ISSN 1816-0301 (Print)
ISSN 2617-6963 (Online)

ИНФОРМАТИКА

INFORMATICS

ТОМ 21
VOL.

2 | 2024

20 лет
ЮБИЛЕЙ ЖУРНАЛА

ОТ РЕДАКЦИИ

В журнале «Информатика» публикуются оригинальные и обзорные статьи, описывающие результаты фундаментальных и прикладных исследований специалистов академического и вузовского профиля в области информатики и информационных технологий.

Основной целью журнала является публикация наиболее значимых новых результатов в указанной области. Приветствуются статьи, описывающие заключительные результаты научных проектов и диссертационных исследований, открывающие новые направления исследований, которые находятся на стыке информатики и других наук.

Журнал рассчитан на широкий круг специалистов в области информатики и информационных технологий.

Основные разделы журнала:

- биоинформатика;
- математическое моделирование;
- защита информации и надежность систем;
- информационные технологии;
- логическое проектирование;
- обработка сигналов, изображений, речи, текста и распознавание образов;
- автоматизация проектирования;
- интеллектуальные системы.

Префикс DOI: 10.37661

Индексирование:

Высшей аттестационной комиссией Республики Беларусь журнал «Информатика» был включен в список научных изданий для опубликования результатов диссертационных исследований.

В декабре 2017 г. включен в базу данных Российского индекса научного цитирования (РИНЦ). С помощью инструментов и сервисов, доступных на платформе eLIBRARY (раздел «Личный кабинет»), можно самостоятельно корректировать список своих публикаций и цитирований в РИНЦ.

В июле 2017 г. включен в базу журналов открытого доступа Directory of Open Access Journals (DOAJ).

С помощью поисковых систем Google Scholar, WorldCat, Соционет можно получить свободный доступ к полному тексту научных публикаций журнала.

Адрес редакции:
ул. Сурганова, 6, к. 305, г. Минск, 220012, Беларусь
Тел. +375 (017) 351 26 22

Editorial address:
Surganova str., 6, of. 305, Minsk, 220012, Belarus
Phone +375 (017) 351 26 22

E-mail: rio@newman.bas-net.by
https://inf.grid.by/jour

THE EDITOR'S NOTE

The journal "Informatics" is a scientific publication in computer sciences and information technologies which reviews the results in basic and applied research of scientists from the universities and scientific centers.

The journal focuses on the most significant and modern papers of research projects results and PhD/DSc thesis in computer sciences.

The journal is edited for the specialists in IT and computer sciences research and application.

The main sections of the journal:

- bioinformatics;
- mathematical modeling;
- information protection and system reliability;
- information technology;
- logical design;
- signal, image, speech, text processing and pattern recognition;
- computer-aided design;
- artificial intelligence methods.

DOI Prefix: 10.37661

Indexation:

the journal "Informatics" is in the list of scientific publications recommended by the Higher Attestation Commission of the Republic of Belarus for scientists to publish the results of PhD/DSc research.

In December 2017 the journal was included in the database of the Russian Science Citation Index (RISC) and provides free access to reviewed electronic scientific paper, improving scientific information traffic and also raising quotation of works of the authors (please use https://elibrary.ru or section for authors https://elibrary.ru_author_tools).

In July 2017 included in the database of open access journals Directory of Open Access Journals (DOAJ).

Using the Google Scholar, WorldCat, Соционет search engine, you can get free access to full text of scientific publications of magazine.

ISSN 1816-0301 (Print)
ISSN 2617-6963 (Online)

ОБЪЕДИНЕННЫЙ ИНСТИТУТ ПРОБЛЕМ ИНФОРМАТИКИ
НАЦИОНАЛЬНОЙ АКАДЕМИИ НАУК БЕЛАРУСИ

ИНФОРМАТИКА

Informatika

Том 21, № 2, апрель-июнь 2024

Ежеквартальный научный журнал

Издается с января 2004 г.

Учредитель и издатель – государственное научное учреждение «Объединенный институт проблем информатики Национальной академии наук Беларуси» (ОИПИ НАН Беларуси)

ИНФОРМАТИКА

Том 21, № 2, апрель-июнь 2024

Ответственный за выпуск *Мойсейчик Светлана Сергеевна*
Редактор *Гончаренко Галина Борисовна*
Компьютерная верстка *Бутевич Ольга Борисовна*

Сдано в набор 20.05.2024. Подписано в печать 18.06.2024. Формат 60×84 1/8. Бумага офсетная. Гарнитура Таймс. Ризография. Усл. печ. л. 12,3. Уч.-изд. л. 12,0. Тираж 40 экз. Заказ 3.

Государственное научное учреждение «Объединенный институт проблем информатики Национальной академии наук Беларуси».
Свидетельство о государственной регистрации издателя, изготовителя, распространителя печатных изданий № 1/274 от 04.04.2014. ЛП № 02330/444 от 18.12.13. Ул. Сурганова, 6, 220012, Минск, Беларусь.

ISSN 1816-0301 (Print)
ISSN 2617-6963 (Online)

THE UNITED INSTITUTE OF INFORMATICS PROBLEMS
OF THE NATIONAL ACADEMY OF SCIENCES OF BELARUS

INFORMATICS

Vol. 21, no. 2, April-June 2024

Published quarterly

Issued since January 2004

Founder and publisher – State Scientific Institution "The United Institute of Informatics Problems of the National Academy of Sciences of Belarus" (UIIP NASB)

INFORMATICS

Vol. 21, no. 2, April-June 2024

Issue Head *Sviatlana S. Maiseichyk*
Editor *Halina B. Hancharenka*
Computer Imposition *Volha B. Butsevich*

Sent for press 20.05.2024. Output 18.06.2024. Format 60×84 1/8. Offset paper. Headset Times. Riesography. Printed sheets 12,3. Publisher's signatures 12,0. Circulation 40 copies. Order 3.

State Scientific Institution "The United Institute of Informatics Problems of the National Academy of Sciences of Belarus".
Certificate on the state registration of the publisher, manufacturer, distributor of printing editions no. 1/274 dated 04.04.2014. License for the press no. 02330/444 dated 18.12.13.
6, Surganov Str., 220012, Minsk, Belarus.

ISSN 1816-0301 (Print)
ISSN 2617-6963 (Online)

СОДЕРЖАНИЕ

ISSN 1816-0301 (Print)
ISSN 2617-6963 (Online)

CONTENTS

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ
INFORMATION TECHNOLOGIES

УДК 004.65;004.75;004.5;004.91
https://doi.org/10.37661/1816-0301-2024-21-2-7-23

Оригинальная статья
Original Article

Основы функционирования семантического портала ядерных знаний BelNET

С. Н. Сытова✉, В. В. Гавриловец, А. П. Дунец, А. Н. Коваленко, С. В. Черепица

Институт ядерных проблем
Белорусского государственного университета,
ул. Бобруйская, 11, Минск, 220006, Беларусь
✉*E-mail: sytova@inp.bsu.by*

Аннотация

Цели. Рассмотрена возможность использования семантических технологий для развития и совершенствования системы управления контентом научно-образовательного портала eLab-Science и созданного на ее основе белорусского портала ядерных знаний BelNET (Belarusian Nuclear Education and Training Portal, https://belnet.by/).

Методы. Разработаны оригинальные алгоритмы автоматической систематизации – размещения записей контента в таксономии портала на основе семантических технологий и формирования списка ключевых слов. Использованы такие понятия семантических технологий, как таксономия (иерархическая структура портала), тезаурус, глоссарий.

Результаты. Разработанные алгоритмы реализованы и протестированы с использованием инструмента полнотекстового поиска и оригинального белорусского глоссария по ядерной и радиационной безопасности.

Заключение. Описанные принципы организации и алгоритмы на базе семантических технологий, лежащие в основе функционирования системы управления контентом научно-образовательного портала eLab-Science и созданного на ее базе белорусского портала ядерных знаний BelNET, позволяют эффективно реализовывать размещение записей контента в таксономии портала, а также автоматически формировать набор ключевых слов создаваемого ресурса.

Ключевые слова: ядерные знания, управление ядерными знаниями, информационная система, свободное программное обеспечение, тезаурус, глоссарий, таксономия

Благодарности. Работа выполняется в рамках мероприятия 13 «Выполнение работ по оказанию научно-технической поддержки Министерству по чрезвычайным ситуациям Республики Беларусь в области обеспечения ядерной и радиационной безопасности» подпрограммы 3 «Научное обеспечение эффективной и безопасной работы Белорусской атомной электростанции и перспективных направлений развития атомной энергетики» Государственной программы «Наукоемкие технологии и техника» на 2021–2025 гг.

Для цитирования. Основы функционирования семантического портала ядерных знаний BelNET / С. Н. Сытова [и др.] // Информатика. – 2024. − Т. 21, № 2. – С. 7–23.
https://doi.org/10.37661/1816-0301-2024-21-2-7-23

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию | Received 05.12.2023
Подписана в печать | Accepted 21.03.2024
Опубликована | Published 28.06.2024

Basics of the semantic portal of nuclear knowledge BelNET functioning

Svetlana N. Sytova✉, Viktar V. Haurylavets, Andrei P. Dunets, Anton N. Kavalenka, Siarhei V. Charapitsa

Institute for Nuclear Problems
of Belarusian State University,
st. Bobruiskaya, 11, Minsk, 220006, Belarus
✉*E-mail: sytova@inp.bsu.by*

Abstract

Objectives. The possibility of using semantic technologies for the development and improvement of the content management system of the scientific and educational portal eLab-Science and the Belarusian nuclear knowledge portal BelNET (Belarusian Nuclear Education and Training Portal, https://belnet.by/) created on its basis is being considered.

Methods. Original algorithms for automatic systematization have been developed, such as placing content records in the portal taxonomy based on semantic technologies and generating a list of keywords. The following concepts of semantic technologies are used: taxonomy (hierarchical structure of the portal), thesaurus, glossary.

Results. The developed algorithms were implemented and tested using a full-text search tool and the original Belarusian glossary on nuclear and radiation safety.

Conclusion. The described basic principles of organization and algorithms based on semantic technologies, which underlie the functioning of the content management system of the scientific and educational portal eLab-Science and the Belarusian nuclear knowledge portal BelNET, created on its basis, make it possible to effectively implement the placement of content records in the portal taxonomy, as well as automatically generate a set of keywords for the resource being created.

Keywords: nuclear knowledge, nuclear knowledge management, information system, free software, thesaurus, glossary, taxonomy

Acknowledgements. The work is carried out within the framework of the activity 13 "Performing work to provide scientific and technical support to the Ministry of Emergency Situations of the Republic of Belarus in the field of ensuring nuclear and radiation safety" of Subprogram 3 "Scientific support for the effective and safe operation of the Belarusian nuclear power plant and promising directions for the development of nuclear energy" of the State Program "High-tech technologies and equipment" for 2021–2025.

For citation. Sytova S. N., Haurylavets V. V., Dunets A. P., Kavalenka A. N., Charapitsa S. V. *Basics of the semantic portal of nuclear knowledge BelNET functioning*. Informatika [*Informatics*], 2024, vol. 21, no. 2, pp. 7−23 (In Russ.). https://doi.org/10.37661/1816-0301-2024-21-2-7-23

Conflict of interest. The authors declare of no conflict of interest.

Введение. Ядерные знания [1] связаны как с исследованиями и разработками, так и с промышленным использованием ядерных технологий и включают широкий спектр энергетических и неэнергетических применений. Международное агентство по атомной энергии (МАГАТЭ) с начала 2000-х гг. активно разрабатывает методологию и руководящие документы для планирования, разработки и реализации программ управления ядерными знаниями [2].

Согласно терминологии МАГАТЭ [3] управление ядерными знаниями: получение, сбор, передача, сохранение, поддержание и использование знаний, а также обмен ими – имеет важное значение для развития и поддержания технических знаний и компетенций, необходимых для ядерно-энергетических программ и различных ядерных технологий. Деятельность МАГАТЭ в этой области содействует ядерному образованию, предоставляя поддержку, возможности для налаживания связей и обмена опытом [4]. В последние годы одним из наиболее действенных инструментов в менеджменте ядерных знаний являются порталы ядерных знаний, создаваемые и развиваемые при поддержке МАГАТЭ, а также разнообразные корпоративные системы современных ядерных знаний [5].

В соответствии с материалами МАГАТЭ [6] семантическая технология, лежащая в основе веб-поиска и управления онлайн-информацией, в обязательном порядке должна использоваться в ядерной области для помощи экспертам и всем заинтересованным сторонам в поддержании, сохранении и обмене ядерными знаниями. Применение семантических технологий помогает в интеграции различных источников данных, автоматизации индексации ресурсов, облегчает поиск информации эффективным и экономичным способом, повышает устойчивость управления сложными и междисциплинарными системами ядерной энергетики. В настоящее время МАГАТЭ развивает различные инициативы в области семантических технологий, которые могут принести пользу в области менеджмента ядерных знаний. Это касается как проектов МАГАТЭ, так и национальных программ [6].

В качестве яркого примера можно привести созданную в 1970 г. под эгидой МАГАТЭ Международную ядерную информационную систему INIS (The International Nuclear Information System, https://www.iaea.org/resources/databases/inis), используемую более чем в 130 странах мира. Репозиторий INIS содержит библиографические ссылки, научные и технические отчеты, материалы конференций, патенты и тезисы во всех областях деятельности МАГАТЭ, включая ядерную технику и технологии, ядерную безопасность и радиационную защиту, гарантии и нераспространение, применение ядерных и изотопных методов, ядерную физику и физику высоких энергий, ядерную и радиационную химию, ядерные применения в науках о жизни, правовые аспекты, экологические и экономические аспекты ядерных и неядерных источников энергии. При предметной классификации (категоризации) каждая запись в базах данных INIS отнесена к определенной предметной категории, а также к одной или нескольким вторичным тематическим категориям.

В Беларуси в настоящее время формируется полноценная система управления ядерными знаниями, основу которой составляет портал ядерных знаний BelNET [7–10]. Его цели полностью соответствуют подходам МАГАТЭ к менеджменту ядерных знаний.

Статья посвящена становлению активно развивающегося в настоящее время портала ядерных знаний BelNET как семантического портала. Необходимость использования семантических технологий в оригинальной системе управления контентом портала BelNET на основе свободного программного обеспечения вызвана тем, что ранее разработчики столкнулись с отрицательным опытом при ручной систематизации (определении разделов и подразделов портала, в которых должна быть размещена запись) создаваемых на портале BelNET новых записей. В результате этого многие разделы таксономии (иерархической структуры портала), которая оказалась очень большой и сложной, до сих пор остаются пустыми либо слабозаполненными. Однако очевидно, что многие ресурсы могли бы быть размещены в нескольких разделах и стать более доступными для читателей.

Теоретические основы семантических технологий. Семантическая технология является одним из бурно развивающихся алгоритмических направлений современной прикладной математики и информационных технологий. Она включает в себя широкий спектр инструментов, стандартов и методологий, позволяющих обрабатывать информацию в зависимости от ее контекста и значения. Назовем основные понятия семантических технологий, используемые в работе: онтология, глоссарий, тезаурус, таксономия.

Онтология используется для подробной формализации области знаний с помощью концептуальной схемы, которая состоит из структуры данных, содержащей все релевантные классы объектов, их связей и правил (теоремы, ограничения), принятых в этой области. Основные

сферы применения онтологий – моделирование бизнес-процессов, семантическая паутина (https://www.w3.org/standards/semanticweb/) и искусственный интеллект. Данные и онтология с правилами вывода вместе представляют собой базу знаний [11] предметной области.

Глоссарий – это словарь узкоспециализированных терминов в отрасли знаний с толкованием, иногда переводом на другой язык, комментариями и примерами. Он не использует дополнительные связи между терминами и может рассматриваться как онтология с пустым множеством отношений.

Тезаурус – это словарь с дополнительными отношениями, охватывающий понятия, определения и термины области знаний или сферы деятельности, которые подчиняются семантическим отношениям между терминами. Обычные простейшие таксономические отношения в тезаурусах составляют несколько уровней отношений типа выше-ниже [12]. Специализированный тезаурус может разрабатываться экспертами или строиться с помощью программных средств [13]. Для создания тезаурусов существуют специальные государственные и международные стандарты[1, 2]. На основе тезауруса может быть создана таксономия (иерархическая структура) портала. Отметим правило [14], что если в тезаурусе нет подходящего дескриптора для поиска полезного понятия, то следует предложить и ввести в тезаурус новый.

В качестве примера тезауруса приведем разработанный МАГАТЭ многоязычный тезаурус для системы INIS (https://inis.iaea.org/search/thesaurus.aspx) на арабском, китайском, английском, французском, немецком, японском, русском и испанском языках, предоставляющий переводы тысяч технических терминов, которые помогают в навигации и поиске по коллекции INIS. Тезаурус дает возможность пользователям БД INIS индексировать и искать литературу на нескольких языках. Объем английской версии тезауруса [15] в настоящий момент составляет 31 301 термин.

Для онлайн-поиска в поисковых запросах могут использоваться контролируемые термины (дескрипторы) – ключевые слова, произвольные текстовые слова или их комбинация, предназначенные для предметного индексирования с контролируемой терминологией по заголовку ресурса и свободному тексту (например, аннотации, реферату, полному тексту ресурса). Для единообразия такие дескрипторы должны входить в тезаурус или глоссарии. При выборе релевантных ссылок из результатов поиска очень полезными элементами являются реферат, заголовок и дескрипторы исследуемого ресурса.

Таксономия – это иерархическая структура портала, которая может быть построена на основании семантических технологий, в частности на основании одного или нескольких тезаурусов [14].

Оригинальный тезаурус портала BelNET. При работе над таксономией портала BelNET было принято решение придерживаться комбинированного подхода с использованием наработок МАГАТЭ и большого багажа знаний белорусских экспертов.

Для внедрения семантических технологий на портале ядерных знаний BelNET разработан тезаурус (дерево категорий). Его верхний уровень изображен на рис. 1. Здесь находятся разделы «Глоссарий по ядерной и радиационной безопасности», «Научные глоссарии», «Организации», «География», «Персоналии», «Календарь и события». Объем этих глоссариев не должен быть большим. При необходимости внесения нового термина глоссарии всегда могут быть дополнены. Оптимальный объем тезауруса в настоящий момент составляет примерно две-три тысячи терминов.

