ниже параметрические гипотезы о закономерности, используя лишь компетенции в той или иной предметной области и в области ИТ. Ранее был выполнен ряд исследований [6], [7], [8], где при построении гипотезы о закономерности происходит выбор параметрического пространства  $Y$  нужного вида аналогично тому, как это имеет место в схеме преобразования Хафа. Возникающая в  $Y$  вторичная кластерная структура  $c^t \in C^T$  в этом случае содержит информацию о повторяемости закономерности и о типичных значениях параметров. Каждая реализация закономерности (как и представляющий её вектор в пространстве параметров) называется обобщенным прецедентом (ОП), т.е., прецедентом закономерности [6].

## 2. МОДЕЛЬ ЗАКОНОМЕРНОСТИ

Пусть  $X, X \subset R^N$  - выборка оцифрованных данных

большого объёма. Основным объектом будет служить кортеж точек  $x = \{x^1, \dots, x^M\}$ ,  $x^m \in X$ ,  $m = 1, \dots, M$ , для которого выполняются условия, сформулированные экспертом:

$$P_l(x^1, \dots, x^M) = P_l(x^1_1, \dots, x^1_N, \dots, x^M_1, \dots, x^M_N), l = 1, 2, \dots, L.$$

Набор условий  $P = \{P_1, \dots, P_L\}$  представляет собой формулировку гипотезы о наличии закономерности. Если условиями  $P$  задана некоторая пространственная форма расположения точек кортежа  $x = \{x^1, \dots, x^M\}$  в пространстве признаков, то она называется базовым кластером. Для проверки гипотезы требуется выполнить перебор всех вариантов  $x \subset X$ , если нужно обнаружить максимальное число подтверждений закономерности.

Например, если  $P = \{P_1, \dots, P_L\}$ ,  $L = N - 2$ , - система полиномиальных уравнений, то тогда в пространстве  $R^N$  определена двумерная алгебраическая поверхность  $R$ ,  $R \subset R^N$ , и проявление закономерности в точке

$x \in X$  соответствует выполнению условия  $x \in R$ . Если, к тому же, эта система приводится к виду  $x_l = F_l(x^1_1, x^1_2)$ ,  $l = 3, \dots, N$ , и функции  $F_3, \dots, F_N$  не имеют особенностей, то проекция выборки  $X$  на поверхность  $R$  может быть взаимно однозначно отображена на плоскость  $(x^1_1, x^1_2)$ , в частности, на экран монитора. Пример весьма условный, но он указывает на возможный способ задействования зрительной компетенции и/или интуиции эксперта.

## 3. ПРИМЕРЫ ПРИМЕНЕНИЯ МОДЕЛИ

Нас интересуют задачи, где комбинации параметров могут выступать в роли малой волны. В частности, это может быть сам кортеж  $x = \{x^1, \dots, x^M\}$ ,  $x^m \in X$ ,  $m = 1, \dots, M$ , описывающий форму базового кластера. Пусть  $F = (F_1, \dots, F_N)$  - список сложных объектов,  $f = (f_1, \dots, f_N)$ , - перечень вариантов их поведения. Требуется реконструировать  $f$  по имеющейся статистике, когда для наблюдения доступны лишь некоторые интегральные параметры вида  $X(F)$ . Построенная система  $c^t \in C^T$  будет содержать информацию о правильности понимания экспертом поведения объектов  $F = (F_1, \dots, F_N)$ , а также о типичных значениях параметров  $f = (f_1, \dots, f_N)$ .

Данная постановка рассматривалась в [8]. Список  $F$  представлял закономерности стока в регионах бассейна реки, где измерение стока затруднено или невозможно. Выборка представляла собой временной ряд  $x(t) = (x_0, x_1(t), \dots, x_N(t))$ , признаковое пространство  $R^{N+1}$  содержало вычисленные по осадкам уровни влаги в регионах  $(x_1(t), \dots, x_N(t))$ , а также объем стока реки в устье  $x_0(t)$  как единственный измеряемый параметр.

Закономерность стока  $i$ -го региона определяется как функция зависимости расхода от уровня влаги  $x_i$ . Её аппроксимацией может служить вектор  $f_i = (f_i^1, f_i^2, \dots, f_i^N)$ , построенный из дискретных значений расхода в фиксированных точках на шкале уровня влаги. Этот способ не единственный (Рис.1), но предполагается, что он одинаков для всех индексов  $i$ , и  $f$  теперь представляет собой  $N \times D$  матрицу  $f = \{f_i^d\}$ , где  $D$  - число параметров объекта  $F_i$ , в нашем случае, позиций на шкале уровня. Проверять гипотезу  $f = \{f_i^d\}$  необходимо во всех точках ряда  $x(t) = (x_0, x_1(t), \dots, x_N(t))$ , где  $\lambda$  - некоторый порог подтверждения:  $\sum (f_i(t) - \sum f_i^d(t))^2 < \lambda$ .

Форма функции стока служит аналогом малой волны, которая реализуется в различные моменты  $t$ . Эта форма может проявиться в чистом виде на временной шкале,

например, когда наблюдается одномоментное большое повышение уровня  $x_i(i^*)$  и последующий период свободного стока. Как было отмечено, даже в таких случаях прямое измерение стока не всегда возможно, и адекватность представления функции стока  $i$ -го региона вектором  $f_i$  будет проявляться в виде кластера плотности реализаций  $f$  как ОП в окрестности гиперплоскости  $f_i$ . Заметим, что тем самым определяются не только точки реализаций, но и правильные варианты формы волны. Например, если пользователя интересуют лишь различия формы стока с точки зрения паводковой опасности, то могут быть приемлемыми 2-параметрическое описание

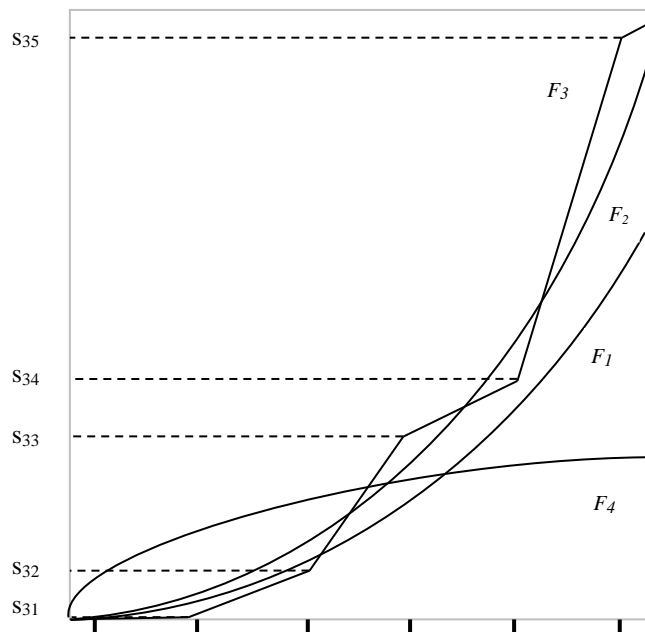


Рис. 1. Параметрическое представление формы числовой зависимости.  $F_1$  и  $F_2$  соответствуют представлению вида  $ax^b$  для выпуклой функции,  $b>1$ ; для  $F_4$  – также  $ax^b$ , но для вогнутой функции,  $b<1$ ; вариант  $F_3$  – кусочно-линейная аппроксимация в виде вектора значений  $s_{id}$  функции в фиксированных точках

вида  $ax^b$  и классификация регионов лишь по знаку неравенства:  $b<1$  или  $b>1$ .