Глоссарий по ядерной и радиационной безопасности включает 525 терминов. Он специально разработан для портала BelNET на основе нескольких глоссариев МАГАТЭ, Госкорпорации «Росатом», НАТО, а также белорусских нормативно-правовых документов в области ядерной и радиационной безопасности. Глоссарий предназначен для обеспечения алгоритмов функцио-

[1]Тезаурус информационно-поисковый одноязычный. Правила разработки, структура, состав и форма представления : ГОСТ 7.25-2001. – Введ. 01.07.02. – М. : Изд-во стандартов, 2001. − 14 с.

[2]Информация и документация. Тезаурусы и взаимосвязь с другими словарями. Ч. 1. Тезаурусы для выдачи информации : ИСО 25964-1:2011. – Введ. 08.08.11. – М. : Изд-во стандартов, 2011. − 160 с.

нирования портала ядерных знаний, а также облегчения понимания и использования основных терминов в области ядерной и радиационной безопасности с учетом белорусской специфики.

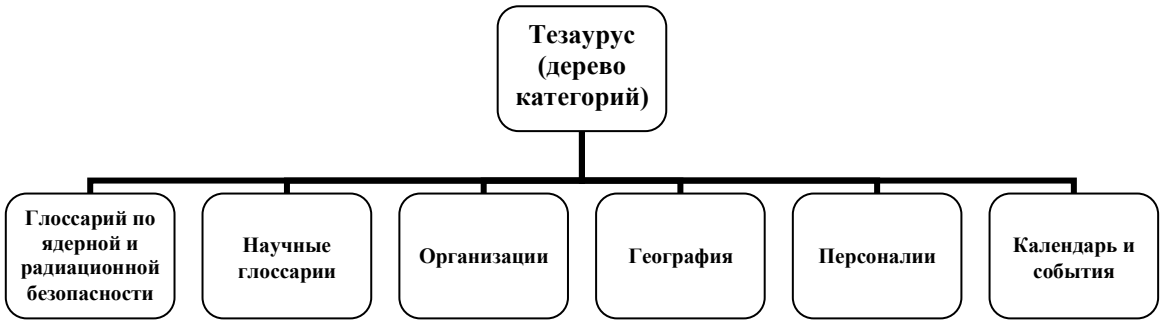


Рис. 1. Верхний уровень тезауруса

Fig. 1. Top level of the thesaurus

Глоссарий включает основные термины (русское и английское название) в области ядерной и радиационной безопасности с учетом белорусской специфики, а также менеджмента ядерных знаний, некоторых основ ядерной физики, физических единиц, основ администрирования и регулирования, учета ядерных материалов, работы с радиоактивными отходами и др. Белорусская специфика отражена через предложение значения терминов в национальных нормативно-правовых актах, а также расшифровку некоторых ключевых понятий и описание некоторых белорусских организаций, в том числе обладающих ядерными установками. Смысл предлагаемых терминов дается предельно кратко, только через главное определение. Дальнейшие подробности могут быть найдены по ссылкам на соответствующие термины.

Отличие глоссария по ядерной и радиационной безопасности от глоссария [16] заключается в преимущественном использовании терминологии белорусских национальных нормативно-правовых актов и выбранном одном значении каждого термина вне зависимости от языка.

В состав научных глоссариев BelNET входят глоссарии по физике, химии, информационным технологиям, техническим терминам, биологии, медицине, науках о Земле, общественным наукам (рис. 2). Изначально предполагалось, что объем создаваемых глоссариев не должен превышать 200 терминов, пригодных для использования в качестве ключевых слов.

В настоящее время специально для портала BelNET разработаны следующие глоссарии: «Единицы физических величин» (55 терминов), «Электричество и магнетизм» (129 терминов), «Квантовая физика» (85 терминов), «Атомная и ядерная физика» (200 терминов), «Астрофизика» (65 терминов), «Радиационное материаловедение, нанотехнологии» (124 термина), «Информационные технологии» (90 терминов), «Технические термины» (150 терминов). Некоторые термины встречаются в нескольких глоссариях, обеспечивая дополнительные горизонтальные отношения в тезаурусе.

Разделы «География», «Организации», «Персоналии», «Календарь и события» (рис. 1) помимо глоссариев включают справочники основных понятий, которые в составе соответствующих глоссариев играют роль предметных категорий.

В разделе «Организации» помимо всех необходимых справочников представлены следующие глоссарии: «Международные и межправительственные организации» (33 термина), белорусские – «Министерства и ведомства» (65 терминов), «Предприятия и организации» (26 терминов), «Научные организации» (30 терминов), «ВУЗы» (44 термина).

Раздел «География» включает подразделы «Регионы мира» и «Регионы Беларуси» (в том числе все административные районы по областям), а также важные географические объекты. Глоссарий «Страны мира» содержит записи о 176 странах мира – членах МАГАТЭ и КНДР, которая прекратила свое членство в МАГАТЭ в 1994 г.

Раздел «Персоналии» подразделяется на глоссарии «Зарубежные персоны» и «Белорусы». В эти глоссарии вошли данные о нобелевских лауреатах в области физики, химии и др. – специалистах в области ядерных знаний, ведущих ученых, руководителях ведомств, организаций и предприятий различного уровня.

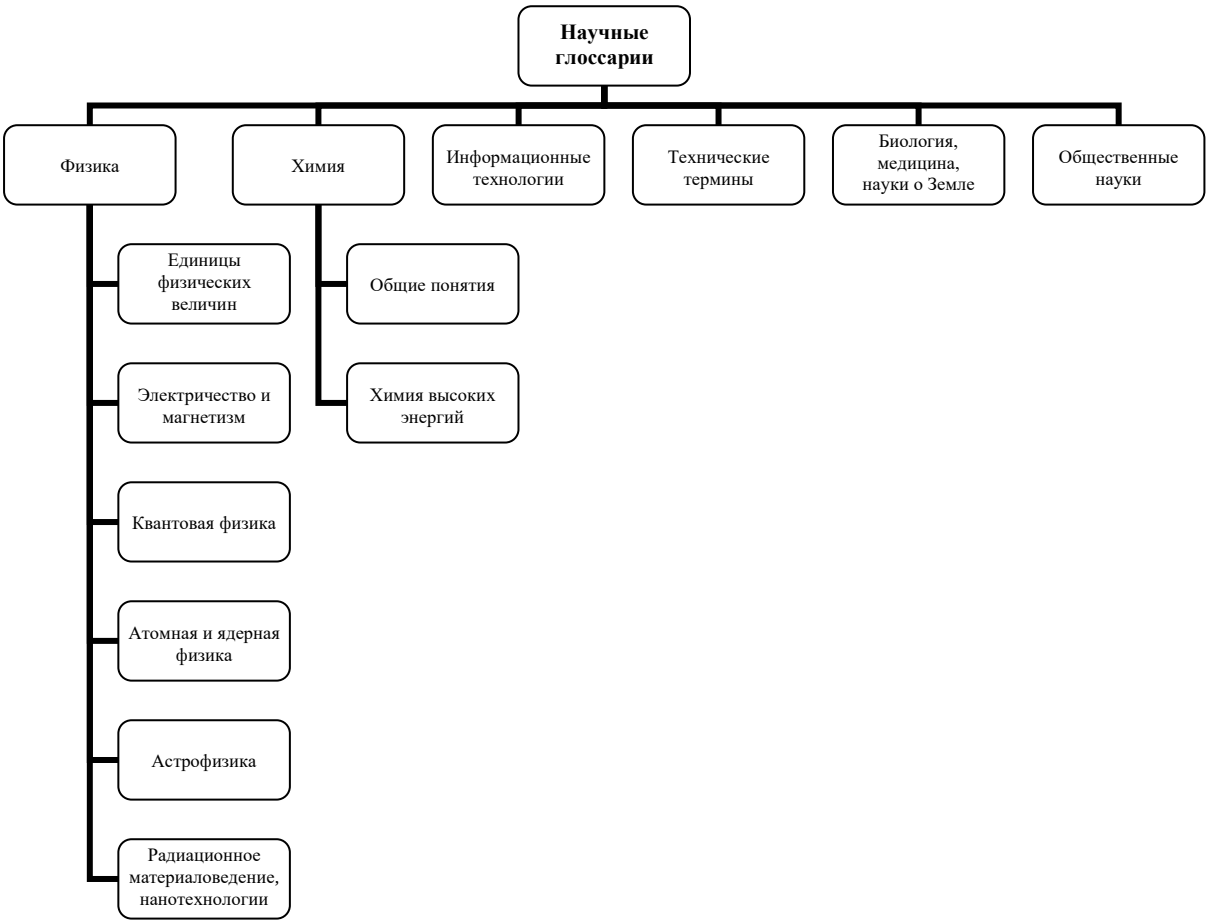


Рис. 2. Структура раздела «Научные глоссарии»
Fig. 2. Structure of section "Scientific glossaries"

Понятно, что разработка такого количества глоссариев является сложной задачей, и эта работа продолжается.

Методы. Алгоритм автоматического размещения ресурса в системе управления контентом eLab-Science. Главная идея алгоритма автоматического размещения ресурса в системе управления контентом eLab-Science заключается в использовании разработанного в системе управления контентом научно-образовательного портала eLab-Science [17, 18], на основе которой создан портал BelNET, полнотекстового поиска в отношении любых вновь создаваемых ресурсов, а также ресурсов, размещенных на портале ранее до внедрения семантических технологий. Это касается и записей в глоссариях.

В eLab-Science в кабинете создателя ресурса специальные кнопки «Индексировать» и «Систематизировать» позволяют пользователю на основе полнотекстового поиска по терминам всех глоссариев тезауруса получать автоматически предлагаемый системой список разделов портала, куда система рекомендует поместить материал, а также список ключевых слов.

На рис. 3 показана диаграмма декомпозиции eLab-Science в обозначениях IDEF1 [19] следующих компонентов системы и их связей: A1 – обслуживание пользователя, A2 – создание ресурса, A3 – автоматическая проверка ресурса, A4 – отображение ресурса на портале, A5 – систематизация ресурса (индексирование, систематизация ресурса, определение уровня доступа к ресурсу).

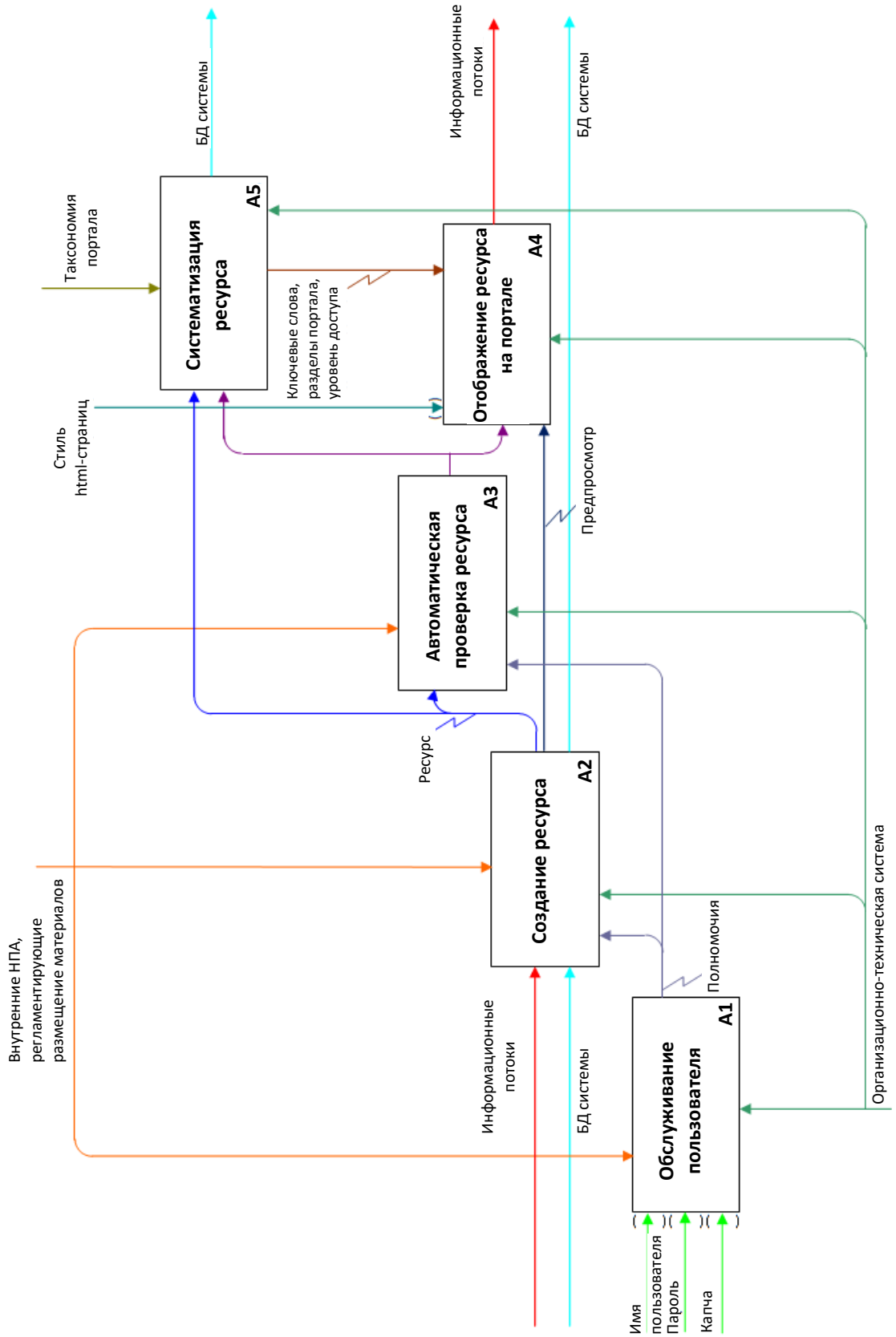


Рис. 3. Схема функциональной структуры портала BelNET в обозначениях IDEF1

Fig. 3. Diagram of the functional structure of portal BelNET in IDEF1 notation

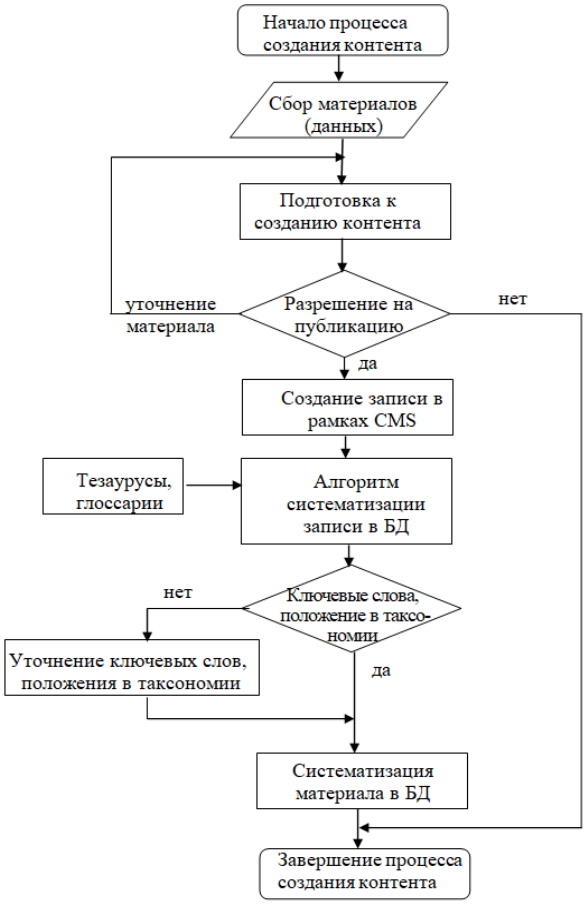


Рис. 4. Алгоритм размещения информационного ресурса на портале ядерных знаний
Fig. 4. Algorithm for posting an information resource on the nuclear knowledge portal

На рис. 4 изображен алгоритм в виде блок-схемы. Создатель новой записи (ресурса, материала, элемента глоссария) начинает ее размещение в своем кабинете, далее система производит автоматический анализ текста (полнотекстовый поиск по всем дескрипторам тезауруса) и пользователю предлагается утвердить набор ключевых слов и положение создаваемой записи в структуре портала (один или несколько разделов). Автор утверждает эти данные либо дополнительно предлагает свои варианты, после чего происходит окончательная систематизация ресурса на портале.

Результаты и обсуждение. Рассмотрим реализацию предложенного семантического алгоритма. Как было отмечено выше, глоссарий – это перечень терминов предметной области (ПрО) с их определениями. В состав глоссария, как правило, включаются термины, которые часто используются в узкой предметной области. Более широкий и системный перечень терминов и понятий ПрО называется тезаурусом. Глоссарий, как правило, составляется на основе ограниченного набора текстов ПрО и предназначен для решения некоторой частной задачи. Составление тезауруса – серьезная системная работа, требующая значительных ресурсов и привлечения экспертов ПрО.

В списке проблем, которые приходится решать при составлении тезауруса, можно указать следующие:

1. Полнота покрытия понятийного поля ПрО – необходимо охватить всю область знаний, не упустив ни одной части.

2. Верификация определений – сверка и согласование формулировок определений экспертами.

3. Структурирование терминов – организация понятийного поля в иерархическую, древовидную или другую систему взаимосвязей с учетом семантики ПрО.

4. Установка границ понятийного поля ПрО.

Отметим, что нередко разработчики тезауруса начинают ощущать себя энциклопедистами в стиле Дидро и д'Аламбера или пытаются превзойти Википедию. Необходимо учитывать, что полезный тезаурус – это терминология конкретной ПрО, собираемая и систематизируемая для решения конкретных практических задач. Попытка «объять необъятное» может ухудшить практическую полезность результатов работы и привести к неоправданным трудозатратам.

В разработанном и импортированном в информационную систему глоссарии (рис. 5) имеются следующие колонки:

– начальная буква – как в словаре;
– термин на русском языке;
– термин на английском языке;
– определения термина;
– ссылки на первоисточники;
– ссылки на категории рубрикатора.

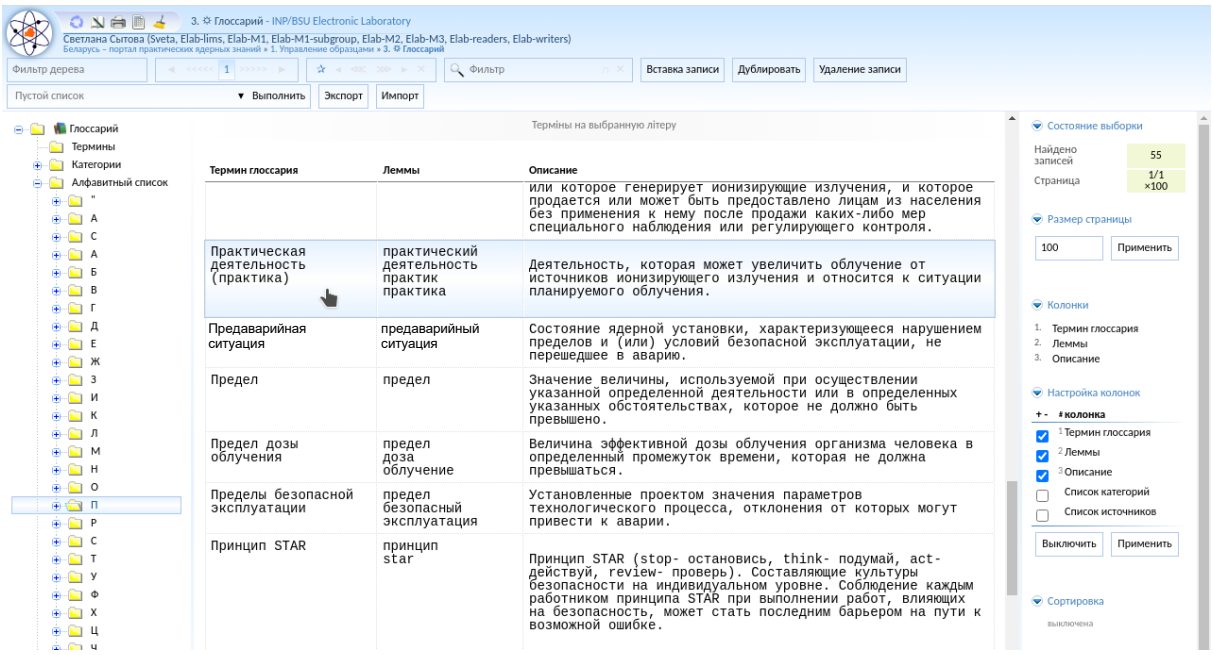


Рис. 5. Результаты импорта глоссария в информационную систему

Fig. 5. Results of the glossary import into the information system

Пример части списка категорий рубрикатора выглядит так:

1. Физические процессы и явления
   1.1. Общие понятия
   1.2. Единицы физических величин
2. Ионизирующие излучения
3. Атомное ядро и элементарные частицы
   3.1. Элементарные частицы
   3.2. Радионуклиды и химические элементы
   3.3. Ядерные процессы
4. Радиоактивность и радиоактивное вещество
5. Источники ионизирующего излучения (ИИИ)
   5.1. Закрытые ИИИ

5.2. Открытые ИИИ
5.3. Генерирующее оборудование
5.4. Работа с ИИИ.

На основе глоссария разработана концептуальная модель БД, в которой будет храниться вся информация для последующей автоматической обработки (рис. 6).

В приведенной модели требует пояснения только один момент – соотношение «Понятие» и «Синоним». Если посмотреть на рис. 5, то в колонке «Термин глоссария» таблицы формулировка термина имеет два варианта написания: практическая деятельность и практика. Такие случаи типичны для терминов рассматриваемой предметной области. В качестве примеров можно привести:

– тепловыделяющий элемент (ТВЭЛ);

– технико-экономическое обоснование (ТЭО);

– экспертиза безопасности в области использования атомной энергии и источников ионизирующего излучения (экспертиза безопасности).

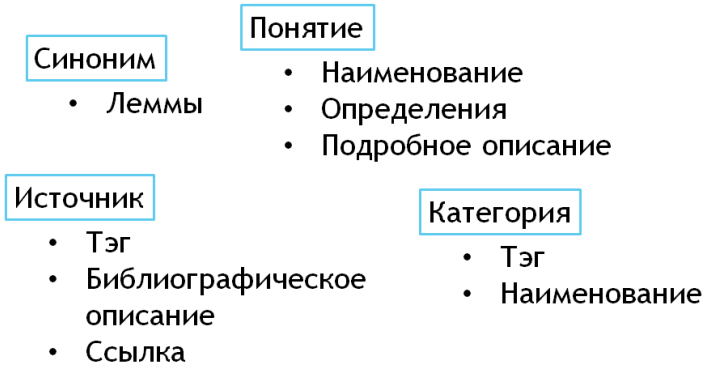


Рис. 6. Концептуальная модель БД глоссария
Fig. 6. Conceptual model of the glossary database

Таким образом, при поиске ключевых слов в тексте необходимо отдельно обнаруживать ТВЭЛ и «тепловыделяющий элемент», но соотносить их с одними и теми же определениями (понятиями) из ПрО. Соответственно, ТВЭЛ и «тепловыделяющий элемент» – это синонимы, которые концептуально в рамках модели относятся к понятию «тепловыделяющий элемент (ТВЭЛ)».

На рис. 6 особый интерес представляет колонка таблицы «Леммы». Это результат специальной обработки отдельных слов из состава понятия для последующего поиска данного понятия в тексте на естественном языке. Так, понятие «тепловыделяющий элемент (ТВЭЛ)» может упоминаться в тексте в различных словоформах в составе, например, следующих предложений:

– …. ТВЭЛы загружены в реактор …;

– …. в хранилище для тепловыделяющих элементов …

Все эти варианты использования понятия система должна находить в текстах и корректно обрабатывать. Для этого производится морфологический анализ слов понятия и выделяются леммы, т. е. проводится лемматизация.

Для поиска лемм, которые являются основой слова, его неизменяемой частью, выражающей лексическое значение, используются алгоритмы из состава библиотеки проекта PHPMorphy (https://github.com/cijic/phpmorphy). Словари этой библиотеки основаны на разработках проекта «Автоматическая обработка текста» (http://aot.ru/docs/sokirko/Dialog2004.htm). Если в словаре нет соответствующего слова, то применяется алгоритм стемминга – обработки словоформы с использованием эвристических правил для получения основы слова [20, 21]. В системе управления контентом eLab-Science применяется алгоритм Snowball (https://snowballstem.org/),

который поддерживает разные языки и реализован для большого числа средств разработки. Исходные тексты реализаций доступны в Интернете по открытым лицензиям.

В результате проблема словоформ решается, но возникает сложность с различными семантиками (смысловыми нагрузками) отдельных слов в тексте. В качестве примера можно привести анализ тестовой статьи «Атом» (рис. 7), где были выявлены следующие ключевые слова: конечное состояние, процесс, работа, система, УЕ, элементы.

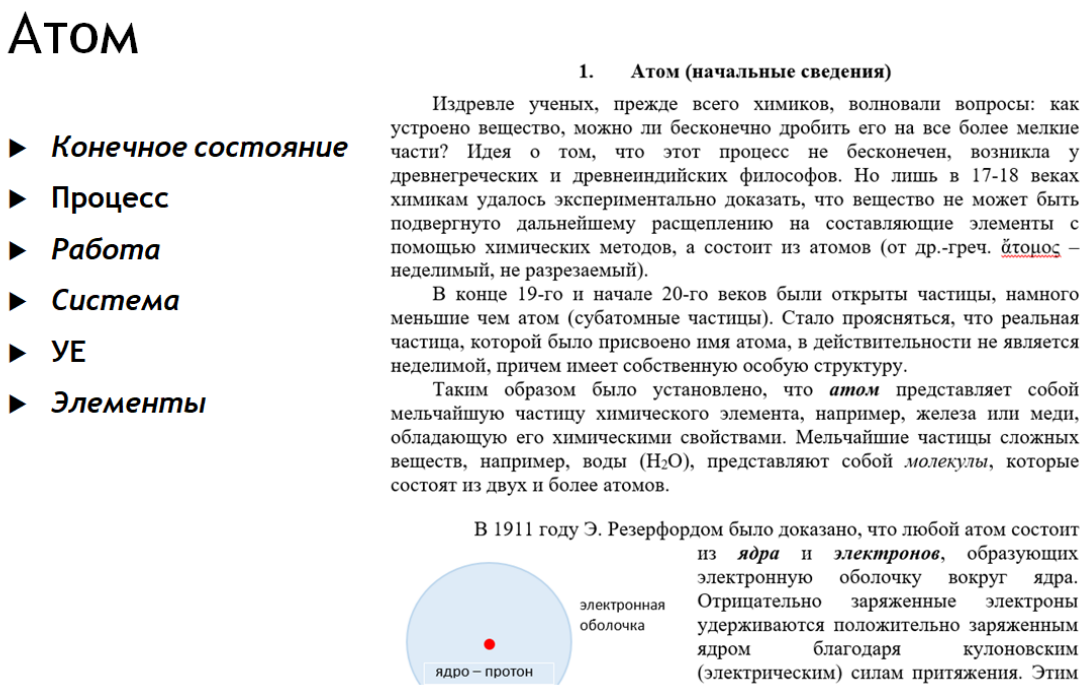


Рис. 7. Обработка тестовой статьи «Атом» с выделением ключевых слов

Fig. 7. Processing the test article "Atom" with the keywords highlighting

Со словом «процесс» получилась следующая семантическая коллизия. В глоссарии BelNET значение термина «процесс» выглядит так: «Процесс – последовательность действий или операций, в особенности ряд последовательных стадий изготовления продукта или некоторых других операций. Ряд взаимосвязанных или взаимодействующих операций, которые преобразуют вкладываемые ресурсы в конечные результаты».

В анализируемом тексте «Атом» есть фраза: «Издревле ученых, прежде всего химиков, волновали вопросы: как устроено вещество, можно ли бесконечно дробить его на все более мелкие части? Идея о том, что этот процесс не бесконечен, возникла у древнегреческих и древнеиндийских философов». То есть семантика в тексте и семантика в определении для слова «процесс» различаются.

С термином УЕ связана еще одна коллизия. В тексте статьи УЕ − это условная единица, а в глоссарии BelNET УЕ − это учетная единица (специальный термин из области обработки и хранения ядерных отходов).

На рис. 8 показан результат обработки тестовой статьи «Ядро». Здесь имеет место следующее нарушение семантики для термина «фон». В тексте статьи фраза с этим понятием звучит так: «…их взаимное электрическое отталкивание даже на фоне сильного ядерного притяжения». Слово «фон» тут является связующим внутри предложения («синтаксический сахар», как говорят лингвисты), но в рамках ПрО ядерных технологий понятие «радиационный фон» − это важнейший термин, который имеет строго определенную семантику и контекст применения.

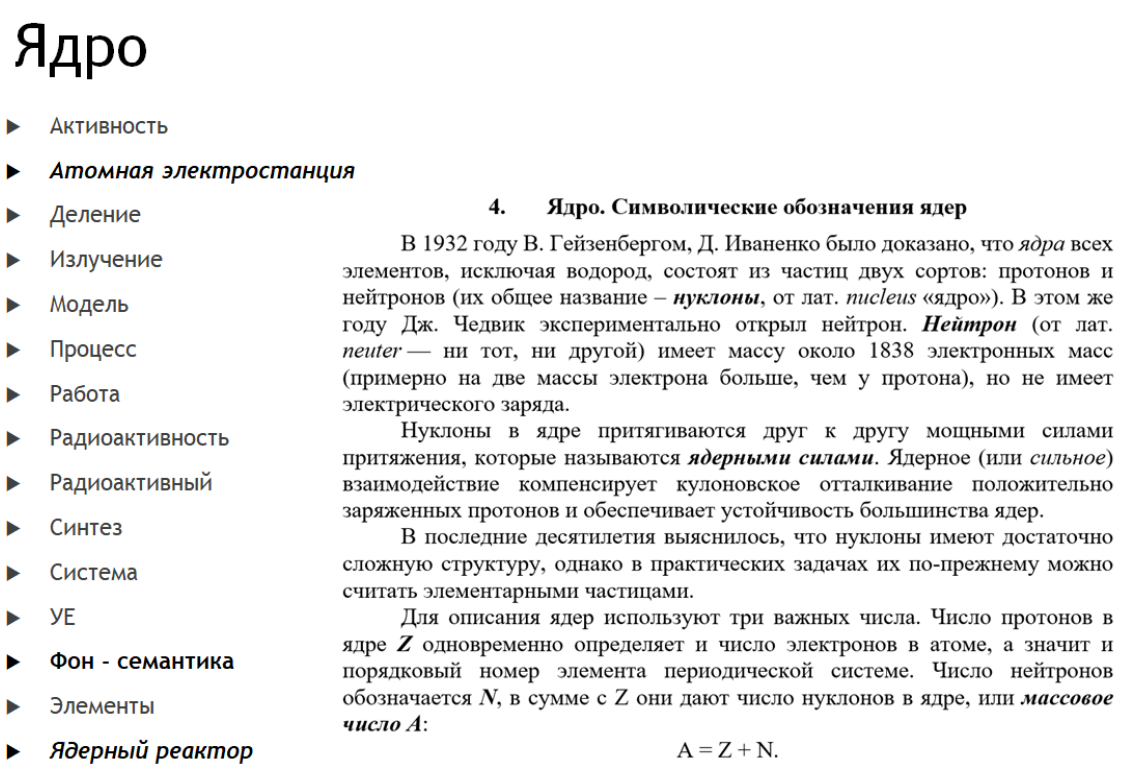


Рис. 8. Обработка статьи «Ядро» с выделением ключевых слов

Fig. 8. Processing the article "Nucleus" test with the keywords highlighting

Очевидно, что в вышеперечисленных и других аналогичных случаях проблему различия семантики в тексте и семантики термина из глоссария может решить только автор ресурса.

Вызывает интерес тот факт, что ключевые слова «атом» (рис. 7) и «ядро» (рис. 8) в данных текстах алгоритмом определены не были, так как тестируемый глоссарий по ядерной и радиационной безопасности попросту не содержит терминов «атом» и «ядро». Это означает, что для адекватной работы алгоритма необходимо использовать несколько тематических глоссариев.

Можно привести еще некоторое количество примеров подобного вида. На практике эта особенность преодолевается тем, что, как указывалось ранее, алгоритмы выделения ключевых слов работают в полуавтоматическом режиме. Список найденных ключевых слов предлагается пользователю, который сам выбирает, включать их в метаданные своего текста или не включать. Соответственно, пользователь из списка в левой колонке выберет то, что относится к его тексту, и исключит из предложенного списка те ключевые слова, которые не относятся к его предметной области.

Понятно, что глоссарии не должны содержать очень подробные «мелкие» термины. Это, с одной стороны, замедлит работу алгоритма из-за неоправданного объема глоссария, а с другой − выдаст большое количество терминов, которые не должны быть «ключевыми словами», и автоматически заставит пользователя исключать большое количество терминов из предложенного автоматически сформированного списка ключевых слов.

В отношении алгоритма определения положения создаваемого ресурса в таксономии портала начинает работать колонка «Список категорий», связанная с ключевым словом и с соответствующим одним или несколькими разделами в таксономии.

Оптимизация. Для проведения оптимизации алгоритма обработки задача, которая решается при поиске ключевых слов, может быть сформулирована как поиск пересечения нескольких цепочек лемм: цепочка лемм документа и набор цепочек лемм ключевых слов. В итоге необходима выборка цепочек, результаты пересечений которых имеют мощность (число элементов), отличную от нуля. В общем случае у этой задачи высокий уровень вычислительной сложности,

что может быть проблемой для больших текстов и ограничивать максимальный размер глоссария. В текущей реализации предлагаются следующие способы оптимизации: буферизация редко меняющихся данных и минимизация числа запросов к БД.

Буферизация редко меняющихся данных – это простой и широко используемый способ ускорения поисковых алгоритмов. Он основан на том, что после разработки содержимое глоссария и (или) тезауруса меняется очень редко. Такие изменения, как правило, представляют собой исправление отдельных ошибок, недоработок и опечаток. Оно проводится как процедура согласования между пользователем, который обнаружил недоработку, автором, внесшим соответствующий термин, и сотрудником, ответственным за сайт. Это в реальности не происходит динамически и за несколько секунд. Поэтому можно разрабатывать алгоритм обработки, предполагая, что в течение одного сеанса обработки данные глоссария меняться не будут и можно разместить их в буфере в оперативной памяти на старте алгоритма, предварительно подготовив данные для поиска вхождений цепочек лемм. Далее, пока обрабатывается текущая группа текстов, содержимое этого буфера не меняется.

Следует отметить, что объем данных типичного глоссария после подготовки невелик с точки зрения аппаратных возможностей современного компьютера: 12 Мб для глоссария из 525 терминов. Как аргумент для использования алгоритма обработки можно привести и то, что в случае обнаружения серьезной недоработки в каком-либо глоссарии либо внедрения нового глоссария потребуется обработка всех текстов заново. Это в любом случае приведет к перезапуску системы и полной перезагрузке всех данных.

Для минимизации числа запросов к БД было проведено испытание системы в разных аппаратных конфигурациях. Оно показало, что целесообразно загружать все данные, относящиеся к конкретному тексту сразу, т. е. лучше загрузить весь массив лемм документа, чем осуществлять фильтрацию по таблице отдельных ключевых слов.

При работе с реальным глоссарием формируются сотни (около 600 в конкретном тесте) простых запросов с фильтрацией по строковому значению. Эти запросы очень быстро выполняются сервером, но при этом каждый из них влечет за собой затраты на обмен данными: передачу запроса в БД и получение данных оттуда. Эта величина постоянна, но сотни повторений создают расход процессорного ресурса, который будет меньше, если применить альтернативный метод: загрузить все леммы текста и выполнить операции поиска в оперативной памяти, не привлекая БД.

В таблице и на рис. 9 показаны результаты нагрузочного испытания реализованного алгоритма для выяснения поведения системы на текстах различного объема. Тестируемый глоссарий содержит 525 терминов. В качестве проверочных текстов использовались реальные документы различных объемов и форматов из состава нормативной базы, которые размещены в открытом доступе на сайте Департамента по ядерной и радиационной безопасности Министерства чрезвычайных ситуаций Республики Беларусь (Госатомнадзор, https://gosatomnadzor.mchs.gov.by/), в интересах которого проводится данная работа.