Пусть теперь  $X, X \subset R^N$  – выборка записей параметров индивидов в популяции, например, антропометрических показателей для жителей страны, города, и т.д. Если эксперта-практика интересует отношение «предок-потомок» для  $x^1, x^2$  из  $X, x^1 \neq x^2$ , то определим меру близости кортежа к носителю закономерности сходства:

$$\rho(x, R) = \sum_1^N \sigma_n, \quad \sigma_n = \begin{cases} 1, & |x_n^1 - x_n^2| \leq \varepsilon_n \\ 0, & |x_n^1 - x_n^2| > \varepsilon_n \end{cases}$$

где  $\varepsilon_n$  – набор границ на шкалах антропометрических показателей. Включим условие  $x^1 \neq x^2$  в список  $P$  и будем отображать отметки о наличии закономерности точками на плоскости  $(\rho, \chi)$ , где  $\chi = |x_n^1 - x_n^2|, n'$  – параметр «возраста». Тогда в окрестности некоторых точек  $(\rho=w, \chi \approx 25), 1 < w \leq N$ , на данной плоскости будут наблюдаться вторичные кластеры с повышенной плотностью отметок, поскольку антропометрические

показатели в паре «предок-потомок» часто близки. Представим теперь в виде малой волны гипотезу о возрастных изменениях, приспособив определение меры близости для сравнения двух кортежей:

$$\rho(x, x^*) = \sum_1^N \sigma_n, \quad \sigma_n = \begin{cases} 1, & |x_n^1 + \alpha_n(x) - x_n^2| \leq \varepsilon_n \\ 0, & |x_n^1 + \alpha_n(x) - x_n^2| > \varepsilon_n \end{cases}$$

где вектор-функция  $\alpha(x)$  выступает в роли малой волны, задающей коррекцию эталонного кортежа  $x^*$  согласно возрасту. Постановка  $\alpha(x)=0$  соответствовала ситуации идеального сходства предка и потомка.

Отметим, что в  $x$  содержатся абсолютные координаты объектов в пространстве  $R^N$ , и они могут использоваться для формирования  $Y$  как подходящего параметрического пространства, в котором представлены также различные координатные аспекты поведения ОП.

#### 4. ЗАКЛЮЧЕНИЕ

В работе представлен новый подход к проблеме поиска закономерностей в прикладных данных, в основе которого лежит понятие ОП. Показано, что при этом полезным также оказывается использование некоторых свойств вейвлет-преобразования. В целом, при работе со сложными данными, центральным моментом и в этом случае оказывается адекватное встраивание прикладных компетенций в числовые модели.

#### БЛАГОДАРНОСТИ

Работа выполнена при частичной поддержке РФФИ, проект 20-01-00609.

#### ЛИТЕРАТУРА

- [1] Davies, E.R. Advanced Methods and Deep Learning in Computer Vision / E.R. Davies, M. A. Turk – Oxford: Elsevier, 2022. –562 p. doi.org/10.1016/B978-0-12-822109-9.00002-3.
- [2] Nixon, M.S. Feature Extraction for Image Processing and Computer Vision / M.S. Nixon, A.S. Aguado – Elsevier Ltd., 2020. – 626 p. DOI: 10.1016/C2017-0-02153-5.
- [3] Аникеев, Ф.А. Эффективная реализация быстрого преобразования Хафа с использованием сопроцессора СРСА / Ф.А. Аникеев, Г.О. Райко, Е.Е. Лимонова, М.А. Алиев, Д.П. Николаев // Программирование. – 2021. – Т. 5. – С. 3-11. DOI: 10.31857/S0132347421050022.
- [4] Rinoshika, A. Application of multi-dimensional wavelet transform to fluid mechanics / A. Rinoshika, H. Rinoshika // Theoretical and Applied Mechanics Letters. – 2020. – Vol. 10(2). – P. 98-115. DOI: 10.1016/j.taml.2020.01.017
- [5] De Mauro, A. Understanding Big Data Through a Systematic Literature Review: The ITMI Model / A. De Mauro, M. Greco, M. Grimaldi // International Journal of Information Technology & Decision Making. – 2019. – Vol. 18(4), – P. 1433-1461
- [6] Ryazanov, V. Analogues of Image Analysis Tools in the Problems of Finding Latent Regularities in Big Applied Data / V. Ryazanov, A. Vinogradov // Pattern Recognition and Image Analysis. – 2022. – Vol.32(3). – P. 639-644. DOI: 10.1134/S105466182203035X.
- [7] Ryazanov, V. Dealing with Realizations of Hidden Regularities in Data as Independent Generalized Precedents," / V. Ryazanov, A. Vinogradov // IEEE Xplore Proceedings of 2021 International Conference on Information Technology and Nanotechnology (ITNT). – 2021. – P. 1-3. DOI: 10.1109/ITNT52450.2021.
- [8] Naumov, V.A. Analysis and prediction of hydrological series based on generalized precedents / V.A. Naumov, E.A. Nelyubina, V.V. Ryazanov, A.P. Vinogradov // Book of abstracts of the 12-th Int. Conf. Intelligent Data Processing (IDP-12). – 2018. – P.178-179.

# Алгоритм обнаружения и выделения сигналов в сильно зашумленных потоках данных

В.А. Засов

Самарский государственный университет путей сообщения  
Самара, Россия  
vzasov@mail.ru

М.В. Ромкин

Научно производственный центр «ИНФОТРАНС»  
Самара, Россия  
romkinmaks@rambler.ru

**Аннотация** — Предложено адаптивное устройство обнаружения и выделения сигналов в сильно зашумленных потоках данных. Особенностью устройства является прогнозирование интервалов адаптации, позволяющее работать в режиме потоковой обработки данных.

**Ключевые слова** – алгоритм, адаптивный, помехи, обнаружение, прогнозирование, интервал, поток

## 1. ВВЕДЕНИЕ

При извлечении информации из сильно зашумленных потоков данных актуальной задачей является обнаружение и выделение сигналов из аддитивной сигнала смеси с помехами. Предложенные в [1,2] адаптивные подавители помех (АПП) с адаптацией только в интервалах между импульсами полезного сигнала позволяют подавлять мощные коррелированные с полезными импульсными сигналами помехи. Эти АПП работают в режиме пакетной обработки данных, что ограничивает их применение в системах реального времени. Применяемый в АПП алгоритм обучения, прерывает на время обучения передачу обработанных пакетов сигналов на выход АПП, что ограничивает функциональные возможности устройства.

В работе предлагается алгоритм и реализующее его адаптивное устройство обнаружения и выделения сигналов (АОВС) в сильно зашумленных потоках данных работающее в режиме потоковой обработки данных без перерывов потока во время обучения в случаях изменения параметров импульсных полезных сигналов.

## 2. СТРУКТУРНАЯ СХЕМА И АЛГОРИТМ РАБОТЫ АДАПТИВНОГО УСТРОЙСТВА ОБНАРУЖЕНИЯ И ВЫДЕЛЕНИЯ СИГНАЛОВ

Структурная схема предлагаемого АОВС приведена на рис. 1.