Результаты испытаний работы системы под нагрузкой

Results of the system performance under load

Число символов *Number of symbols*	Число лемм в тексте *Number of lemmas in the text*	Число найденных терминов *Number of terms found*	Время обработки, с *Processing time, s*
15 438	1208	6	0,03
24 921	1951	15	0,05
182 764	17 669	43	0,45
4 870 803	165 478	113	45,50

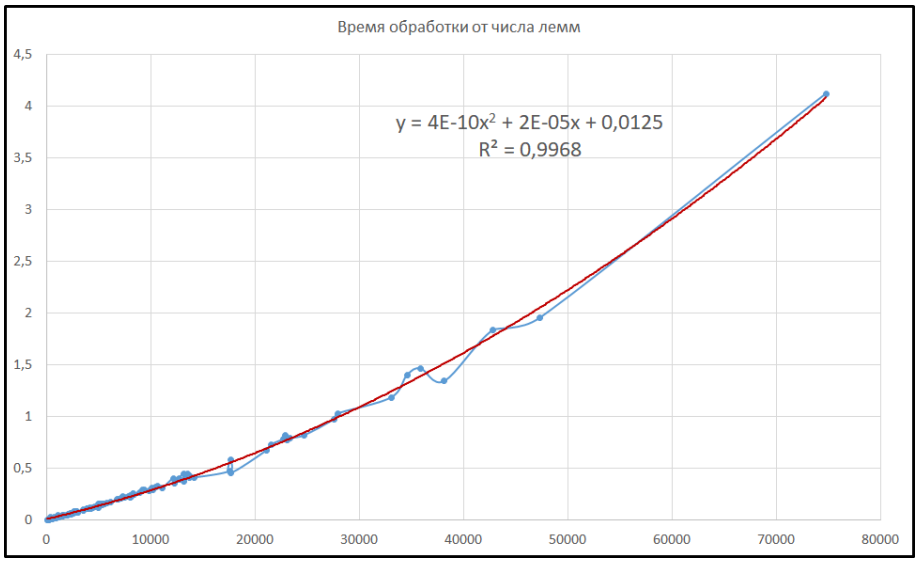


Рис. 9. Результаты испытаний работы системы и квадратичный тренд

Fig. 9. Results of the system performance and quadratic trend

На рис. 9 видно, что зависимость времени обработки от числа лемм скорее является квадратичной с $R^2 = 0{,}9968$, чем линейной с $R^2 = 0{,}9486$, при аппроксимации функцией $y = 4 \cdot 10^{-5}x - 0{,}08$.

Компьютер, на котором проводились испытания, был оснащен процессором Intel i5-7200U (тактовая частота 2,5 ГГц). Объем оперативной памяти практически не влиял на результат, так как требуемый объем ОЗУ не превысил 50 Мб. Оказалось, что даже для очень больших документов (в рассмотренном примере самый большой документ содержал около 200 страниц) время работы находится в приемлемых рамках – 4,1 с. Типичные документы – статьи и т. п. объемом несколько страниц – обрабатываются менее секунды.

Поскольку от рассмотренных алгоритмов не требуется мгновенный результат, работа может быть выполнена в фоновом режиме. Также очевидно, что размещение огромных документов одной записью зачастую не очень оптимально с точки зрения читабельности и наглядности материала. Понятно, что гораздо лучше использовать разбиение материала на главы и разделы.

Заключение. Портал BelNET постоянно развивается и бесперебойно функционирует с 2016 г. За прошедшее время число посетителей портала исчисляется тысячами, не только из Республики Беларусь и стран СНГ, но и со всего мира, со всех континентов, о чем свидетельствуют счетчики посещений портала, установленные в системе.

Подчеркнем, что процесс наполнения портала информацией и заполнения базы знаний, разработки специальных материалов для системы дистанционного обучения любого портала, тем более портала ядерных знаний, трудоемкий и длительный. И в этом смысле работа над BelNET находится в самом начале.

Можно констатировать также, что результаты реализации разработанных семантических алгоритмов, описанные в данной статье, являются очень хорошими, а их внедрение на портале BelNET позволяет отнести его к семантическим порталам.

Вклад авторов. *С. Н. Сытова* осуществила постановку проблемы, научное руководство ее решением, разработала структуру тезауруса и некоторые глоссарии, подготовила статью к публикации. *А. П. Дунец* реализовал семантические алгоритмы, включая полнотекстовый поиск в системе, а также провел оптимизацию работы и тестирование семантических алгоритмов. *А. Н. Коваленко* и *В. В. Гавриловец* разработали ядро системы eLab-Science. *С. В. Черепица* разработал структуру тезауруса и составил некоторые глоссарии.

Список использованных источников

1. Maintaining Knowledge, Training and Infrastructure for Research and Development in Nuclear Safety: INSAG-16. – Vienna : IAEA, 2003. – 19 p.

2. Knowledge Management for Nuclear Industry Operating Organizations. IAEA-TECDOC-1510. – Vienna : IAEA, 2006. – 185 p.

3. Knowledge Management and Its Implementation in Nuclear Organizations. IAEA Nuclear Energy Series No. NG-T-6. 10. – Vienna : IAEA, 2016. – 52 p.

4. Managing Nuclear Safety Knowledge: National Approaches and Experience. Safety Reports Series No. 105. – Vienna : IAEA, 2021. – 45 p.

5. Арсентьев, С. В. Корпоративная система современных ядерных знаний / С. В. Арсентьев // Глобальная ядерная безопасность. – 2023. – № 1(46). – С. 92–103.

6. Exploring Semantic Technologies and Their Application to Nuclear Knowledge Management. IAEA Nuclear Energy Series No. NG-T-6.15. – Vienna : IAEA, 2021. – 62 p.

7. Сытова, С. Н. Система управления ядерными знаниями в Республике Беларусь / С. Н. Сытова // Журнал БГУ. Физика. – 2022. − № 2. – С. 87–98.

8. Управление ядерными знаниями в системе научно-технической информации Республики Беларусь / С. Н. Сытова [и др.] // Развитие информатизации и государственной системы научно-технической информации (РИНТИ-2022) : докл. XXI Междунар. науч.-техн. конф., Минск, 17 нояб. 2022 г. – Минск : ОИПИ НАН Беларуси, 2022. – С. 265–269.

9. Nuclear knowledge management in the Republic of Belarus / S. Sytova [et al.] // Nonlinear Dynamics and Applications. – 2022. – Vol. 28. – P. 440–449.

10. Белорусский портал ядерных знаний BelNET: вчера, сегодня, завтра / С. Н. Сытова [и др.] // Сахаровские чтения 2023 года: экологические проблемы XXI века : материалы 23-й Междунар. науч. конф., Минск, Беларусь, 18–19 мая 2023 г. : в 2 ч. – Минск, 2023. – Ч. 2. – С. 158–162.

11. Онтологии и тезаурусы: модели, инструменты / Б. В. Добров [и др.]. − М. : Бином. Лаборатория знаний, 2009. − 173 с.

12. Лукашевич, Н. В. Тезаурусы в задачах информационного поиска / Н. В. Лукашевич. − М. : Изд-во МГУ, 2011. − 512 с.

13. Кириллович, А. В. Программная система для разработки многоязычного тезауруса / А. В. Кириллович, А. М. Баширов, А. Р. Гатиатуллин // Программные продукты и системы. – 2018. – Т. 31. – С. 112–120.

14. UNESCO SC/W/255. Guidelines for the Establishment and Development of Monolingual Thesauri. – N. Y. : UNESCO, 1973. − 37 p.

15. INIS Thesaurus. English Version. IAEA-INIS Reference Series. IAEA-INIS-01 (2018/09). – Vienna : IAEA, 2018. – 1312 p.

16. Михалевич, М. М. О лингвистических аспектах подготовки национального глоссария по ядерной и радиационной безопасности Республики Беларусь / М. М. Михалевич, Н. Н. Тушин // Сахаровские чтения 2022 года: экологические проблемы XXI века : материалы 22-й Междунар. науч. конф., Минск, Беларусь, 19–20 мая 2022 г. : в 2 ч. – Минск, 2022. – Ч. 1. – С. 176−179.

17. Информационная система eLab для аккредитованных испытательных лабораторий на основе свободного программного обеспечения / С. Н. Сытова [и др.] // Информатика. – 2017. − № 3(55). – С. 49–61.

18. Сытова, С. Н. Информационная система eLab в науке, практике, образовании / С. Н. Сытова. – Минск : Изд. центр БГУ, 2021. – 202 с.

19. Марка, Д. А. Методология структурного анализа и проектирования SADT : пер. с англ. / Д. А. Марка, К. МакГоуэн. – М. : Метатехнология, 1993. – 240 с.

20. Lovins, J. B. Development of a stemming algorithm / J. B. Lovins // Mechanical Translation and Computational Linguistics. – 1968. – Vol. 11. – P. 22–31.

21. Frakes, W. B. Strength and similarity of affix removal stemming algorithms / W. B. Frakes, C. J. Fox // SIGIR Forum. – 2003. – Vol. 37. – P. 26–30.

References

1. *Maintaining Knowledge, Training and Infrastructure for Research and Development in Nuclear Safety: INSAG-16*. Vienna, IAEA, 2003, 19 p.

2. *Knowledge Management for Nuclear Industry Operating Organizations*. *IAEA-TECDOC-1510*. Vienna, IAEA, 2006, 185 p.

3. *Knowledge Management and Its Implementation in Nuclear Organizations. IAEA Nuclear Energy Series No. NG-T-6. 10*. Vienna, IAEA, 2016, 52 p.

4. *Managing Nuclear Safety Knowledge: National Approaches and Experience. Safety Reports Series No. 105*. Vienna, IAEA, 2021, 45 p.

5. Arsentev S. V. *Corporate system of modern nuclear knowledge*. Global'naja jadernaja bezopasnost' [*Global Nuclear Security*], 2023, no. 1(46), pp. 92–103 (In Russ.).

6. *Exploring Semantic Technologies and Their Application to Nuclear Knowledge Management. IAEA Nuclear Energy Series No. NG-T-6.15*. Vienna, IAEA, 2021, 62 p.

7. Sytova S. N. *Nuclear knowledge management system in the Republic of Belarus*. Zhurnal Belorusskogo gosudarstvennogo universiteta. Fizika [*Journal of the Belarusian State University. Physics*], 2022, no. 2, pp. 87–98 (In Russ.).

8. Sytova S. N., Bartkevich A. R., Verenich K. A., Gavrilovec V. V., Gurachevskij V. L., ..., Cherepica S. V. *Nuclear knowledge management in the scientific and technical information system of the Republic of Belarus*. Razvitie informatizacii i gosudarstvennoj sistemy nauchno-tehnicheskoj informacii (RINTI-2022) : doklady XXI Mezhdunarodnoj nauchno-tehnicheskoj konferencii, Minsk, 17 nojabrja 2022 g. [*Development of Informatization and the State System of Scientific and Technical Information (RINTI-2022) : Reports of the XXI International Scientific and Technical Conference, Minsk, 17 November 2022*]. Minsk, Ob"edinennyj institut problem informatiki Nacional'noj akademii nauk Belarusi, 2022, pp. 265–269 (In Russ.).

9. Sytova S., Charapitsa S., Kavalenka A., Dunets A., Haurilavets V. Nuclear knowledge management in the Republic of Belarus. *Nonlinear Dynamics and Applications*, 2022, vol. 28, pp. 440–449.

10. Sytova S. N., Bartkevich A. R., Verenich K. A., Gavrilovec V. V., Dunec A. P., ..., Cherepica S. V. *Belarusian nuclear knowledge portal BelNET: yesterday, today, tomorrow*. Saharovskie chtenija 2023 goda: jekologicheskie problemy XXI veka : materialy 23-j Mezhdunarodnoj nauchnoj konferencii, Minsk, Belarus', 18–19 maja 2023 g. : v 2 chastjah [*Sakharov Readings 2023: Environmental Problems of the XXI Century : Materials of the 23rd International Scientific Conference, Minsk, Belarus, 18–19 May 2023 : in 2 Parts*]. Minsk, 2023, part 2, p. 158–162 (In Russ.).

11. Dobrov B. V., Ivanov V. V., Lukashevich N. V., Solov'ev V. D. Ontologii i tezaurusy: modeli, instrumenty. *Ontologies and Thesauri: Models, Tools.* Moscow, Binom. Laboratoriya znanij, 2009, 173 p. (In Russ.).

12. Lukashevich N. V. Tezaurusy v zadachah informacionnogo poiska. *Thesauruses in Information Retrieval Tasks.* Moscow, Izdatel'stvo Moskovskogo gosudarstvennogo universiteta, 2011, 512 p. (In Russ.).

13. Kirillovich A. V., Bashirov A. M., Gatiatullin A. R. *Software system for developing a multilingual thesaurus*. Programmnye produkty i sistemy [*Software & Systems*], 2018, vol. 31, pp.112–120 (In Russ.).

14. *UNESCO SC/W/255. Guidelines for the Establishment and Development of Monolingual Thesauri*. New York, UNESCO, 1973, 37 p.

15. *INIS Thesaurus. English Version. IAEA-INIS Reference Series. IAEA-INIS-01 (2018/09).* Vienna, IAEA, 2018, 1312 p.

16. Mihalevich M. M., Tushin N. N. *On the linguistic aspects of the preparation of the national glossary on nuclear and radiation safety of the Republic of Belarus*. Saharovskie chtenija 2022 goda: jekologicheskie problemy XXI veka : materialy 22-j Mezhdunarodnoj nauchnoj konferencii, Minsk, Belarus', 19–20 maja 2022 g. : v 2 chastjah [*Sakharov Readings 2022: Environmental Problems of the XXI Century : Materials of the 22nd International Scientific Conference, Minsk, Belarus, 19–20 May 2022 : in 2 Parts*]. Minsk, 2022, part 1, pp. 176−179 (In Russ.).

17. Sytova S. N., Dunets A. P., Kovalenko A. N., Mazanik A. L., Sidorovich T. P., Charapitsa S. V. *Information system eLab for accredited testing laboratories*. Informatika [*Informatics*], 2017, no. 3(55), pp. 49–61 (In Russ.).

18. Sytova S. N. Informacionnaya sistema eLab v nauke, praktike, obrazovanii. *Information System eLab in Science, Practice, Education.* Minsk, Izdatel'skij centr Belorusskogo gosudarstvennogo universiteta, 2021, 202 p. (In Russ.).

19. Marca D. A., McGowan C. L. *Sadt: Structured Analysis and Design Techniques*. McGraw-Hill, 1987, 392 p.

20. Lovins J. B. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 1968, vol. 11, pp. 22–31.

21. Frakes W. B., Fox C. J. Strength and similarity of affix removal stemming algorithms. *SIGIR Forum*. 2003, vol. 37, pp. 26–30.

Информация об авторах

Сытова Светлана Николаевна, кандидат физико-математических наук, доцент, заведующий лабораторией, Институт ядерных проблем Белорусского государственного университета.
E-mail: sytova@inp.bsu.by
https://orcid.org/0000-0002-2476-9979

Гавриловец Виктор Васильевич, научный сотрудник, Институт ядерных проблем Белорусского государственного университета.
E-mail: bycel@tut.by
https://orcid.org/0000-0002-9452-7465

Дунец Андрей Петрович, старший научный сотрудник, Институт ядерных проблем Белорусского государственного университета.
E-mail: dunets@gmail.com
https://orcid.org/0009-0006-0980-7697

Коваленко Антон Николаевич, старший научный сотрудник, Институт ядерных проблем Белорусского государственного университета.
E-mail: anton.kavalenka@gmail.com
https://orcid.org/0000-0002-0320-2092

Черепица Сергей Вячеславович, кандидат физико-математических наук, доцент, ведущий научный сотрудник, Институт ядерных проблем Белорусского государственного университета.
E-mail: svcharapitsa@gmail.com
https://orcid.org/0000-0001-9657-1948

Information about the authors

Svetlana N. Sytova, Ph. D. (Phys.-Math.), Assoc. Prof., Head of the Laboratory, Institute for Nuclear Problems of the Belarusian State University.
E-mail: sytova@inp.bsu.by
https://orcid.org/0000-0002-2476-9979

Viktar V. Haurylavets, Researcher, Institute for Nuclear Problems of the Belarusian State University.
E-mail: bycel@tut.by
https://orcid.org/0000-0002-9452-7465

Andrei P. Dunets, Senior Researcher, Institute for Nuclear Problems of the Belarusian State University.
E-mail: dunets@gmail.com
https://orcid.org/0009-0006-0980-7697

Anton N. Kavalenka, Senior Researcher, Institute for Nuclear Problems of the Belarusian State University.
E-mail: anton.kavalenka@gmail.com
https://orcid.org/0000-0002-0320-2092

Siarhei V. Charapitsa, Ph. D. (Phys.-Math.), Assoc. Prof., Leading Researcher, Institute for Nuclear Problems of the Belarusian State University.
E-mail: svcharapitsa@gmail.com
https://orcid.org/0000-0001-9657-1948

МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ
MATHEMATICAL MODELING

УДК 519.24
https://doi.org/10.37661/1816-0301-2024-21-2-24-35

Оригинальная статья
Original Article

Применение моделей копул в анализе акций фондового рынка

А. М. Кендысь✉, Н. Н. Труш

Белорусский государственный университет,
пр. Независимости, 4, Минск, 220030, Беларусь
✉*E-mail: kendyslesha@gmail.com*

Аннотация

Цели. Целью исследования является применение моделей копул для анализа акций российского фондового рынка и описания изменения зависимости между акциями до и во время коронавирусной инфекции (COVID-19).

Методы. Приводится алгоритм использования копул и функций языка программирования R при его реализации. Для описания динамики финансовых рядов используется модель ARMA-GJR-GARCH (ARMA-Glosten-Jagannathan-Runkle-GARCH, модель авторегрессии – скользящего среднего Глостен – Джаганнатан – Ранкл с обобщенной авторегрессионной условной гетероскедастичностью). Осуществляется подбор оптимальных семейств и параметров моделей копул. Проверяется адекватность полученных моделей и анализируются результаты исследования взаимосвязи между данными рядами.

Результаты. Разработан алгоритм для относительно нового подхода использования копул в связке с моделью ARMA-GJR-GARCH. Подход применен для исследования влияния коронавируса в контексте российской экономики. Выявлено, что в период COVID-19 зависимость между различными акциями фондового рынка возрастает. Показано, что эффект волатильности финансовых рядов увеличивается после вспышки пандемии.

Заключение. Алгоритм исследования с помощью моделей копул в связке с моделью ARMA-GJR-GARCH показал свою целесообразность. Данный подход можно использовать и с применением других моделей GARCH-типа для исследования финансов и других сфер.

Ключевые слова: копула, модель ARMA-GJR-GARCH, фондовый рынок, акции, коронавирусная инфекция, математическое моделирование

Для цитирования. Кендысь, А. М. Применение моделей копул в анализе акций фондового рынка / А. М. Кендысь, Н. Н. Труш // Информатика. – 2024. − Т. 21, № 2. – С. 24–35.
https://doi.org/10.37661/1816-0301-2024-21-2-24-35

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию | Received 27.12.2023
Подписана в печать | Accepted 21.03.2024
Опубликована | Published 28.06.2024

Application of copula models in stock market analysis

Alexey M. Kendys✉, Mikolai M. Troush

Belarusian State University,
av. Nezavisimosti, 4, Minsk, 220030, Belarus
✉*E-mail: kendyslesha@gmail.com*

Abstract
Objectives. The objective of the study is to use copula models to analyze shares of the Russian stock market and describe changes in the relationship between the shares before and during the coronavirus infection (COVID-19).
Methods. An algorithm for using copulas and functions of the R programming language in its implementation is presented. To model the dynamics of financial series the ARMA-GJR-GARCH process (autoregressive moving average Glosten-Jagannathan-Runkle model with generalized autoregressive conditional heteroskedasticity) is used. The selection of optimal families and parameters of copula models is carried out. The adequacy of the obtained models is checked and the results of the study of the relationship between the series are analyzed.
Results. An algorithm has been developed for a relatively new approach to using copulas in conjunction with the ARMA-GJR-GARCH model. The approach was used to study the impact of coronavirus in the context of the Russian economy. It is revealed that during the COVID-19 period the dependence between different stocks increases. It is shown that the effect of volatility in financial series increases after the outbreak of the pandemic.
Conclusion. The research algorithm using copula models in conjunction with the ARMA-GJR-GARCH process has shown its feasibility. This approach can be used with other GARCH-type models to study finance and other areas.

Keywords: copula, ARMA-GJR-GARCH model, stock market, shares, coronavirus infection, mathematical modelling

For citation. Kendys A. M., Troush M. M. *Application of copula models in stock market analysis*. Informatika [*Informatics*], 2024, vol. 21, no. 2, pp. 24–35 (In Russ.). https://doi.org/10.37661/1816-0301-2024-21-2-24-35

Conflict of interest. The authors declare of no conflict of interest.

Введение. В настоящее время активное исследование и применение копул происходит в таких областях, как, например, управление финансами, анализ страховых рисков, моделирование макроэкономических процессов. Э. Склар ввел копулы в 1959 г. как мощный инструмент для моделирования зависимости между переменными. Копулы – это функции, которые позволяют «связывать» многомерные функции распределения с их одномерными маргинальными функциями распределения. Данное представление оказывается удобнее, чем использование совместных функций распределений изучаемых показателей. Это обусловлено тем, что, с одной стороны, в копулах выделены маргинальные распределения показателей, что, конечно же, играет важную роль при исследовании реальных совокупностей, а с другой – выделена структура зависимости между найденными маргинальными распределениями.

Копулы активно применяются для управления финансовыми рисками, поскольку позволяют не только определять совместное распределение с помощью частных (маргинальных) функций и вида взаимосвязи, но и моделировать неэллипсообразные многомерные распределения.

В книге [1] описываются базовые принципы применения копул в финансах, а работы [2, 3] иллюстрируют современный подход к применению моделей копул в этой области. В настоящей статье копулы используются для исследования влияния коронавирусной инфекции на акции российского фондового рынка. Изучение последствий COVID-19 на экономику уже было проведено в различных публикациях [4–7]. Хотя копулы уже использовались в некоторых исследованиях для анализа последствий коронавирусной инфекции, у работ, фокусирующихся на российском фондовом рынке, есть ряд недостатков, к тому же использование модели ARMA-GJR-GARCH в связке с копулами является относительно новым подходом. Настоящая работа направлена на дополнение существующих исследований с применением современных методов матема-

тического моделирования, а также внесение нового вклада в литературу об исследовании влияния COVID-19 в контексте российской экономики. Помимо этого, в работе представлена разработка алгоритма исследования на языке R, который можно использовать в дальнейшем для изучения других данных и сравнения эффективности различных моделей.

Теория копул. Для начала приведем определение копулы и основную теорему теории копул.

Функция $C(u,v)$, принимающая значения на $[0;1]$, называется копулой двух переменных u и v, где $u,v \in [0;1]$ (т. е. $C(u,v)$ действует из $[0;1]^2$ в $[0;1]$), если она удовлетворяет следующим условиям [8, с. 10]:

1) $C(u,0)=0,\ C(0,v)=0$;

2) $C(1,v)=v,\ C(u,1)=u$;

3) $C(u_2,v_2)+C(u_1,v_1)-C(u_2,v_1)-C(u_1,v_2)\geq 0$, где $u_1,v_1,u_2,v_2 \in [0;1]$ и $u_1 \leq v_1,\ u_2 \leq v_2$.

Аналогичное определение можно привести и для копулы размерности n [8, с. 45].

Теорема Склара [1, с. 24]. *Функцию распределения случайного вектора* $X=(X_1,\ldots,X_n)$ *со значениями в* $\mathbb{R}^n$ *обозначим* $F_{X_1\ldots X_n}(x_1,\ldots,x_n)=P\{X_1<x_1,\ldots,X_n<x_n\}$. *Пусть* $F_{X_j}(x_j)=P\{X_j<x_j\},\ j=\overline{1,n}$ *– маргинальные функции распределения отдельных компонент. Тогда существует такая n-мерная копула* $C(u_1,\ldots,u_n)$, *что для любых* $x_j \in \mathbb{R},\ j=\overline{1,n}$, *справедлива формула*

$$F_{X_1\ldots X_n}(x_1,\ \ldots,x_n)=C\left(F_{X_1}(x_1),\ \ldots,\ F_{X_n}(x_n)\right). \tag{1}$$

Если функции $F_{X_j}(x_j)$ *непрерывны,* $j=\overline{1,n}$, *то такая копула* C *единственная.*

Таким образом, копула позволяет перейти от одномерных распределений нескольких случайных величин к их совместному распределению.

Методология. Перед тем как измерять зависимость между финансовыми рядами, извлечем стандартизированные остатки данных для учета некоторых свойств финансовых временных рядов. Используем связку моделей ARMA(p, q)-GJR-GARCH(1, 1) [9], применяемую для стационарных временных рядов.

Модель авторегрессии – скользящего среднего, или ARMA(p, q), где p, q – целые числа, задается для временного ряда r_t в виде равенства [9]

$$r_t = c + \varepsilon_t + \sum_{i=1}^{p} \alpha_i r_{t-i} + \sum_{i=1}^{q} \beta_i \varepsilon_{t-i}, \tag{2}$$

где c – константа (отвечает за среднее значение); $\{\varepsilon_t\}$ – остатки модели ARMA; α_i и β_j – авторегрессионные (AR) коэффициенты и коэффициенты скользящего среднего (MA) соответственно (являются действительными числами), $i=\overline{1,p},\ j=\overline{1,q};\ t \in \mathbb{Z}$.

Сильная модель GJR-GARCH(1, 1), введенная Глостеном, Джаганнатаном и Ранклом, использует понятие гетероскедастичности и для процесса ε_t из равенства (2) задается соотношениями [9]

$$\varepsilon_t = \sigma_t z_t,\ \sigma_t^2 = \omega + (\alpha + \gamma I_{t-1})\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2, \tag{3}$$

где z_t – последовательность независимых одинаково распределенных случайных величин с нулевым средним, единичной дисперсией и некоторым заданным распределением;

$I_{t-1} = \begin{cases} 0, \ \varepsilon_{t-1} \geq 0 \\ 1, \ \varepsilon_{t-1} < 0 \end{cases}$; σ_t^2 – условная дисперсия; $\omega, \alpha, \beta, \gamma$ – параметры модели, где $\omega > 0, \ \alpha \geq 0, \ \beta \geq 0, \ \gamma \geq 0$. Параметр γ относится к модификации GJR и описывает эффект финансового рычага. Величины z_t и σ_t независимы.

В отличие от сильной модели слабая модель GJR-GARCH(1, 1) используется для процессов, у которых не существует вторых, а иногда и первых моментов.

В качестве распределения z_t будем рассматривать скошенное распределение Стьюдента. Для оценивания зависимости между стандартизированными остатками z_t будем применять копулы [3]. Найдем оптимальные семейства копул и их параметры и, используя их как оценки совместных распределений остатков, исследуем зависимость между рядами. Вычисления произведем с помощью языка программирования R [10]. Опишем алгоритм исследования и укажем основные функции, использованные в языке R для получения результатов.

Алгоритм исследования:

1. Преобразование исходных данных. Удаление тренда (функция *detrend*), стандартизация данных (функция *scale*).

2. Предварительный статистический анализ рассматриваемых временных рядов. Вывод описательных статистик (функция *descr*).

3. Проверка рядов на стационарность и отсутствие нормальности. В частности, тест на нормальность Харке – Бера (функция *col_jarquebera*) и некоторые тесты на стационарность: расширенный тест Дики – Фуллера (или ADF-тест, функция *adf.test*), тест Филлипса – Перрона (или PP-тест, функция *pp.test*) и тест Квятковского – Филлипса – Шмидта – Шина (или KPSS-тест, функция *kpss.test*, проверка на тренд).

4. Построение графиков Кендалла для предварительного анализа зависимости данных (функция *BiCopKPlot*).

5. Оценка оптимальных параметров *p* и *q* модели ARMA(*p*, *q*) (функция *auto.arima*).

6. Построение модели ARMA(*p*, *q*)-GJR-GARCH(1, 1) для полученных параметров *p* и *q* (функция *ugarchfit*).

7. Извлечение стандартизированных остатков из построенной модели (функция *residuals*).

8. Извлечение эффекта волатильности (функция *sigma*).

9. Построение графиков квантиль-квантиль для полученных остатков, основанных на скошенном распределении Стьюдента (функция *plot* с аргументом *which=9*).

10. Приведение стандартизированных остатков к псевдонаблюдениям с помощью функции *pobs*.

11. Выбор оптимальных моделей копул для описания зависимости между остатками рядов и оценка их параметров (функция *BiCopSelect*).

12. Генерация данных с помощью построенных копул и сравнение результатов с исходными данными (функция *BiCopSim*).

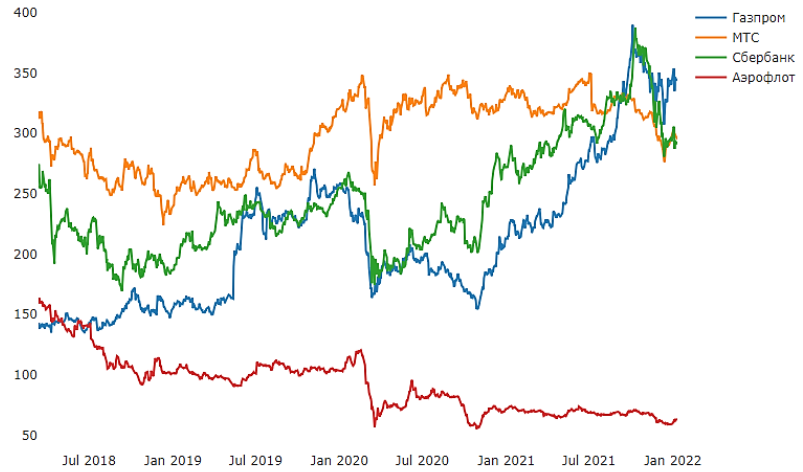
13. Сравнительный анализ зависимости рядов до коронавирусной инфекции и во время инфекции, основанный на полученных результатах и оптимальных копулах.

Исходные данные и их преобразование. Исследуются цены на акции следующих российских компаний: Лукойла, Газпрома, Магнита, МТС (Мобильные ТелеСистемы), Сбербанка, Аэрофлота. Соответственно, исследуются акции из шести различных областей: нефти, газа, торговли, телекоммуникаций, финансов и транспорта. Ежедневные данные были собраны с финансовой платформы *investing.com*[1] в период с 12.03.2018 по 11.01.2022.

На рис. 1, *a* представлены исходные данные для компаний Лукойл и Магнит, а на рис. 1, *b* – для компаний Газпром, МТС, Сбербанк, Аэрофлот (разные графики из-за большой разницы в значениях).

[1]Investing.com [Electronic resource] // Fusion Media Limited. – Mode of access: https://ru.investing.com. – Date of access: 10.09.2023.

a)



b)

Рис. 1. Цены на акции компаний Лукойл и Магнит (*a*) и Газпром, МТС, Сбербанк и Аэрофлот (*b*)

Fig. 1. Share prices of Lukoil and Magnit (a) and Gazprom, MTS, Sberbank and Aeroflot (b)

Периодом до пандемии считаем даты с 12.03.2018 по 29.01.2020, а даты с 30.01.2020 по 11.01.2022 считаем периодом пандемии.

Обозначим через $P(t)$ цены на акции, а через $R(t)$ – логарифмические доходности в момент времени t:

$$R(t) = \ln \frac{P(t)}{P(t-1)}. \tag{4}$$

Из доходностей удаляется линейный тренд и полученные данные стандартизируются.

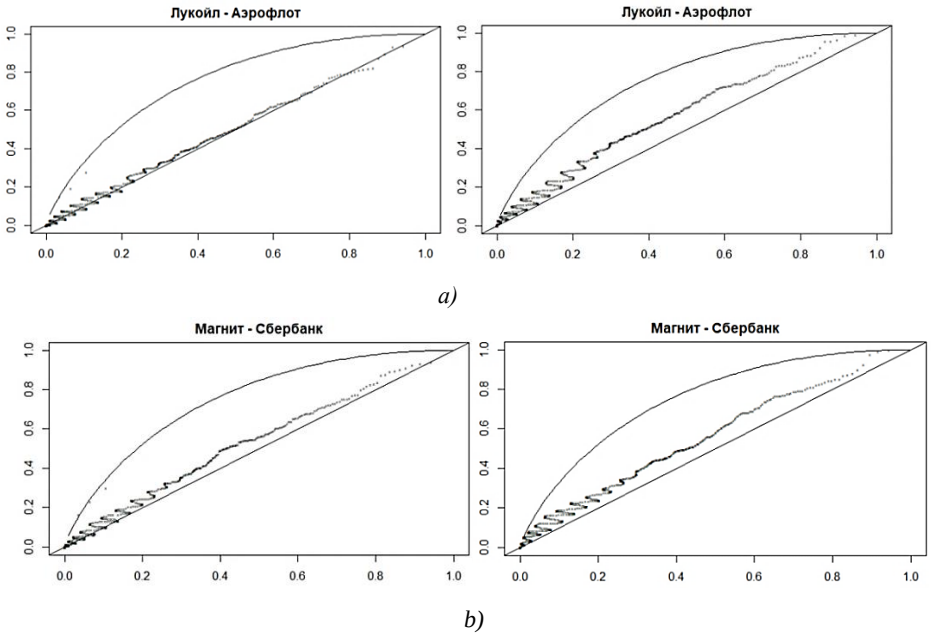
Описательные статистики и предварительный анализ зависимости. Для полученных доходностей были найдены описательные статистики, а также проведен тест на их нормальность Харке – Бера (приведено значение статистики), расширенный тест Дики – Фуллера (или ADF-тест, приведено значение статистики), тест Филлипса – Перрона (или PP-тест, приведено значение статистики) и тест Квятковского – Филлипса – Шмидта – Шина (или KPSS-тест, приведено значение статистики). Все результаты даны в табл. 1. Статистики тестов ADF и PP вычислены с уровнем значимости 0,01.

Таблица 1
Описательные статистики и результаты тестов
Table 1
Descriptive statistics and test results

	Статистики *Statistics*									
Компания *Company*	Выборочн. среднее *Mean*	Мин. *Min*	Макс. *Max*	Стат. отклонение *Std. dev.*	Коэф-т асимметр. *Skewness*	Коэф-т эксцесса *Kurtosis*	Тест Харке – Бера *Jarque-Bera test*	ADF	PP	KPSS
До пандемии										
Лукойл	0	–6,08	5,19	1	–0,03	4,11	341,8	–7,99	–427,36	0,05
Газпром	0	–4,64	8,85	1	1,85	15,32	4993,2	–7,7	–476,96	0,07
Магнит	0	–3,46	4,21	1	0,26	1,96	83,1	–7,6	–478,25	0,04
МТС	0	–5,73	3,3	1	–0,59	4,15	376,7	–9,03	–505,01	0,02
Сбербанк	0	–9,69	4,19	1	–1,89	18,47	7144,8	–7,97	–454,61	0,06
Аэрофлот	0	–5,5	4,11	1	0,1	4,82	470,8	–9,19	–421,81	0,03
Во время пандемии										
Лукойл	0	–8,7	6,15	1	–0,9	15,45	5010	–8,02	–520,92	0,07
Газпром	0	–5,25	3,62	1	–0,47	3,05	212	–7,25	–496,9	0,1
Магнит	0	–7,49	3,38	1	–1,05	8,87	1720	–7,59	–529,28	0,06
МТС	0	–6,43	5,42	1	–1,43	11,23	2782,4	–7,83	–521,95	0,03
Сбербанк	0	–4,93	5,97	1	–0,16	5,34	593,1	–7,29	–521,44	0,14
Аэрофлот	0	–6,15	4,59	1	–0,33	6,24	818	–5,95	–451,05	0,05

Из результатов тестов можно сделать вывод, что все ряды стационарны, не имеют трендов и не являются нормальными.

Далее для доходностей строятся графики Кендалла, основанные на двумерных копулах, с целью предварительного анализа зависимости между рядами. График Кендалла (Kendall Plot, или сокращенно K-Plot) – это аналог графика квантиль-квантиль для копул. Некоторые из графиков изображены на рис. 2.



a)

b)

Рис. 2. Графики Кендалла для Лукойла-Аэрофлота до (слева) и во время (справа) пандемии (*a*) и для Магнита-Сбербанка до (слева) и во время (справа) пандемии (*b*)

Fig. 2. Kendall plots for Lukoil-Aeroflot before (left) and during (right) the pandemic (a) and for Magnit-Sberbank before (left) and during (right) the pandemic (b)

Как видно из графиков, зависимость после пандемии либо увеличивается по сравнению с зависимостью до пандемии, либо остается примерно такой же.

Применение модели ARMA-GJR-GARCH. Для доходностей вычисляются оптимальные параметры p и q для модели ARMA(p, q). Результаты приведены в табл. 2.