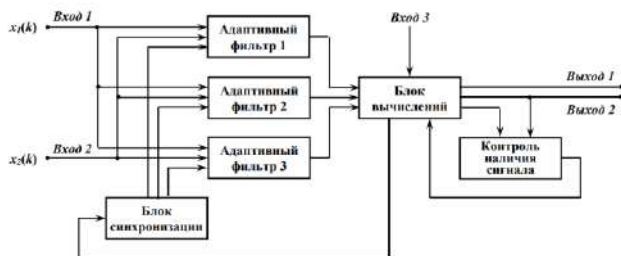


Рис. 1. Адаптивное устройство обнаружения и выделения сигналов (АОВС)

На вход 1 устройства поступает аддитивная смесь  $x_1(k)$  полезного импульсного сигнала и помех (коррелированных с полезным сигналом и шума), на вход 2 – сумма  $x_2(k)$  коррелированных с полезным сигналом помех и шума.

Предлагаемое в работе устройство отличается от известных [1,2] тем, что вычисление моментов времени начала интервалов адаптации производится в процессе приема зашумленного потока сигналов  $x_1(k)$  после обнаружения импульсных сигналов в потоке. Таким образом, исключается временная задержка, обусловленная в известных устройствах [1,2] обработкой записанных в память пакетов входных сигналов, и обеспечивается масштаб реального времени. Для устранения перерывов потока сигналов на выходе 2 и надежной привязки интервала адаптации к интервалам между импульсами предлагается использовать три адаптивных фильтра АФ1, АФ2 и АФ3 и формировать три тестовых импульса ТИ1, ТИ2 и ТИ3, причем ТИ2 в центре группы, ТИ1 опережает ТИ2 и ТИ3 отстает от ТИ2. Тестовые импульсы ТИ определяют интервалы адаптации АФ1, АФ2 и АФ3 в потоке входных сигналов. Сравнивая нормированные мощности сигналов на выходах пар фильтров АФ1 и АФ2 или АФ2 и АФ3 можно определить направление смещения во времени интервалов адаптации для надежного попадания в интервалы между импульсами. Алгоритм работы АОВС описывается следующими шагами.

**Шаг 1:** Проверка наличия или отсутствия полезного импульсного сигнала (обнаружение сигнала) в зашумленном потоке данных  $x_1(k)$ .

Проверка осуществляется путем сравнения в интервалах времени, задаваемыми контрольными импульсами КИ1, КИ2 и КИ3, нормированных мощностей сигналов на выходах АФ1, АФ2 и АФ3. Возможны следующие состояния: попадание интервалов ТИ для фильтров АФ на фронт или срез импульсного полезного сигнала, попадание интервалов ТИ на импульс или интервал между импульсами полезного сигнала. Только состояние, характеризующее равенством нормированных мощностей сигналов на выходах трёх фильтров АФ1, АФ2 и АФ3 указывает на гарантированное попадание интервалов адаптации ТИ на импульс или интервал между импульсами полезного сигнала. При неравенстве нормированных мощностей сигналов на выходах фильтров АФ1, АФ2 и АФ3 осуществляется повторная адаптация со смещением во времени.

Далее производится определение по уровню мощности выходов АФ1, АФ2 и АФ3 ступенчатых изменений в сигналах. При наличии ступенчатых изменений (наличия фронтов и срезов) принимается решение об обнаружении импульсных полезных сигналов в потоке данных на входе 1 и осуществляется переход к шагу 2. При отсутствии ступенчатых изменений в течение определенного временного интервала принимается решение об отсутствии импульсных полезных сигналов в потоке данных, выдается сообщение на выход 1 АОВС и шаг 1 повторяется;

**Шаг 2:** На основе определенных ступенчатых

изменений сигналов производится измерение периодов сигналов, длительностей импульсов и вычисление скважности, значения которых записываются в память АОВС и передаются на выход 3.

Если вычисленная в блоке вычислений скважность равно величине скважности, заданной на входе 3 АОВС, принимается решение о попадании интервалов адаптации АФ1, АФ2 и АФ3 в интервал между импульсами. В противном случае принимается решение о попадании интервалов адаптации АФ1, АФ2 и АФ3 на импульс сигнала. Далее осуществляется переход к шагу 3.

**Шаг 3:** Производится вычисление прогнозируемого времени начала интервала адаптации  $t_{прог}$ .

Это время рассчитывается от начала определенного на шаге 2 периода импульсного полезного сигнала. При попадании интервалов адаптации АФ1, АФ2 и АФ3 в интервал между импульсами паузу сигнала  $t_{прог}$  равно

$$t_{прог} = t_{имп} + \frac{t_{пер} - t_{имп}}{2} - 2 \cdot t_{ТИ},$$

где  $t_{имп}$  – длительность импульсов,  $t_{пер}$  – период импульсов,  $t_{ТИ}$  – длительность интервала адаптации АФ.

Если интервал адаптации АФ1, АФ2 и АФ3 попадает на импульс время  $t_{прог}$  вычисляется так

$$t_{прог} = \frac{t_{имп}}{2} - 2 \cdot t_{ТИ}.$$

Вычисленное время  $t_{прог}$  записывается в таймер блока синхронизации и прогнозирует время, когда на входе 1 АОВС появится интервал между импульсами. Тогда по сигналу таймера блока синхронизации запускается процесс адаптации и вычисляются весовые коэффициенты АФ1, АФ2 и АФ3. После завершения адаптации осуществляется переход к шагу 4.

**Шаг 4:** Производится вычисление  $t_{прог}$  начала очередного интервала адаптации на основе сравнения нормированных мощностей сигналов на выходах АФ1, АФ2 и АФ3. Если эти величины равны, это указывает на ситуацию, при которой интервал адаптации находится внутри интервала между импульсами. В этом случае время  $t_{прог}$  начала адаптации, определенное на шаге 3, не изменяется.

Если величины нормированных мощностей на выходах АФ1 и АФ2 или АФ2 и АФ3 не равны, то время начала адаптации находится на границе интервала между импульсами. В первом случае время начала адаптации, вычисленное на шаге 3, требуется уменьшить, во втором – увеличить на величину длительности ТИ, после чего записать  $t_{прог}$  в таймер блока синхронизации и перейти к шагу 5.

**Шаг 5:** Производится сравнение параметров (периода сигналов и длительности импульса) выделенного импульсного полезного сигнала с выхода 2 с записанными в память на шаге 3 такими же параметрами. При равенстве параметров переходим на шаг 4, при неравенстве параметров переходим на шаг 1.

### 3. РЕЗУЛЬТАТЫ КОМПЬЮТЕРНОГО МОДЕЛИРОВАНИЯ

Эффективность предложенных в работе решений подтверждается результатами компьютерного моделирования приведенными на рис. 2. При

моделировании применялся алгоритм адаптации Recursive Least Squares (RLS) [3,4], число весовых коэффициентов АФ1, АФ2, АФ3 равно 32.