Таблица 2
Оптимальные параметры модели
Table 2
Optimal model parameters

Компания *Company*	Параметры *Options*			
	p	q	p	q
	До пандемии		Во время пандемии	
Лукойл	0	0	0	0
Газпром	0	0	0	0
Магнит	3	1	0	0
МТС	0	0	0	0
Сбербанк	2	4	0	0
Аэрофлот	2	2	2	1

Далее была построена модель ARMA(p, q)-GJR-GARCH(1, 1), основанная на формулах (2) и (3). На рис. 3 показаны некоторые графики стандартизированных остатков и исходных рядов для сравнения. Видно, что стандартизированные остатки отличаются от исходных рядов только результатом эффекта волатильности.

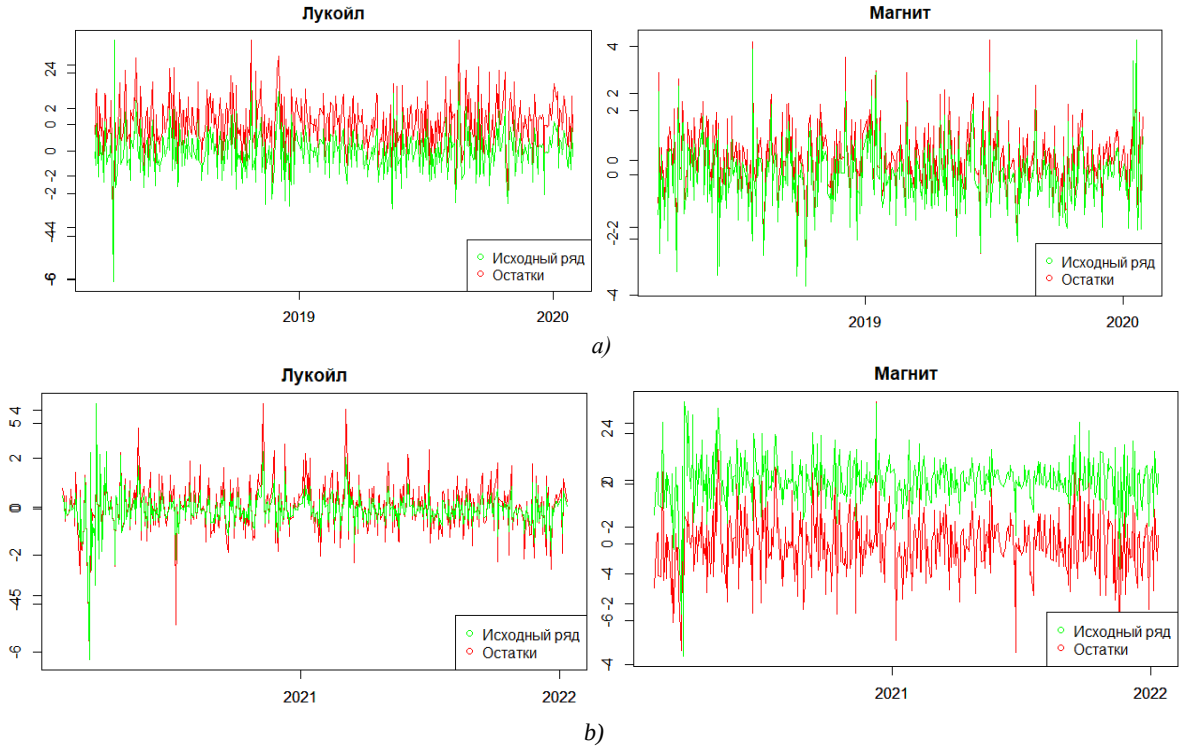
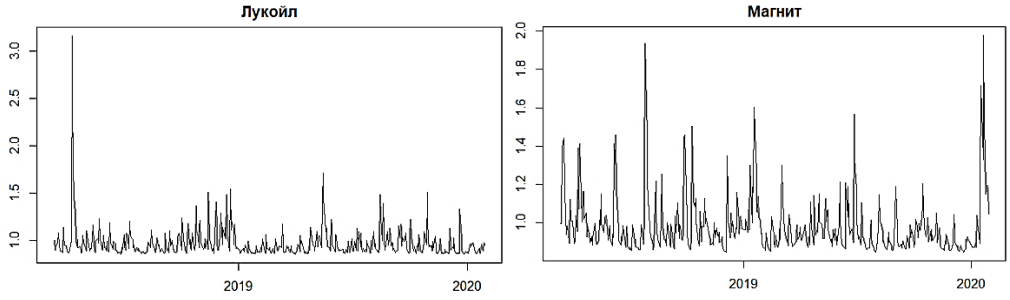


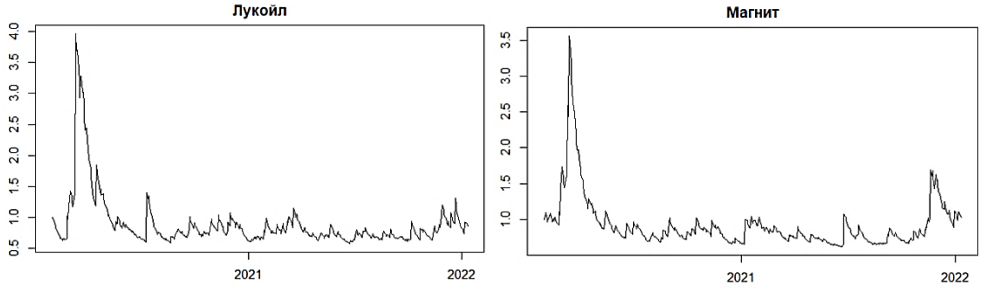
Рис. 3. Графики исходных рядов и стандартизированных остатков для Лукойла и Магнита до пандемии (*a*) и во время пандемии (*b*)

Fig. 3. Plots of initial series and standardized residuals for Lukoil and Magnit before the pandemic (a) and during the pandemic (b)

Волатильность, оцененная моделью, изображена на рис. 4. Видно, что эффект волатильности увеличивается после вспышки пандемии.



a)

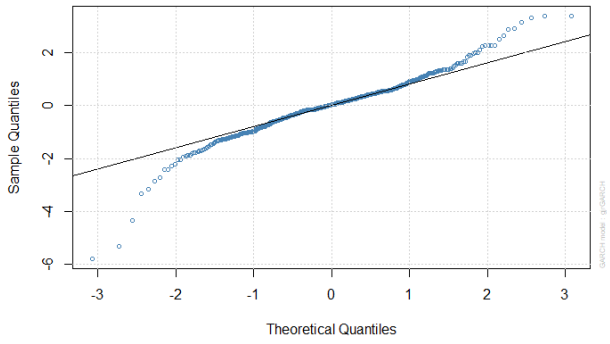


b)

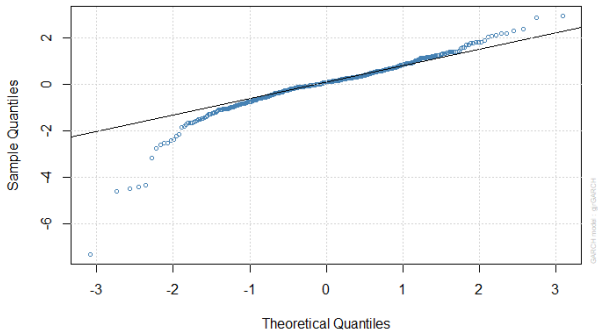
Рис. 4. Графики волатильности для Лукойла и Магнита до пандемии (*a*) и во время пандемии (*b*)

Fig. 4. Volatility plots for Lukoil and Magnit before the pandemic (a) and during the pandemic (b)

Для полученных стандартизированных остатков выведены графики квантиль-квантиль для проверки соответствия распределения оцененному скошенному распределению Стьюдента. На рис. 5 данные графики приведены для ряда МТС.



a)



b)

Рис. 5. График квантиль-квантиль для ряда МТС до пандемии (*a*) и во время пандемии (*b*)

Fig. 5. Quantile-quantile plot for the MTS series before the pandemic (a) and during the pandemic (b)

Представленные графики позволяют сделать вывод, что, хотя скошенное распределение Стьюдента отчасти соответствует данным, оно недостаточно точно описывает толстые хвосты реального распределения процесса. Соответственно, есть смысл использовать копулы для более точного описания совместного распределения данных и для более подробного анализа реальной зависимости между остатками рядов.

Подбор копул. Стандартизированные данные преобразуются в псевдонаблюдения и для преобразованных остатков подбираются подходящие копулы и их параметры. Для подбора наиболее подходящей копулы функция рассматривает целый ряд семейств, таких как *t*-копулы Стьюдента, копулы Гаусса, Клейтона, Франка, Тауна, Джо, а также BB1-, BB6-, BB7- и BB8-копулы [8]. В табл. 3 приведена вся информация о подобранных копулах, в том числе их параметры, название семейства, зависимость в нижних и правых хвостах (LTD и UTD соответственно), коэффициент корреляции Кендалла (Tao), критерий максимального логарифмического правдоподобия (LL, приведено значение критерия) и информационный критерий Акаике (AIC, приведено значение критерия). Критерии помогают понять, насколько хорошо копула аппроксимирует данные.

Таблица 3
Информация о копулах

Table 3
Copula information

Копула *Copula*	Параметр 1 *Parameter 1*	Параметр 2 *Parameter 2*	Семейство *Family*	Tao	LTD	UPD	LL	AIC
До пандемии								
Лукойл-Газпром	0,47	5,97	Student’s t	0,31	0,16	0,16	59,07	–114,14
Лукойл-Магнит	1,12	–	Survival Gumbel	0,11	0,15	0	8,43	–14,85
Лукойл-МТС	0,27	8,48	Student’s t	0,17	0,04	0,04	19,35	–34,69
Лукойл-Сбербанк	0,11	1,17	Survival BB1	0,19	0,19	0,01	24,05	–44,11
Лукойл-Аэрофлот	1,08	–	Survival Joe	0,05	0,1	0	3,4	–4,79
Газпром-Магнит	1,14	–	Survival Gumbel	0,12	0,16	0	10,63	–19,26
Газпром-МТС	2,83	0,7	Survival BB8	0,26	0	0	38,01	–72,01
Газпром-Сбербанк	5,36	0,5	Survival BB9	0,33	0	0	61,84	–119,68
Газпром-Аэрофлот	1,12	–	Survival Gumbel	0,11	0,15	0	10,93	–19,86
Магнит-МТС	1,21	–	Survival Gumbel	0,17	0,23	0	20,6	–39,2
Магнит-Сбербанк	1,35	0,35	Rotated Tawn type 1 180 degrees	0,14	0,18	0	16,42	–28,84
Магнит-Аэрофлот	0,1	4,16	Student’s t	0,06	0,09	0,09	11,66	–19,33
МТС-Сбербанк	2,33	–	Frank	0,24	0	0	32,35	–62,69
МТС-Аэрофлот	1,12	–	Survival Gumbel	0,11	0,15	0	10,04	–18,08
Сбербанк-Аэрофлот	1,11	–	Survival Gumbel	0,1	0,13	0	8,73	–15,46
Во время пандемии								
Лукойл-Газпром	0,22	1,57	Survival BB1	0,43	0,45	0,13	122,57	–241,14
Лукойл-Магнит	1,39	–	Survival Joe	0,18	0,35	0	30,97	–59,95
Лукойл-МТС	1,35	0,25	Survival BB7	0,24	0,33	0,06	44,04	–84,08
Лукойл-Сбербанк	0,56	–	Gaussian	0,38	0	0	89,89	–177,79
Лукойл-Аэрофлот	1,37	–	Survival Gumbel	0,27	0,34	0	52,02	–102,05
Газпром-Магнит	1,73	0,94	Survival BB8	0,23	0	0	37,29	–70,57
Газпром-МТС	1,65	0,47	Rotated Tawn type 2 180 degrees	0,24	0,31	0	50,21	–96,41
Газпром-Сбербанк	1,66	–	Survival Gumbel	0,4	0,48	0	110,34	–218,67
Газпром-Аэрофлот	1,39	–	Survival Gumbel	0,28	0,36	0	54,65	–107,29
Магнит-МТС	1,44	–	Survival Joe	0,2	0,38	0	38,08	–74,16
Магнит-Сбербанк	1,68	0,44	Rotated Tawn type 2 180 degrees	0,24	0,3	0	48,72	–93,44
Магнит-Аэрофлот	0,48	–	Clayton	0,19	0,23	0	30,65	–59,3
МТС-Сбербанк	1,34	–	Survival Gumbel	0,25	0,32	0	43,63	–85,27
МТС-Аэрофлот	1,23	–	Survival Gumbel	0,19	0,25	0	26,25	–50,5
Сбербанк-Аэрофлот	3,03	0,78	Survival BB8	0,34	0	0	71,43	–138,87

Далее с помощью построенных копул генерируются новые данные и сравниваются с исходными. Некоторые результаты показаны на рис. 6.

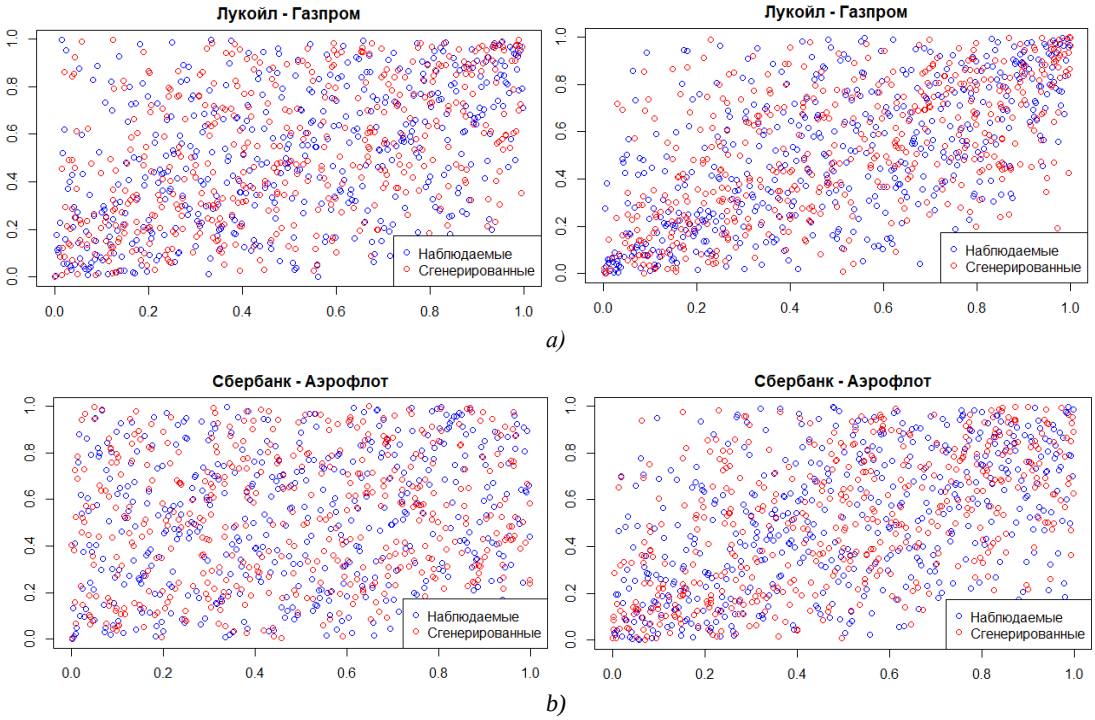


Рис. 6. Графики данных для Лукойла-Газпрома до (слева) и во время (справа) пандемии (*a*) и для Сбербанка-Аэрофлота до (слева) и во время (справа) пандемии (*b*)

Fig. 6. Data plots for Lukoil-Gazprom before (left) and during (right) the pandemic (a) and for Sberbank-Aeroflot before (left) and during (right) the pandemic (b)

На рис. 6 видно, что копулы адекватно описывают данные, так как наблюдаемые и сгенерированные наблюдения согласуются друг с другом. Также можно заметить, что зависимость после вспышки пандемии для исследуемых рядов увеличивается. Например, для Сбербанка-Аэрофлота данные до пандемии распределены по всей области равномерно, т. е. не имеют явной зависимости, а данные во время пандемии больше сконцентрированы посередине и направлены по диагонали вверх, что свидетельствует об умеренной положительной зависимости.

Анализ зависимости. На рис. 7 изображен график изменения зависимости до и после вспышки COVID-19 (на основании коэффициента корреляции Кендалла из табл. 3).

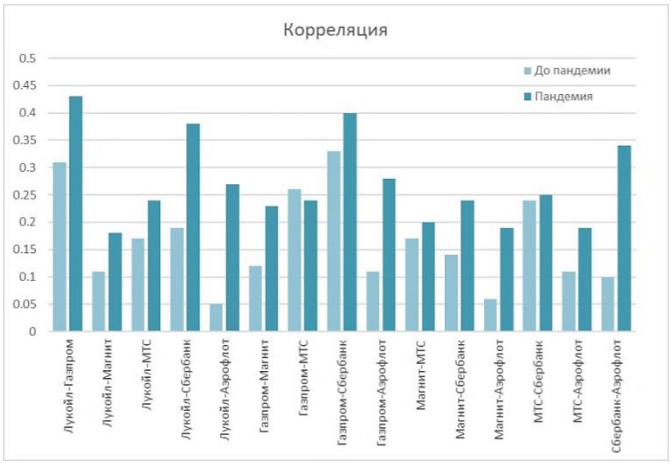


Рис. 7. Коэффициент корреляции в период до пандемии и во время пандемии

Fig. 7. Correlation coefficient before the pandemic and during the pandemic

Из графика видно, что почти для всех рядов зависимость увеличивается после вспышки COVID-19. Особенно это справедливо для компании Аэрофлот. Данный факт объясняется тем, что во время пандемии многие рейсы были отменены. Соответственно, акции авиалиний подверглись сильному изменению.

Заключение. В настоящей работе исследована зависимость между акциями некоторых российских компаний с помощью моделей копул. При моделировании финансовых рядов была применена модель ARMA-GJR-GARCH(1,1) и выявлено увеличение эффекта волатильности после коронавирусной инфекции. Выполнен оптимальный подбор семейств и параметров копул. При проверке адекватности модели осуществлена генерация значений на основе построенных моделей и получено детальное описание зависимости между акциями фондового рынка.

Из результатов проведенных исследований можно сделать вывод, что в период пандемии COVID-19 зависимость между различными акциями фондового рынка увеличивается. Полученный результат может быть связан с волнением на фондовом рынке, а также с негативным влиянием пандемии на инвестиции, поскольку многие страны, в том числе и Россия, столкнулись с экономическим спадом в этот период.

К похожему результату пришли и авторы других работ, например [4, 7]. Отсюда можно сделать вывод, что использованный алгоритм исследования показал свою целесообразность.