На рис. 2 приведены: а) – сигнал  $x_I(k)$  на входе 1

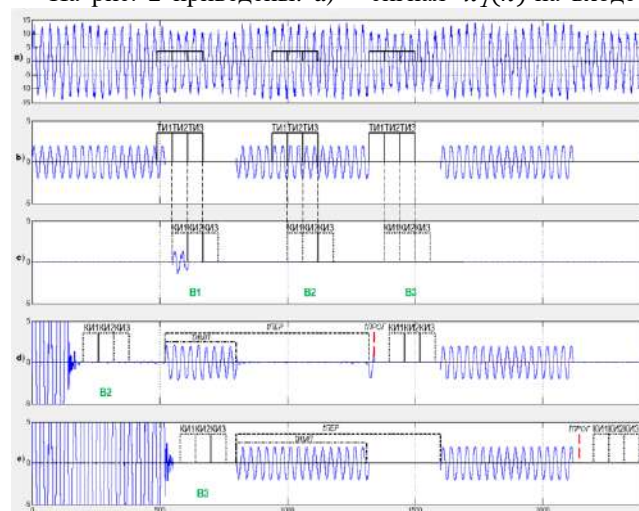


Рис. 2. Результаты компьютерного моделирования адаптивного устройства обнаружения и выделения сигналов

устройства, представляющий собой аддитивная смесь импульсного полезного сигнала амплитудой с коррелированной помехой и белым шумом, уровень; б) – варианты возможного попадания тестовых импульсов ТИ1, ТИ2 и ТИ3 соответственно на срез импульса полезного сигнала в зашумленном сигнале  $x_I(k)$  (В1), на импульс полезного сигнала в сигнале  $x_I(k)$  (В2) и паузу полезного сигнала в сигнале  $x_I(k)$  (В3); в) – положения контрольных импульсов КИ1, КИ2 и КИ3 на выходах соответствующих адаптивных фильтров для ситуаций В1, В2 и В3; д) – определение прогнозируемого момента времени начала интервала адаптации  $t_{прог}$  (шаг 3 алгоритма) для ситуации В2; е) – определение прогнозируемого времени начала интервала адаптации  $t_{прог}$  (шаг 3 алгоритма) для ситуации В3.

### ЗАКЛЮЧЕНИЕ

Разработаны алгоритм и реализующее его адаптивное устройство обнаружения и выделения сигналов в сильно зашумленных потоках данных. Прогнозирование при вычислении интервалов адаптации обеспечивает работу устройства в режиме потоковой обработки данных без перерывов выходного потока во время обучения в случаях изменения параметров полезных сигналов.

### ЛИТЕРАТУРА

- [1] Засов, В.А. Адаптивный компенсатор помех в импульсных сигналах / В.А. Засов, М.В. Ромкин// Патент на изобретение RU №2735671 от 22.10.2019.
- [2] Zasov, V. Adaptive Cancellation of Interference in Intermittent and Pulse Signals / V. Zasov, M. Romkin// Data Science. Information Technology and Nanotechnology. Proc. of the Int. Conf. ITNT-2021. –2021. IEEEExplore, 2021. DOI: 10.1109/ITNT52450.2021.9649169.
- [3] Haykin, S. Adaptive filter theory (4<sup>th</sup> ed.) / S. Haykin – Prentice Hall, 2001. – 936 p.
- [4] Джиган, В.И. Адаптивная фильтрация сигналов: Теория и алгоритмы / В.И. Джиган. – М.: Техносфера, 2013. – 528 с.

# Сравнение эффективности методов машинного обучения в задаче оценки стоимости недвижимости

Е.О. Агафонова

Самарский национальный исследовательский университет  
им. академика С.П. Королева  
Самара, Россия  
super.kia.140401@gmail.com

А.А. Белоусов

Самарский национальный исследовательский университет  
им. академика С.П. Королева  
Самара, Россия  
adark@narod.ru

**Аннотация**—В данной статье рассматривается проблема оценки стоимости жилой недвижимости. Были собраны и обработаны данные о продаже трехкомнатных квартир в г. Самара. Для решения проблемы оценки были использованы методы машинного обучения Random Forest и Gradient Boosting, среди которых выбран наиболее эффективный.

**Ключевые слова**— машинное обучение, оценка стоимости недвижимости, Random Forest, Gradient Boosting, обработка данных

## 1. ВВЕДЕНИЕ

Рынок недвижимости России является одной из самых динамичных сфер российской экономики. Учитывая огромное количество как внутренних, так и внешних факторов, влияющих на стоимость объектов недвижимости, вопрос оценки жилого имущества играет ключевую роль в сделке купли-продажи [1]. Однако расчет оценки недвижимости достаточно трудоемкая задача. Для решения данной проблемы могут использоваться методы машинного обучения, которые позволяют с высокой точностью определить стоимость недвижимости. Что подтверждает актуальность данного исследования.

Целью исследования является сравнение методов машинного обучения для задачи оценки стоимости жилой недвижимости в г. Самара и определение их эффективности.

## 2. ОСОБЕННОСТИ ПОДГОТОВКИ ДАННЫХ ДЛЯ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Для исследования была выбрана вторичная недвижимость, расположенная в городе Самара. Данные для исследования были получены с интернет-сервиса для размещения объявлений о недвижимости «Авито Недвижимость». Данный сервис был выбран в силу своей популярности у пользователей. На нем размещено наибольшее количество интересующих нас объявлений.

Методом скрапинга были получены данные о продаже 1705 трехкомнатных квартир. Сбор данных был осуществлен при помощи расширения для Google Chrome “Web Scraper”. Данный инструмент позволяет создавать карту сайта из различных типов селекторов, извлекать данные с сайтов с несколькими уровнями навигации.

Точность оценки зависит от набора ценообразующих параметров, с помощью которых происходит идентификация объекта оценки. Определение состава параметров для описания каждого объекта является значимой составляющей формирования исходных данных. Выделим существенные характеристики

объекта, определяющие его рыночную стоимость: числовые переменные – площадь квартиры, жилая площадь, площадь кухни, этаж расположения, количество этажей в доме, высота потолков, год постройки; категориальные переменные – балкон/лоджия, тип санузла, качество ремонта, вид из окна, район, тип дома.

Существенную роль в оценке недвижимости играет не только количество параметров, но и степень влияния каждого из них на рыночную стоимость квартиры. Проведенный анализ данных позволил установить значимость каждого из используемых признаков с точки зрения его влияния на стоимость объекта недвижимости (рисунок 1). Значимость признаков считается при помощи встроенного в алгоритм построения ансамбля деревьев метода feature\_importances. Он основан на вычислении суммарного уменьшения минимизируемого функционала ошибки с помощью ветвлений по рассматриваемому признаку.

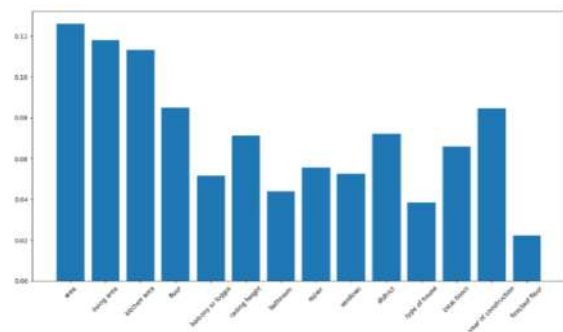


Рис. 1. Значимость признаков с точки зрения его влияния на стоимость объекта

Анализируя полученные результаты можно сделать выводы: наибольшее влияние на стоимость недвижимости оказывает площадь квартиры, жилая площадь и площадь кухни; менее значимыми оказались признаки расположения квартиры на первом или последнем этаже, тип дома и тип санузла.

После создания датасета его необходимо подготовить для методов машинного обучения. Целевой переменной является price. Для корректной работы методов машинного обучения необходимо восстановить пропущенные значения, в нашем случае ставится среднее значение или строка удаляется полностью, если характеристика является качественной. Потом находятся выбросы (результат измерения, сильно выбивающийся из выборки) и удаляются из набора данных. Значения числовых признаков подвергаются типизации, а категориальные признаки – стандартизации (убираются лишние символы и знаки препинания, исправляются

орфографические ошибки, используются только строчные буквы). Затем происходит кодирование категориальных данных методом маркировки – процесс, который ставит в соответствие числовым порядковые значения. Именно маркировка позволяет сохранить порядок, присвоив целочисленные значения, начинающиеся с 0 для значения самого низкого порядка, 1 для следующего порядка и так далее. Методы, которые будут применяться не требуют нормализации данных. Последним шагом в подготовке данных к обучению является разделение выборки на две части: обучающую (80%) и тестовую (20%). Обучающая выборка использовалась для обучения моделей, а тестовая - для определения качества их предсказания.

### 3. РЕЗУЛЬТАТЫ РАБОТЫ МЕТОДОВ

Была написана программа на языке программирования Python, реализующая методы Random Forest и Gradient Boosting. Данные модели были обучены на тестовой выборке, которая составляет 20% от выборки. Для оценки качества моделей использовались следующие метрики:  $R^2$  – коэффициент детерминации, MAE – средняя абсолютная ошибка, MSE – средняя квадратичная ошибка [2].

Для получения наибольшей точности предсказаний в моделях необходимо настроить гиперпараметры. Опытным путем установлено, что наилучший результат предсказания на обучающей выборке был получен при следующих значениях гиперпараметров: `max_features = 0,75`, `min_samples_leaf = 1`, `n_estimators = 150` для Random Forest и `learning_rate = 0,1`, `max_depth = 3`, `n_estimators = 200` для Gradient Boosting.

Результаты расчетов точности на тестовой выборке для методов Random Forest и Gradient Boosting представлены в таблице I.

Таблица I. РЕЗУЛЬТАТЫ РАСЧЕТОВ ТОЧНОСТИ НА ТЕСТОВОЙ ВЫБОРКЕ ДЛЯ РАЗНЫХ МЕТОДОВ

Метод	$R^2$	MAE	MSE
Random Forest	0,991212	182577,08	209110839446,94
Gradient Boosting	0,990235	213013,84	232343126319,58

Анализируя таблицу, можно заметить, что Random Forest превосходит по точности Gradient Boosting по всем метрикам.

Определим зависимость точности оценки от параметров метода обучения, показавшего лучшие результаты - Random Forest.

Минимальное количество объектов в листе способствует борьбе с переобучением. Дерево строится до тех пор, пока количество объектов в листьях остается более заданного пользователем минимального числа. Когда в листе остается один объект, это может привести как к очень точной оценке, так и к большой ошибке. В то же время чем больше объектов в листе, тем больше их разнородность, что тоже может привести к ошибочной оценке. На рисунке 2 заметим, что высокое качество оценки может быть достигнуто, когда в листе не более шести объектов.

Чем больше деревьев решений, тем точнее должен быть результат. Однако на рисунке 3 заметно, что при достижении числа деревьев, равного десяти значение метрик качества меняется незначительно.

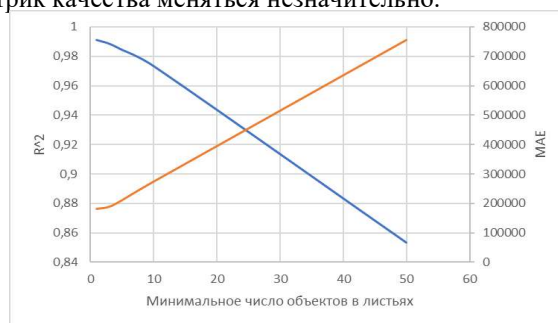


Рис. 2. Зависимость  $R^2$  и MAE от числа объектов в листьях

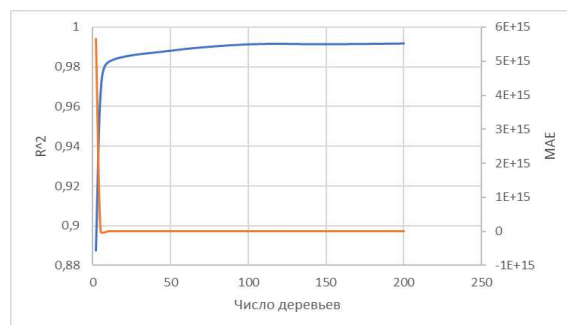


Рис. 3. Зависимость  $R^2$  и MAE от числа деревьев

### ЗАКЛЮЧЕНИЕ

В статье была рассмотрена проблема оценки стоимости жилой недвижимости по ее характеристикам. Для решения этой задачи были реализованы методы машинного обучения Random Forest и Gradient Boosting. Сравнение методов по качеству, которое измерялось с помощью метрик регрессии, показало, что наибольшая точность предсказания у Random Forest. Были определены зависимости точности оценки от параметров метода Random Forest. Графики зависимостей показали, что высокое качество оценки может быть достигнуто, когда в листе не более шести объектов и при числе деревьев равном десяти. В ходе исследований были собраны и проанализированы данные, размещенные на сайте «Авито Недвижимость» о продаже жилой недвижимости в г. Самара, которые были использованы для обучения методов машинного обучения. Научная новизна исследования состоит в применении методов машинного обучения для решения задачи оценки стоимости трехкомнатных квартир в г. Самара.

### ЛИТЕРАТУРА

- [1] Сурков, Ф.А. Сравнение временных рядов и нейросетевых методов в задаче прогнозирования стоимости и оценки недвижимости / Ф.А. Сурков, Н.В. Петкова, С.Ф. Суховский // Моделирование, оптимизация и информационные технологии. – 2018. – Т.6, №3.
- [2] Chugh, A. MAE, MSE, RMSE, Coefficient of Determination, Adjusted R Squared — Which Metric is Better? [Electronic resource]. — Access mode: <https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e> (15.10.2021).

# Annotation of mathematical formulas in PDF documents

Konstantin Nikolaev  
*Federal State Institution of the Federal Research Center NIISI*  
RAS  
Kazan, Russia  
konnikolaeff@yandex.ru

Olga Nevzorova  
*Kazan Federal University*  
Kazan, Russia  
onevzoro@gmail.com

**Abstract**—This article provides an overview of existing solutions for semantic analysis of mathematical documents, and also presents a method for automatic semantic analysis of documents in PDF format. This method searches for local variables in the text of the article, extracts their definitions and connects concepts with formulas. The advantage of the method over the existing ones is independence from the markup of the original PDF document, which expands the scope of the method. We provide estimates of recall, precision and F-measure for algorithms for finding variables and linking local variables with formulas. The resulting semantic markup of the document will be used to create a collection of documents suitable for the semantic formula search service, which is part of the set of services of the Lobachevskii-DML digital publishing system.

**Keywords**—*semantic analysis, PDF, document processing, scientific journals, Lobachevskii-DML*

## I. INTRODUCTION

Semantic search is focused on searching in collections of semantic publications, which are documents with semantic markup of text components. Mathematical texts are highly structured, with the presence of fixed semantics components, such as theorems, proofs, formulas, etc.

The task of searching for documents on mathematical formulas is relevant for conducting scientific research, preparing articles, studying mathematical disciplines. The paper [1] describes a semantic search engine for mathematical formulas, which uses a data set based on a collection of scientific articles of the journal "Izvestiya Vuzov. Mathematics" for 1997-2009. This article discusses new improved algorithms for constructing a data set for semantic search by mathematical formulas, which will qualitatively improve the search results.