Предложенный подход может быть применен для дальнейшего исследования экономических последствий вспышек заболеваний, особенно актуальной на данный момент вспышки коронавируса, на фондовый рынок и крупные компании России и других стран.

Вклад авторов. *А. М. Кендысь* сформировал данные для исследования, разработал программное обеспечение, участвовал в анализе полученных результатов, подготовил текст статьи. *Н. Н. Труш* руководил проектом, анализировал полученные результаты, утвердил окончательный вариант статьи для публикации.

Список использованных источников

1. Cherubini, U. Copula Methods in Finance / U. Cherubini, E. Luciano, W. Vecchiato. – England : John Wiley & Sons, 2004. – 293 p. https://doi.org/10.1002/9781118673331

2. Щетинин, Е. Ю. О методах количественного анализа финансовых показателей компании в условиях высокой рискованности инвестиций / Е. Ю. Щетинин // Управление финансовыми рисками. – 2020. – Vol. 2, no. 2. – P. 108–119. https://doi.org/10.36627/2221-7541-2020-2-2-108-119

3. Dewick, P. R. Copula modelling to analyse financial data / P. R. Dewick, S. Liu // J. of Risk and Financial Management. – 2022. – Vol. 15, no. 3. – P. 104. https://doi.org/10.3390/jrfm15030104

4. Ghaemi, A. M. Sector-by-sector analysis of dependence dynamics between global large-cap companies and infectious diseases: A time-varying copula approach in EBOV and COVID-19 episodes / A. M. Ghaemi, H. R. Tavakkoli, M. M. Rashidi // PLOS ONE. – 2021. – Vol. 16, no. 11. – P. e0259282. https://doi.org/10.1371/journal.pone.0259282

5. Financial contagion intensity during the COVID-19 outbreak: A copula approach / R. Benkraiem [et al.] // Intern. Review of Financial Analysis. – 2022. – Vol. 81, no. 7. – P. 102136. https://doi.org/10.1016/j.irfa.2022.102136

6. Sahu, P. K. Gold price and exchange rate in pre and during Covid-19 period in India: Modelling dependence using copulas / P. K. Sahu, D. P. Bal, P. Kundu // Resources Policy. – 2022. – Vol. 79, no. 4. – P. 103126. https://doi.org/10.1016/j.resourpol.2022.103126

7. Голованов, О. А. Ретроспективный анализ влияния пандемии COVID-19 на социально-экономическое развитие региона (на примере Свердловской области) / О. А. Голованов, А. Н. Тырсин // Прикладная математика и вопросы управления. – 2023. – № 1. – С. 61–71. https://doi.org/10.15593/2499-9873/2023.1.04

8. Nelsen, R. B. An Introduction to Copulas / R. B. Nelsen. – 2nd ed. – N. Y. : Springer-Verlag, 2006. – 272 p. https://doi.org/10.1007/0-387-28678-0

9. Francq, C. GARCH Models Structure, Statistical Inference and Financial Applications / C. Francq, J. Zakoian. – England : John Wiley & Sons, 2010. – 504 p. https://doi.org/10.1002/9780470670057

10. Elements of Copula Modeling with R / M. Hofert [et al.]. – Switzerland : Springer, 2018. – 267 p. https://doi.org/10.1007/978-3-319-89635-9

References

1. Cherubini U., Luciano E., Vecchiato W. *Copula Methods in Finance*. England, John Wiley & Sons, 2004, 293 p. https://doi.org/10.1002/9781118673331

2. Shchetinin E. Yu. *On methods of quantitative analysis of the company's financial indicators under conditions of high risk of investments*. Upravlenie finansovy`mi riskami [*Financial Risk Management*], 2020, vol. 2, no. 2, pp. 108–119 (In Russ.). https://doi.org/10.36627/2221-7541-2020-2-2-108-119

3. Dewick P. R., Liu S. Copula modelling to analyse financial data. *Journal of Risk and Financial Management*, 2022, vol. 15, no. 3, p. 104. https://doi.org/10.3390/jrfm15030104

4. Ghaemi A. M., Tavakkoli H. R., Rashidi M. M. Sector-by-sector analysis of dependence dynamics between global large-cap companies and infectious diseases: A time-varying copula approach in EBOV and COVID-19 episodes. *PLOS ONE*, 2021, vol. 16, no. 11, p. e0259282. https://doi.org/10.1371/journal.pone.0259282

5. Benkraiem R., Garfatta R., Lakhal F., Zorgatid I. Financial contagion intensity during the COVID-19 outbreak: A copula approach. *International Review of Financial Analysis*, 2022, vol. 81, no. 7, p. 102136. https://doi.org/10.1016/j.irfa.2022.102136

6. Sahu P. K., Bal D. P., Kundu P. Gold price and exchange rate in pre and during Covid-19 period in India: Modelling dependence using copulas. *Resources Policy*, 2022, vol. 79, no. 4, p. 103126. https://doi.org/10.1016/j.resourpol.2022.103126

7. Golovanov O. A., Tyrsin A. N. *Retrospective analysis of the impact of the COVID-19 pandemic on the socio-economic development of the region (on the example of the Sverdlovsk region)*. Prikladnaya matematika i voprosy` upravleniya [*Applied Mathematics and Control Sciences*], 2023, no. 1, pp. 61–71 (In Russ.). https://doi.org/10.15593/2499-9873/2023.1.04

8. Nelsen R. B. *An Introduction to Copulas*, 2nd edition. New York, Springer-Verlag, 2006, 272 p. https://doi.org/10.1007/0-387-28678-0

9. Francq C., Zakoian J. *GARCH Models Structure, Statistical Inference and Financial Applications*. England, John Wiley & Sons, 2010, 504 p. https://doi.org/10.1002/9780470670057

10. Hofert M., Kojadinovic I., Mächler M., Yan J. *Elements of Copula Modeling with R*. Switzerland, Springer, 2018, 267 p. https://doi.org/10.1007/978-3-319-89635-9

Информация об авторах

Кендысь Алексей Максимович, студент факультета прикладной математики и информатики БГУ.
E-mail: kendyslesha@gmail.com
https://orcid.org/0009-0001-0779-9156

Труш Николай Николаевич, доктор физико-математических наук, профессор; профессор кафедры теории вероятностей и математической статистики факультета прикладной математики и информатики БГУ.
E-mail: troushnn@bsu.by
https://orcid.org/0000-0002-1473-0894

Information about the authors

Alexey M. Kendys, Student of the Faculty of Applied Mathematics and Computer Science, Belarusian State University.
E-mail: kendyslesha@gmail.com
https://orcid.org/0009-0001-0779-9156

Mikolai M. Troush, D. Sc. (Phys.-Math.), Prof.; Prof. at the Department of Probability Theory and Mathematical Statistics, Faculty of Applied Mathematics and Computer Science, Belarusian State University.
E-mail: troushnn@bsu.by
https://orcid.org/0000-0002-1473-0894

БИОИНФОРМАТИКА
BIOINFORMATICS

УДК 519.23
https://doi.org/10.37661/1816-0301-2024-21-2-36-53

Оригинальная статья
Original Article

Дооперационное прогнозирование Т-стадии рака желудка на базе моделей порядковой регрессии

О. В. Красько[1]✉, М. Ю. Ревтович[2], А. И. Потейко[3]

[1]*Объединенный институт проблем информатики Национальной академии наук Беларуси, ул. Сурганова, 6, Минск, 220012, Беларусь*
✉*E-mail: krasko@newman.bas-net.by*

[2]*Белорусский государственный медицинский университет, пр. Дзержинского, 83, Минск, 220083, Беларусь*
E-mail: mihail_revtovich@yahoo.com

[3]*Республиканский научно-практический центр им. Н. Н. Александрова, аг. Лесной, Минский район, 223040, Беларусь*
E-mail: drpatseika@gmail.com

Аннотация

Цели. Исследование порядковых регрессий, представленных набором бинарных логистических регрессий, и их применение в клинической практике при Т-стадировании рака желудка.

Методы. Использовались методы статистических моделей порядковой регрессии, оценки эффективности модели и анализа выживаемости.

Результаты. Основные модели порядковой регрессии были изучены и применены к клиническим данным рака желудка. К известным прогностическим критериям по классификации TNM в многофакторной регрессионной модели добавлены некоторые клинические предикторы, результаты представляются перспективными для персонализированного подхода при планировании объема лечения для повышения его эффективности.

Заключение. Проведенное исследование показало, что комплексное использование порядковых моделей наряду с мультиномиальными дает дополнительную информацию, которая помогает понять поведение латентной переменной в сложных процессах онкологических заболеваний. Клиническая часть исследования создает предпосылки к дифференцированному подходу к предоперационному планированию объема лечения пациентов с одинаковой Т-стадией на основе результатов моделирования.

Ключевые слова: порядковые регрессионные модели, метрики производительности и классификации моделей, TNM-классификация, дооперационное Т-стадирование рака желудка, анализ выживаемости

Для цитирования. Красько, О. В. Дооперационное прогнозирование Т-стадии рака желудка на базе моделей порядковой регрессии / О. В. Красько, М. Ю. Ревтович, А. И. Потейко // Информатика. – 2024. – Т. 21, № 2. – С. 36–53. https://doi.org/10.37661/1816-0301-2024-21-2-36-53

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию | Received 11.03.2024
Подписана в печать | Accepted 26.04.2024
Опубликована | Published 28.06.2024

Preoperative prediction of gastric cancer T-staging based on ordinal regression models

Olga V. Krasko[1]✉, Mikhail Yu. Reutovich[2], Aliaksandr I. Patseika[3]

[1]The United Institute of Informatics Problems of the National Academy of Sciences of Belarus, st. Surganova, 6, Minsk, 220012, Belarus
✉E-mail: krasko@newman.bas-net.by

[2]Belarusian State Medical University, av. Dzerzhinsky, 83, Minsk, 220083, Belarus
E-mail: mihail_revtovich@yahoo.com

[3]N. N. Alexandrov National Cancer Centre of Belarus, Lesnoy, Minsk District, 223040, Belarus
E-mail: drpatseika@gmail.com

Abstract

Objectives. Study of ordinal regressions presented via the set of binary logistic regressions and their application in clinical practice for T-staging of gastric cancer.

Methods. Methods of ordinal regression statistical models, model performance assessment, and survival analysis were used.

Results. Basic ordinal regression models have been studied and applied to the clinical data of gastric cancer. Some clinical predictors have been added to the well-known prognostic criteria according to the TNM classification in the multifactor regression model, results seem appropriate for a personalized approach when planning the treatment volume for improving efficacy.

Conclusion. The study showed that the analysis of ordinal models, along with multinomial ones, provides additional information that helps to understand the behavior of the latent variable in the complex cancer processes. The clinical part of the study facilitates a differentiated approach to preoperative planning of the treatment volume for patients with the same T-stage, based on modeling results.

Keywords: ordinal regression models, model performance and classifications metrics, TNM-descriptors, preoperative T-staging of gastric cancer, survival analysis

For citation. Krasko O. V., Reutovich M. Yu., Patseika A. I. *Preoperative prediction of gastric cancer T-staging based on ordinal regression models*. Informatika [*Informatics*], 2024, vol. 21, no. 2, pp. 36–53 (In Russ.). https://doi.org/10.37661/1816-0301-2024-21-2-36-53

Conflict of interest. The authors declare of no conflict of interest.

Введение. Использование регрессионных моделей в медицине связано с необходимостью анализа биомедицинских данных. Медицинские исследователи все чаще используют регрессионные модели при анализе биомедицинских данных. Особенно популярными становятся логистические регрессии. За последние годы были решены вычислительные задачи оценки этих моделей, и в настоящее время модели можно легко оценить с помощью статистических пакетов. Однако их интерпретация нуждается в пояснении. Правильная интерпретация предполагает анализ после оценки, который преобразует оцененные параметры модели в более полезную информацию о влиянии различных ковариат на зависимую переменную, измеренную в номинальной шкале.

Набор логистических регрессий может решать задачи моделирования переменных, которые измеряются в порядковой шкале. Целью настоящей статьи является краткое определение и описание порядковых регрессий, построенных на наборе бинарных логистических регрессий, демонстрация их использования в клинической практике.

1. Базовые модели для порядковых категориальных переменных. С. С. Стивенс [1] дал первоначальные определения номинальных и порядковых шкал, которые определяют возможности математической обработки переменных, измеренных в этих шкалах. Номинальные шкалы присваивают номера категориям (классам) в виде меток без какого-либо порядка, подразумеваемого числами; порядковая шкала возникает в результате операции ранжирования категорий номинальной шкалы. Поскольку любое преобразование, «сохраняющее порядок», оставляет форму шкалы неизменной, эта шкала имеет структуру того, что можно назвать изотонической или сохраняющей порядок группой [1].

Существует широкое и постоянно растущее разнообразие моделей, которые можно использовать для получения категориальных результатов и прогнозов [2]. В разд. 1 рассматриваются базовые модели с зависимой переменной, измеренной в порядковой шкале. Эти модели интересны сами по себе, а также служат основой для огромного и растущего числа альтернативных моделей, доступных для порядковых переменных (ПП).

Различают несколько базовых порядковых регрессионных моделей, в основе которых лежат предположения, связанные с явлениями и процессами окружающего мира: кумулятивную модель, которая предполагает, что в основе ПП-отклика лежит ненаблюдаемая латентная переменная, а наблюдаемая ПП рассматривается как категоризация основной непрерывной латентной переменной; последовательную модель, которая представляет собой модель латентного процесса и может моделировать динамику его развития; модель смежных категорий, которая может отражать дискретную динамику некоторого латентного процесса; модель стереотипа, которая считает, что один класс ПП считается основным состоянием изучаемого множества объектов, а остальные – степенями отклонения от основного состояния.

Если $Y \in \{1, \ldots, J\}$ означает порядковую категорию ПП, то набор бинарных переменных (БП), описывающий состояние ПП, может быть получен несколькими способами, что и приводит к нескольким вариантам бинарных моделей: кумулятивная модель пропорциональных рисков (proportional odds version of the cumulative logit model, the parallel regression model, grouped continuous model) [3–5], последовательная модель (sequential model, continuation ratio model) [6–9], модель смежных категорий (adjacent categories model) [4, 7, 9–11], модель стереотипной порядковой регрессии (stereotype model) [12, 13].

Все вышеназванные модели предлагают решения через наборы более простых моделей БП, которые образуются на базе категорий ПП. Количество бинарных моделей на единицу меньше, чем число категорий ПП, варианты сведения к набору бинарных моделей представлены в табл. 1. Функция связи бинарных моделей может быть не только логистической, но использование логит-преобразования в этих моделях ведет к понятной интерпретации поведения ковариат и их влияния на изучаемую ПП, что важно в клинических исследованиях. А. Агрести [2] отмечает, что логит-модели являются наиболее популярными для применения в медицинской статистике.

Таблица 1
Построение наборов бинарных моделей для порядковых регрессий
Table 1
Construction of binary models sets for ordinal regressions

Категории упорядоченной зависимой переменной $Y \in \{1,\ldots,J\}$ *Categories of an ordinal dependent variable* $Y \in \{1,\ldots,J\}$	Бинарные сопоставления категорий ПП *Binary comparisons of the ordered categorical variables*			
	Кумулятивная модель пропорциональных рисков *Proportional odds version of the cumulative logit model*	Последовательная модель *Continuation ratio model*	Модель смежных категорий *Adjacent category model*	Модель стереотипа *Stereotype model*
1	1 vs 2,…,*J*	1 vs 1 vs 2,…,*J*	1 vs 2	1 vs *J*
2	1,2 vs 3,…,*J*	2 vs 3,…,*J*	2 vs 3	2 vs *J*
…				
j	1,…,*j* vs *j*+1,…,*J*	*j* vs *j*+1,…,*J*	*j* vs *j*+1	*j* vs *J*
…				
J–1	1,…, *J*–1 vs *J*	*J*–1 vs *J*	*J*–1 vs *J*	*J*–1 vs *J*

1.1. Кумулятивная логит-модель с пропорциональными шансами. При предположении о ненаблюдаемой латентной переменной модель определяет совокупную вероятность того, что латентная переменная порождает категорию ПП, меньшую или равную заданной категории $Y \le j$: $\Pr(Y \le j \mid \mathbf{x}) = \sum_{r=1}^{j} \Pr(Y = r \mid \mathbf{x})$. Данная модель может быть представлена набором функций:

$$g_j(\mathbf{x}) = \log\left[\frac{\Pr(Y \le j \mid \mathbf{x})}{\Pr(Y > j \mid \mathbf{x})}\right] = \beta_{0j} - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x},\ j = 1,\ldots,J-1, \tag{1}$$

где $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_p)^{\mathrm{T}}$ – вектор коэффициентов регрессии, которые считаются одинаковыми для любой бинарной модели набора и не зависят от категории ПП (модель с параллельным уклоном). Коэффициенты смещения отвечают условию $\beta_{01} < \beta_{02} < \ldots < \beta_{0\,J-1}$. Отрицательный знак при $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$ позволяет интерпретировать положительные коэффициенты как «возрастание ковариаты вызывает возрастание категории ПП».

Кумулятивная логит-модель с пропорциональными шансами сравнивает группы категорий ПП, а условную вероятность каждой категории ПП можно получить следующим образом: поскольку $\Pr(Y \le j \mid \mathbf{x}) = 1 - \Pr(Y > j \mid \mathbf{x})$, то с учетом формулы (1) $\Pr(Y \le j \mid \mathbf{x}) = \frac{\exp(g_j(\mathbf{x}))}{1+\exp(g_j(\mathbf{x}))}$. Из этого следует равенство

$$\Pr(Y = j \mid \mathbf{x}) = \begin{cases} \Pr(Y \le j \mid \mathbf{x}),\ j = 1; \\ \Pr(Y \le j \mid \mathbf{x}) - \Pr(Y \le j-1 \mid \mathbf{x}),\ j = 2,\ldots,J-1; \\ 1 - \Pr(Y \le j-1 \mid \mathbf{x}),\ j = J. \end{cases} \tag{2}$$

Экспоненцированный коэффициент регрессии интерпретируется как отношение шансов на единицу изменения соответствующего признака, причем отношение шансов не зависит от категории ПП. Оно определяет, что влияние некоторого признака на шансы оказаться ниже заданной точки отсечения латентной переменной одинаково для всех точек отсечения, и обеспечивает тем самым экономную (в смысле степеней свободы) модель влияния признаков на ПП.

Еще одним преимуществом описываемой модели является то, что она относительно инвариантна к выбору количества категорий ответа, чего нельзя сказать о других моделях порядкового ответа [4].