Mathematical search by formulas can be divided into two categories – search for formulas by structure and by content. Searching for formulas by structure comes down to getting a list of formulas that partially or completely match the structure of the formula specified in the search query. This approach does not take into account the semantics of formulas.

More effective, but difficult to implement, is the search for mathematical articles on the content of the formula. To determine the content of the formula, it is necessary to identify the variables of that formula, and to define mathematical concepts denoted by variables. Additional difficulties are caused by a variety of templates for the design of mathematical documents for different information systems of scientific journals. Currently, the most popular formats for presenting mathematical formulas in scientific articles are: graphic image (articles in pdf format); formulas in Microsoft Word editor; LaTeX format; MathML format. Collections of articles in digital mathematical libraries are presented mainly in PDF format. Text recognition, and, in

particular, mathematical expressions, is the main task of this study. Mathematical formulas extracted during recognition in scientific articles and their descriptions are the source data for building an improved data set for a mathematical search engine.

## II. METHOD OF SEMANTIC ANNOTATION OF FORMULAS IN A PDF DOCUMENT

Most of the existing solutions of semantic annotation of documents strongly depend on the input document, its format and structure [2-6]. The article proposes a universal method for determining the structure of a PDF document and linking variables in the text with the main formulas for determining the semantic content of the document. The data set built on the basis of the developed algorithms will be used to improve the quality of semantic search for mathematical formulas.

Semantic annotation of a formula consists in extracting a formula that meets special requirements from the text of a mathematical article, followed by an analysis of its structural elements and linking the selected formula variables with the legends given in the textual context of the formula.

The main task of semantic annotation of formulas is to develop a software solution allowing to identify a set of variables in the formulas of a mathematical document, and associate variables with mathematical concepts using mathematical ontology. The resulting semantic markup of the document will make possible the creation of a collection of documents suitable for the semantic formula search service, which is part of the set of services of the Lobachevskii-DML digital library.

To solve the problem of semantic annotation of formulas in PDF documents, the following tasks were solved:

- Splitting the document into blocks.
- Highlighting the main formulas and text blocks.
- Search for variables in text blocks.
- Recognition of the main formulas and local variables.
- Linking formulas and local variables.
- Markup of mathematical concepts in text blocks based on OntoMathPRO ontology.
- Linking the selected concepts with the variables of the formula.

These tasks were performed using python programming language. The division into blocks was carried out by analyzing the markup of the document. The main formulas are the formulas highlighted in a separate paragraph. Local variables are located in text blocks. Linking local variables and main formulas was performed by searching for common formulas. Fig. 1 shows an example of linking local variable to the main formula.

**§1. Введение**

Линейные уравнения и операторы с частными интегралами возникают в теории эластичности [4], механики сплошных сред [1; 2; 12], аэродинамики [7], в теории частных дифференциальных уравнений [5; 14] и ряде других задач [8; 28; 29]. Самосопряженные частично интегральные операторы возникают также в теории дискретных операторов Шредингера [15; 22; 26; 27]. Как нам известно, исследованием трансфер-матриц гильбертовских случайных полей на целочисленной решетке (решетчатых моделей квантового поля, моделей статистической физики [13; 31; 32]), а также моделей из теории твердого тела (спиновых волн [3; 17]) приводятся к задаче о спектральном анализе так называемого кластерного оператора [9; 11]. В теории кластерных операторов и теории решетчатых гамильтонианов также возникают частично интегральные операторы. В настоящей работе рассматривается самосопряженный частично интегральный оператор  $H$  типа Фредгольма из теории двухчастичных кластерных операторов и двухчастичных решетчатых гамильтонианов (см. [6; 11]).

Пусть  $\Omega_1 = [a, b]^{\nu_1}$  и  $\Omega_2 = [c, d]^{\nu_2}$  ( $\nu_1, \nu_2 \in \mathbb{N}$ ). В гильбертовом пространстве  $L_2(\Omega_1 \times \Omega_2)$  рассмотрим следующий самосопряженный частично интегральный оператор (ЧИО):

$$H = H_0 - (T_1 + T_2). \quad (1)$$

Fig. 1. Linking the main formula and the variable

Recognition of concepts in the text was performed using the OntoMathPRO ontology concept extraction algorithm. This method is used in the preparation of mathematical educational courses at Kazan Federal University. The main idea of this algorithm is to search for all chains of words in a sentence and compare them with similarly constructed chains in concepts of ontology. The algorithm accepts documents in html format as input, therefore, a method for generating an intermediate html representation of a PDF document was created for its application in this task. In the tags of such a representation, a text representation of the source document was obtained, indicating the number of the block. Fig. 2 shows an example of recognizing concepts from the OntoMathPRO anthology in a text block of a document.

Линейные уравнения (линейное уравнение) и операторы (оператор) с частными интегралами (интеграл) возникают в теории эластичности [4], механики сплошных сред [1; 2; 12], аэродинамики [7], в теории частных дифференциальных уравнений (дифференциальное уравнение) [5; 14] и ряде других задач [8; 28; 29]. Самосопряженные (сопряженный оператор) частично интегральные операторы (интегральный оператор) возникают также в теории дискретных операторов Шредингера [15; 22; 26; 27]. Как нам известно, исследованием трансфер-матриц гильбертовских случайных полей (поле случайное) на целочисленной решетке (решетка) (решетчатых моделей квантового поля (поле), моделей статистической (статистическая модель) физики [13; 31; 32]), а также моделей из теории твердого тела (тело) (спиновых волн [3; 17]) приводятся к задаче о спектральном анализе так называемого кластерного оператора [9; 11]. В теории кластерных операторов (оператор) и теории решетчатых гамильтонианов (гамильтониан) также возникают частично интегральные операторы. В настоящей работе рассматривается самосопряженный (сопряженный оператор) частично интегральный оператор (интегральный оператор)  $H$  типа Фредгольма из теории двухчастичных кластерных операторов (оператор) и двухчастичных решетчатых гамильтонианов (гамильтониан) (см. [6; 11]).

$$H = H_0 - (T_1 + T_2). \quad (1)$$

Пусть  $\Omega_1 = [a, b]^{\nu_1}$  и  $\Omega_2 = [c, d]^{\nu_2}$  ( $\nu_1, \nu_2 \in \mathbb{N}$ ). В гильбертовом пространстве (гильбертово пространство)  $L_2(\Omega_1 \times \Omega_2)$  рассмотрим следующий самосопряженный (самосопряженный оператор) частично интегральный оператор (интегральный оператор) (ЧИО):

Fig. 2. Recognized concepts in the text of the document

The developed algorithm for linking local variables and main formulas has the following estimates: precision - 0,81, recall - 0,72, F-measure - 0,75.

Table 1 shows examples of the main formulas and related local variables.

TABLE I. EXAMPLES OF RECOGNIZED FORMULAS AND VARIABLES

Main formula	Variable	Ontology concept
$\int_{\Omega_1} \varphi_j(\xi) d\mu_j(\xi) = 0, \int_{\Omega_1} \varphi_j^2(\xi) d\mu_j(\xi) = 1$	$\mu_j(\cdot)$	Lebesgue measure
$t \mapsto \frac{n!}{\sqrt{\sum_{i=1}^n (x_i - x_0)^2}} \cdot \text{mes}(U(t))$	$\text{mes}(U(t))$	Lebesgue measure
$H = H_0 - (T_1 + T_2)$	$H$	Integral operator

III. CONCLUSION

The article presents a method of semantic annotation of mathematical documents in PDF format. A method for determining the structure of a document by dividing it into blocks with text and main formulas is described. A method of linking local variables in the text with the main forms has been developed. With the help of the method of marking mathematical concepts in the text, a semantic representation of the main formulas is formed. The developed method is used to prepare a data set for a semantic search engine using formulas in the Lobachevskii-DML digital library.

Future developments are related to the expansion of the method's capabilities, in particular, for linking many variables of a mathematical function, as well as the development of a method for filtering concepts found in a sentence to increase the accuracy of the annotation method. Another direction for increasing the universality of the method can be considered planning the implementation of an OCR module for text recognition in scanned PDF, which is especially relevant for mathematical articles published in the pre-digital era.

ACKNOWLEDGMENT

The study was carried out with the support of the Russian Science Foundation, project No. 21-11-00105.

REFERENCES

- [1] Nevzorova O. The semantic context models of mathematical formulas in scientific papers / O. Nevzorova, A. Kirillovich, V. Nevzorov, K. Nikolaev // CEUR Workshop Proceedings. – 2018. – Vol. 2277. – P. 33-40.
- [2] Bertin M. Hybrid Approach for the Semantic Processing of Scientific Papers / M. Bertin, I. Atanassova // Semantic Publishing Challenge Track in 11 th European Semantic Web Conference (ESWC 2014). – 2014.
- [3] Ciancarini P. Semantic annotation of scholarly documents and citations / P. Ciancarini, A. Di Iorio, A. G. Nuzzolese et al. // Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). – 2013. – Vol. 8249 LNAI. – P. 336-347.
- [4] Ronzano F. Semantify CEUR-WS proceedings: Towards the automatic generation of highly descriptive scholarly publishing linked datasets / F. Ronzano, G. C. Del Bosque, H. Saggion // Communications in Computer and Information Science. – 2014. – Vol. 475. – P. 83-88.
- [5] Ahmad R. Information extraction from PDF sources based on rule-based system using integrated formats / R. Ahmad, M. T. Afzal, M. A. Qadir // Communications in Computer and Information Science. – 2016. – Vol. 641. – P. 293-308.
- [6] Greiner-Petter A. Math-word embedding in math search and semantic extraction / A. Greiner-Petter, A. Youssef, T. Ruas [et al.] // Scientometrics. – 2020. – Vol. 125. (3). – P. 3017-3046.

# Способ темпоральной интерполяции толщины подвергающейся коррозии стенки газопровода согласованной с физической моделью

Р.Р. Габбасов

Самарский национальный исследовательский университет  
им. академика С.П. Королева  
Самара, Россия  
rklug@mail.ru

Р.А. Парингер

Самарский национальный исследовательский университет  
им. академика С.П. Королева  
ИСОИ РАН  
Самара, Россия  
rusparinger@gmail.com

**Аннотация**—Анализ развивающихся во времени процессов играет с развитием вычислительных мощностей всё более важную роль в современном мире. В данной работе рассматривается процесс коррозионного износа стенки газопровода, а именно, задача регрессии значения толщины стенки трубы. Предлагается новый способ интерполяции во времени значений толщины стенки, производимой в согласовании с физическими показателями транспортируемого газового конденсата. Проводятся эксперименты по машинному обучению регрессионных моделей с использованием алгоритма RANSAC, вводятся определения двух мер соответствия обученных моделей физической реальности. Результаты экспериментов показали, что использование предлагаемого способа интерполяции вместо интерполяции сплайнами позволяет добиться увеличения значения первой меры в среднем в 2 раза, а значения второй меры – в 3 раза.

**Ключевые слова**— **ВРЕМЕННЫЕ РЯДЫ, КОРРОЗИОННЫЙ ИЗНОС, ЗАДАЧА РЕГРЕССИИ, RANSAC**

## 1. ВВЕДЕНИЕ

Актуальной во многих областях деятельности человека, в частности, в контексте анализа временных рядов [1, 2, 3] является задача предсказания (регрессии) значений. В данной работе речь идет о регрессии уменьшающейся вследствие коррозии толщины труб, привязанных к двум скважинам одного газового месторождения. Обвязки скважин, рассматриваемых в нашей работе, оборудованы датчиками, снимающими показания физических характеристик проходящего конденсата. Данные с этих датчиков используются в качестве признаков для регрессионных моделей. Замеры толщины производятся в разных местах в разное время на обвязке скважины с помощью ультразвуковой диагностики. Мы предлагаем способ интерполяции толщин во времени, учитывающий физические характеристики потока наряду с данными о промежуточных значениях толщин. Для оценки эффективности применения предлагаемого способа наряду с ним мы используем интерполяцию сплайнами.

## 2. ОПИСАНИЕ И ПОДГОТОВКА НАБОРА ДАННЫХ

В данной работе использовались данные с двух скважин (с названиями «2-2» и «3-1») одного месторождения. Для каждой из них имелись значения 17-ти изменяющихся со временем параметров, снятые с шагом в 1 час с датчиков, расположенных на трубной обвязке скважины: значения давлений и температур в разных местах, содержания в конденсате  $\text{CO}_2$  и его pH. Данные параметры выступали в качестве входных признаков в процессе обучения. Процесс сглаживания

выбросов и заполнения пропусков в этих данных был осуществлён с использованием скользящего окна [4]. Также для каждой скважины имелись значения целевого параметра – сильно разнесенные по времени результаты замеров толщины стенки компонентов обвязки в разных местах обвязки в разное время. В данной работе было рассмотрено два способа интерполяции этого признака (с целью соответствия частоте дискретизации признаков): с использованием квадратичных сплайнов [5] и **предлагаемый нами способ интерполяции**, производимой в согласовании с **физической моделью**, т.е. с **физическими показателями** транспортируемого газового конденсата. Для обоих способов интерполяции соответственно были подготовлены два набора данных для дальнейших экспериментов.

### А. Интерполяция квадратичными сплайнами

Характеристика интенсивности коррозионного процесса связана со **скоростью изменения толщины стенки**, поэтому модели обучались регрессировать значения этой скорости. Так как скорость изменения параметра является производной функции параметра, использование линейной интерполяции привело бы к вырождению скорости изменения толщины между двумя известными исходными временными метками в константу, поэтому используется квадратичная интерполяция.

### Б. Интерполяция, согласованная с физическими показателями

Данный способ интерполяции, который мы предлагаем, основан на использовании расчетов по стандарту NORSOK M-506 [6]. Данный стандарт позволяет рассчитать теоретическую скорость коррозии трубопровода (в мм/год) в фиксированный момент времени на основе значений давления, температуры и pH потока, содержания в нем  $\text{CO}_2$ , а также значений диаметра, шероховатости и напряжения сдвига стенки трубы. Имея эти данные, мы по данной методике получили значения теоретической скорости утонения стенки, а затем использовали их в качестве весов интенсивности истончения между двумя исходными замерами.

## 3. УСЛОВИЯ ЭКСПЕРИМЕНТОВ

Так как рассматриваемые данные являют собой временной ряд, то для каждого целевого значения рассматривалось временное окно признаков размера 3600 (3600 часов = 150 суток = 5 месяцев). В качестве регрессионной модели в данной работе используется модель линейной регрессии с дополнительным использованием алгоритма RANSAC [7] со следующими гиперпараметрами: минимальная доля выбираемых случайных элементов – 0,1, функция ошибок –

квадратичная. Также рассматривались модели регрессора с функцией потерь Хьюбера [8] и регрессора Тейла-Сена [9], однако было установлено, что их обучение занимает слишком много времени относительно времени обучения регрессора с использованием RANSAC при сравнимых с точки зрения точности результатах. Для каждой точки на трубе была обучена собственная регрессионная модель.

#### 4. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

В данном разделе представлено сравнение результатов обучения регрессионных моделей для двух рассматриваемых в работе скважин («2-2» и «3-1») в случае использования двух представленных способов интерполяции значения толщины стенки.

##### А. Схожесть моделей

Мера схожести между двумя моделями высчитывалась следующим образом. С помощью обеих моделей на временном участке между двумя соседними исходными моментами замера толщины (входящих в тренировочную выборку) по 17-ти признакам были регрессированы валидационные значения скорости изменения толщины стенки. Мера схожести моделей определялась как максимум между 0 и значением метрики  $R^2$  между двумя получившимися векторами значений. Далее модели со схожестью более 0,9 между собой объединялись в отдельные группы. Первая мера соответствия обученных моделей физической реальности ( $M_1$ ) определялась как средний размер таких групп моделей (так как физическая связность обвязки ведет к схожести в поведении моделей для разных точек на ней).

Таблица I. РЕЗУЛЬТИРУЮЩИЕ ЗНАЧЕНИЯ МЕРЫ  $M_1$

Скважины	2-2	3-1
Интерполяция квадратичными сплайнами	6,87	4,32
Интерполяция, согласованная с физическими параметрами	10,23	10,57
Соотношение	1,49	2,45

##### Б. Относительная важность признаков

Была оценена относительная значимость признаков для регрессии утонения стенок в точках на различных участках обвязки: около устья, до штуцера, после штуцера. Исходные признаки были сформированы в шесть наборов в зависимости от местоположения соответствующих датчиков на обвязке скважины. Затем для каждой точки замера на обвязке был обучен набор из шести регрессионных моделей, обученных на соответствующих наборах признаков. Далее модели были поделены на три класса: около устья, до штуцера, после штуцера – в зависимости от местоположения соответствующей точки замера на обвязке. Затем для каждого класса для каждого набора признаков было подсчитано среднее по соответствующим моделям значение метрики  $L_1$  между регрессированными моделями и исходными значениями целевого параметра на тестовой выборке. Полученные значения нормализовывались и инвертировались таким образом, чтобы наименьшее значение  $L_1$  соответствовало 1, а большие значения  $L_1$  – числам меньше 1. Полученные числа и являются значимостью признаков. Вторая мера

соответствия обученных моделей физической реальности ( $M_2$ ) определялась на основе полученных значений значимости признаков следующим образом:

$$M_2 = (L_1(\mathbf{i}, \mathbf{i}'))^{-1},$$

где  $L_1(*,*)$  – метрика  $L_1$ ,  $\mathbf{i} \in \mathbb{R}^{18}$  – вектор значимостей признаков для различных классов (6 наборов признаков  $\times$  3 класса моделей),  $\mathbf{i}' \in \mathbb{R}^{18}$  – вектор, состоящий из одного и того же значения, являющегося медианой вектора  $\mathbf{i}$ .

Таблица II. РЕЗУЛЬТИРУЮЩИЕ ЗНАЧЕНИЯ МЕРЫ  $M_2$

Скважины	2-2	3-1
Интерполяция квадратичными сплайнами	0,31	0,24
Интерполяция, согласованная с физическими параметрами	0,77	0,83
Соотношение	2,48	3,46

#### 5. ЗАКЛЮЧЕНИЕ

В данной статье был предложен способ интерполяции во времени значений толщины стенки компонентов обвязки газовых скважин, производимой в согласовании с физическими параметрами проходящего конденсата. Были проведены эксперименты по обучению моделей линейной регрессии с использованием алгоритма RANSAC. Были введены определения двух мер соответствия моделей физической реальности. Результаты показывают, что использование предлагаемого способа интерполяции вместо интерполяции квадратичными сплайнами приводит к повышению меры соответствия моделей физической реальности: значение меры  $M_1$  повысилось в 1,49 и 2,45 раза для скважин 2-2 и 3-1 соответственно, а для  $M_2$  соответствующие повышения составили 2,48 и 3,46 раз. Стоит отметить, что согласованность результатов для обеих предложенных мер говорит о высоком уровне их достоверности.

#### ЛИТЕРАТУРА

- [1] Boori, M.S. Crop growth monitoring through Sentinel and Landsat data based NDVI time-series / M. S. Boori, K. Choudhary, A. V. Kupriyanov // Computer Optics. – 2020. – Vol. 44(3). – P. 409-419.
- [2] Терехин Э. А. Индикация многолетних изменений в растительном покрове залежных земель лесостепи на основе рядов вегетационного индекса NDVI / Э. А. Терехин // Компьютерная оптика. – 2021. – Т. 45, №. 2. – С. 245-252.
- [3] Plotnikov, D. Daily surface reflectance reconstruction using LOWESS on the example of various satellite systems / D. Plotnikov [et al.] // 2022 VIII International Conference on Information Technology and Nanotechnology (ITNT). – IEEE. – 2022. – P. 1-5.
- [4] Yu, Y. Time series outlier detection based on sliding window prediction / Y. Yu [et al.] // Mathematical problems in Engineering. – 2014. – Vol. 2014.
- [5] Sharma, A. Quadratic splines / A. Sharma, J. Tzimbalaro // Journal of Approximation Theory. – 1977. – Vol. 19(2). – P. 186-193.
- [6] NORSOK M-506. CO2 corrosion rate calculation model.
- [7] Derpanis, K.G. Overview of the RANSAC Algorithm / K.G. Derpanis // Image Rochester NY. – 2010. – Vol. 4(1). – P. 2-3.
- [8] Huber, P.J. Robust regression: asymptotics, conjectures and Monte Carlo / P.J. Huber // The annals of statistics. – 1973. – P. 799-821.
- [9] Wilcox, R. A note on the Theil-Sen regression estimator when the regressor is random and the error term is heteroscedastic / R. Wilcox // Biometrical Journal: Journal of Mathematical Methods in Biosciences. – 1998. – Vol. 40(3). – P. 261-268.



Министерство образования  
и науки Самарской области



**САМАРСКИЙ УНИВЕРСИТЕТ**  
SAMARA UNIVERSITY  
Самарский университет



Институт систем обработки изображений РАН



**J-VPE**

**OPTICAL MEMORY  
AND  
NEURAL NETWORKS  
(Information Optics)**

ISSN 2070-7401 (Print), ISSN 2411-0280 (Online)

Институт космических исследований  
Российской академии наук

**СОВРЕМЕННЫЕ ПРОБЛЕМЫ ДИСТАНЦИОННОГО  
ЗОНДИРОВАНИЯ ЗЕМЛИ ИЗ КОСМОСА**

**ИКИ**

физические основы, методы и технологии мониторинга  
окружающей среды, потенциально опасных явлений  
и объектов

**ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ  
(INFORMATION PROCESSES)**  
Электронный научный журнал