1.2. Последовательная порядковая модель. Данная модель сравнивает вероятность каждой категории ПП с вероятностью более высоких категорий ПП:

$$g_j(\mathbf{x}) = \log\left[\frac{\Pr(Y = j \mid \mathbf{x})}{\Pr(Y > j \mid \mathbf{x})}\right] = \beta_{0j} - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x},\ j = 1,\ldots,J-1. \tag{3}$$

Также влияние конкретной ковариаты на ПП можно описать одним коэффициентом (и отношением шансов).

Последовательная порядковая модель сравнивает категорию с группой более высоких категорий ПП. Условная вероятность каждой категории ПП представляется как [13]

$$\Pr(Y = j \mid \mathbf{x}) = \begin{cases} \dfrac{\exp\left(g_1(\mathbf{x})\right)}{1+\exp\left(g_1(\mathbf{x})\right)},\ j = 1; \\ \dfrac{\exp\left(g_j(\mathbf{x})\right)}{\prod\limits_{r=1}^{j}\left[1+\exp\left(g_j(\mathbf{x})\right)\right]},\ j = 2,\ldots,J-1; \\ 1-\sum\limits_{r=1}^{J-1}\Pr(Y = r \mid \mathbf{x}),\ j = J. \end{cases} \tag{4}$$

Модель описывает условия выживания (или прохождения категорий по порядку) в дискретное время, поскольку выживание в момент времени j+1 является релевантным при условии, что субъект дожил до момента времени j.

1.3. Порядковая регрессионная модель смежных категорий. Данная модель представляется в виде набора бинарных моделей:

$$g_j(\mathbf{x}) = \log\left[\frac{\Pr(Y = j \mid \mathbf{x})}{\Pr(Y = j+1 \mid \mathbf{x})}\right] = \beta_{0j} - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x},\ j = 1,\ldots,J-1. \tag{5}$$

Как и в случае модели пропорциональных рисков, эффект воздействия определенной объясняющей переменной на отклик может быть описан одним коэффициентом (отношением шансов). При снятии этого ограничения данная модель фактически переходит в полиномиальную модель.

Порядковая регрессионная модель сравнивает локально две смежные категории порядкового отклика. Интерпретация экспоненцированного коэффициента регрессии остается такой же, как и в предыдущих моделях.

Условная вероятность каждого отклика определяется выражением [13]

$$\Pr(Y = j \mid \mathbf{x}) = \begin{cases} \dfrac{\exp\left(\sum\limits_{r=j}^{J-1} g_r(\mathbf{x})\right)}{1+\sum\limits_{q=1}^{J-1}\exp\left(\sum\limits_{r=q}^{J-1} g_r(\mathbf{x})\right)},\ j = 1,\ldots,J-1; \\ 1-\sum\limits_{q=1}^{J-1}\Pr(Y = q \mid \mathbf{x}),\ j = J. \end{cases} \tag{6}$$

Для категории $Y \in \{1,\ldots,J\}$ регрессионные модели сравнивают вероятность попадания в эту категорию j с вероятностью попадания в следующую по величине категорию $j+1$. Полученные модели описывают эффекты локальных отношений шансов, а не агрегированной кумулятивной вероятности в модели пропорциональных шансов. Модель смежных категорий наиболее подходит, когда интерес сосредоточен на интерпретации эффектов на уровне отдельной категории, а не на кумулятивных категориях.

Модель смежных категорий является специальным случаем базовой мультиномиальной логит-модели (см. разд. 1.5) с исходным вектором ковариат $\mathbf{x}$, умноженным на член, линейный по j:

$$g_j(\mathbf{x}) = \log\left[\frac{\Pr(Y=j\,|\,\mathbf{x})}{\Pr(Y=J\,|\,\mathbf{x})}\right] = \beta_{0j} - \boldsymbol{\beta}^{\mathrm{T}}(J-j)\mathbf{x},\ j=1,\ldots,J-1.$$

1.4. Модель стереотипной порядковой регрессии. Данная модель занимает промежуточное состояние между порядковыми регрессиями и мультиномиальной моделью. Модель описывается набором бинарных моделей:

$$g_j(\mathbf{x}) = \log\left[\frac{\Pr(Y=j\,|\,\mathbf{x})}{\Pr(Y=J\,|\,\mathbf{x})}\right] = \beta_{0j} - \varphi_j\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x},\ j=1,\ldots,J-1, \tag{7}$$

где $1 \equiv \varphi_1 > \varphi_2 > \ldots > \varphi_k \equiv 0$. Данное условие позволяет оценить и отобразить категории ПП в отрезок $[0,1]$ и выявить расположение упорядоченных категорий ПП.

Интерпретация экспоненцированного коэффициента регрессии остается такой же, как и в предыдущих моделях относительно категории J.

Условная вероятность каждой категории рассчитывается по формуле

$$\Pr(Y=j\,|\,\mathbf{x}) = \frac{\exp\left(g_j(\mathbf{x})\right)}{\sum_{r=1}^{J}\exp\left(g_r(\mathbf{x})\right)},\ j=1,\ldots,J. \tag{8}$$

1.5. Мультиномиальная модель. Первые четыре описанные модели работают с упорядоченными категориями. В мультиномиальной модели (polytomous model, multinomial model) допускается отсутствие порядка в изучаемой категориальной переменной, фактически осуществляется переход в номинальную шкалу изменений [1]. Это приводит к набору независимых моделей бинарной логистической регрессии, в которой любая категория (любая метка) может быть выбрана в качестве опорной, и всегда возможно данную опорную категорию считать последней в листе перечислений меток номинальной переменной. Тогда верно равенство

$$g_j(\mathbf{x}) = \log\left[\frac{\Pr(Y=j\,|\,\mathbf{x})}{\Pr(Y=J\,|\,\mathbf{x})}\right] = \beta_{0j} - \boldsymbol{\beta}_j{}^{\mathrm{T}}\mathbf{x},\ j=1,\ldots,J-1. \tag{9}$$

Интерпретация коэффициентов в терминах отношения шансов опирается на выбранную категорию, с которой сравниваются остальные категории мультиномиальной переменной.

Условная вероятность отклика к определенной категории определяется по формуле

$$\Pr\left(Y = j \mid \mathbf{x}\right) = \begin{cases} \dfrac{\exp\left(g_j\left(\mathbf{x}\right)\right)}{1+\sum\limits_{r=1}^{J-1}\exp\left(g_r\left(\mathbf{x}\right)\right)}, \; j = 1,\ldots,J-1; \\ \dfrac{1}{1+\sum\limits_{r=1}^{J-1}\exp\left(g_r\left(\mathbf{x}\right)\right)}, \; j = J. \end{cases} \tag{10}$$

В первых четырех моделях накладывается ограничение на параметры, которые должны быть идентичны для каждой бинарной регрессии (модели с параллельными уклонами, в которых различаются только параметры смещения β_0), в последней это ограничение снято. Тем не менее мультиномиальная модель, модель стереотипной порядковой регрессии и регрессионная модель смежных категорий связаны друг с другом [4, 14, 15].

Необходимо заметить, что описанные выше модели являются базовыми, на их основе возможно построение более сложных соотношений ковариат. В статистической литературе появляются варианты моделей гетерогенного выбора. Например, модель логистического ответа с ограничениями пропорциональности (logistic response model with proportionality constraints) [16] конструирует модель таким образом, что при оценке определяются мультипликативные скаляры, выражающие пропорциональное изменение эффекта определенных предикторов в зависимости от изменения категории. Связь между категориями и линейным предиктором определяется выражением

$$g_j\left(\mathbf{x}\right) = \log\left[\frac{\Pr\left(Y = j \mid \mathbf{x}\right)}{1-\Pr\left(Y = j \mid \mathbf{x}\right)}\right] = \beta_{0j} - \lambda_j \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}, \; j = 1,\ldots,J-1. \tag{11}$$

Также в работе [16] предложена модель, названная логистической моделью с частичными ограничениями пропорциональности (logistic response model with partial proportionality constraints). Эта модель более универсальна, чем модель (11). Связь линейного предиктора имеет вид

$$g_j\left(\mathbf{x}\right) = \log\left[\frac{\Pr\left(Y = j \mid \mathbf{x}\right)}{1-\Pr\left(Y = j \mid \mathbf{x}\right)}\right] = \beta_{0j} - \lambda_j \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + \boldsymbol{\gamma}_j{}^{\mathrm{T}}\boldsymbol{\omega}, \; j = 1,\ldots,J-1,$$

где $\mathbf{x}$ и $\boldsymbol{\omega}$ – векторы ковариат; $\boldsymbol{\beta}$ – вектор коэффициентов, которые фиксированы для всех категорий ПП; $\boldsymbol{\gamma}$ – вектор коэффициентов, которые изменяются в зависимости от категории ПП.

Таким образом, можно получить различные гибкие конструкции наборов логистических регрессий на базе исходных моделей порядковых регрессий. Большой таксономический обзор моделей ПП приведен в работе [17].

2. Анализ пригодности модели. Анализ соответствия модели данным и ее производительности – необходимый этап после подбора модели. Модель должна достаточно адекватно описывать данные, а также давать прогнозы возможной точности. Для тщательного изучения имеется множество оценок соответствия модели данным, пригодности модели при дальнейшем использовании ее в процессе прогнозирования. Существует несколько категорий оценки пригодности (производительности) моделей. Некоторые из них универсальны, некоторые представляют интерес только для задач классификации:

– традиционные тесты на пригодность;

– информационные критерии;

– анализ остатков;

– тесты семейства Хосмера – Лемешова;

– матрица классификации (матрица рассогласования).

К традиционным метрикам можно отнести статистики R^2 и псевдо-R^2, девиацию, отношения правдоподобия. Они характеризуют в первую очередь тот факт, что наличие модели позволяет снизить неопределенность в отношении зависимой переменной.

Тесты по информационным критериям, таким как AIC, BIC, HQIC, и их всевозможные модификации относятся, скорее всего, к сравнению различных моделей, построенных для одной и той же зависимой переменной, а не к внутренней непротиворечивости модели данным. Анализ остатков, наоборот, описывает отклонение данных от конкретной модели, выявление отклонений и артефактов в данных.

Хосмер и Лемешов [18] разработали ряд остатков, которые можно использовать для бинарной логистической модели, а также критерий согласия, который является одним из лучших способов оценки соответствия логистических моделей данным. Тест Хосмера – Лемешова – это статистический тест степени соответствия данным для моделей логистической регрессии. Он оценивает, соответствует ли наблюдаемая частота событий ожидаемой частоте событий, разделяя результаты прогноза на подгруппы по децилям (10 подгрупп, но возможно разделение на другое число подгрупп). Модели, для которых ожидаемая и наблюдаемая частоты событий в подгруппах схожи, считаются хорошо откалиброванными. Также предложены варианты тестов согласия специально для моделей порядковой регрессии [19, 20]. Сравнительный анализ таких тестов приведен в работе [21], для порядковых регрессий предлагается проводить три теста: тест Хосмера и Лемешова, тест Липсица [19] и тест Пулькстениса – Робинсона [20].

Матрица классификации (матрица рассогласования) – один из тестов соответствия, впервые предложенный Карлом Пирсоном в 1904 г. Он получил распространение в машинном обучении для задач классификации. Данный метод дает интерпретируемое представление о качестве прогнозирования по модели, хотя не отвечает на вопросы о качестве соответствия модели данным, возможно сравнение моделей по качеству прогноза. На основании матрицы классификации для каждой категории определяется число:

– объектов, которые правильно классифицированы моделью в данную категорию; это число истинно положительных (true positive, TP) случаев в категории;

– ложноположительных (false positive, FP) случаев в данной категории, классифицируемых моделью в данную категорию, но согласно данным они относятся к любой другой категории;

– объектов, которые правильно классифицированы моделью не в данную категорию; это число истинно отрицательных (true negative, TN) случаев в категории;

– ложноотрицательных (false negative, FN) случаев в данной категории, которые модель не классифицирует в данную категорию, но согласно данным они относятся к этой категории.

На основании данных четырех характеристик строятся метрики качества классификации: точность, чувствительность, специфичность, прогностическая ценность положительного и отрицательного результатов и др. [22]. Для клинических моделей точность, чувствительность, специфичность, прогностическая ценность положительного и отрицательного результатов имеют четкую интерпретацию и позволяют судить о прогностических возможностях модели не специалисту по машинному обучению.

3. Содержательная постановка задачи. Современная стратегия радикального лечения рака желудка (РЖ) предполагает наряду с выполнением хирургического лечения проведение адъювантной полихимиотерапии (АПХТ) или периоперационной полихимиотерапии (ППХТ), включающей в себя проведение курсов химиотерапии как до, так и после операции. Кроме того, при инфильтративных формах РЖ сообщается о необходимости дополнения стандартного объема лечебных мероприятий интраперитонеальной ПХТ (проводится во время операции) [23, 24].

Учитывая необходимость дооперационного планирования объема противоопухолевого лекарственного лечения, в частности дооперационного блока ППХТ, а также интраперитонеальной

ПХТ, проведение которой более результативно одновременно с радикальным хирургическим лечением, целесообразно уже в дооперационном периоде иметь представление о степени местной распространенности опухолевого процесса, описываемого дескрипторами классификации TNM (Tumor Node Metastasis – международная классификация стадий злокачественных новообразований).

Таким образом, может быть сформулирована следующая прикладная задача: на основе прецедентных данных предварительных дооперационных анализов и исследований прогнозировать глубину инвазии первичной опухолью стенки желудка (Т-дескриптор) на этапе дооперационного обследования для последующего планирования объема противоопухолевого лечения, т. е. провести Т-стадирование до патоморфологического исследования.

4. Описание исследуемой когорты и основные факторы прогноза. В исследование включены данные 1054 пациентов, которым в период 2008–2021 гг. было проведено радикальное лечение по поводу местно-распространенного рака желудка. В 100 % случаев диагноз был подтвержден морфологически и установлена категория Т-дескриптора.

Локализация опухоли в желудке, а также макроскопическая форма роста первичной опухоли оценивались по данным дооперационной фиброгастродуоденоскопии (ФГДС) и (или) компьютерной томографии (КТ) брюшной полости. Размеры опухолей измерены до 1 мм по данным КТ и (или) ФГДС, при наличии расхождений в размерах учитывалось значение, полученное после выполнения эндоскопического исследования. Глубина инвазии (клинический Т-дескриптор, сТ) опухоли определялась до операции ориентировочно на основании данных КТ брюшной полости по наличию признаков инвазии серозной оболочки (сТ4a) или распространения опухоли на соседние структуры (сТ4b). При отсутствии признаков, позволявших заподозрить распространенный опухолевый процесс, опухоль классифицировалась как сТ2 или сТ3. Кроме того, в случае описания эндоскопических признаков раннего РЖ опухоль стадировалась как сТ1, а для опухолей более 1 см в диаметре – как сТ2. Гистологическая структура опухоли уточнялась по результатам дооперационного морфологического исследования, в котором оценивался уровень тромбоцитов и фибриногена сыворотки до операции.

После операции проводилась патогистологическая оценка глубины инвазии первичной опухолью стенки желудка (рТ-дескриптор) на основании гистологического исследования удаленного желудка. При этом оценка категории рТ выполнялась в соответствии с классификацией опухолей по системе TNM восьмого пересмотра, согласно которой: Т1 – опухоль поражает собственную пластинку слизистой оболочки, мышечную пластинку слизистой оболочки или подслизистый слой; Т2 – опухоль поражает мышечную оболочку; Т3 – опухоль поражает субсерозный слой; Т4 – опухоль прорастает в серозную оболочку и (или) распространяется на соседние структуры [23].

На первом шаге рассмотрено соотношение клинического (дооперационного) и патогистологического (постоперационного) Т-дескрипторов (сТ и рТ соответственно). Установлено несоответствие стадии сТ, определяемой по результатам дооперационного обследования, и стадии рТ, определяемой по результатам послеоперационного морфологического исследования. Соотношение категорий сТ и рТ представлено в табл. 2 и на рис. 1.

Таблица 2
Соотношение клинического и гистологического Т-дескрипторов

Table 2
The correlation between clinical and histological T-descriptors

	pT1	pT2	pT3	pT4	Всего *Total*
cT1	51 (22,6)	2 (1,0)	0	0	53
cT2	175 (76,8)	99 (48,3)	53 (32,5)	8 (1,7)	335
cT3	2 (0,9)	103 (50,2)	100 (61,3)	236 (51,5)	441
cT4	0	1 (0,5)	10 (6,1)	214 (46,7)	225
Всего	228	205	163	458	1054

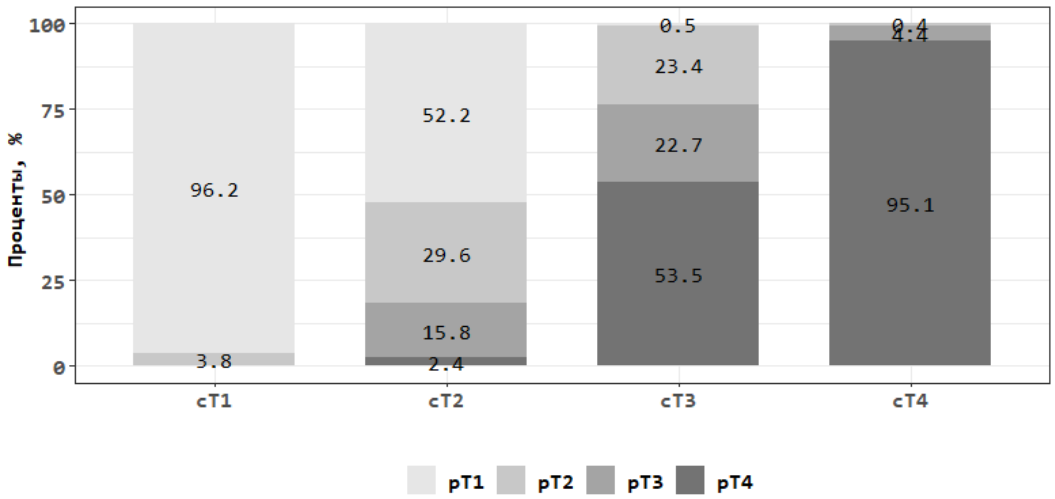


Рис. 1. Соотношение клинического и гистологического Т-стадирований

Fig. 1. The correlation between clinical and histological T-staging

Матрица рассогласования приведена в табл. 3. Каппа Коэна [23] составила 0,273, что можно трактовать как незначительное соответствие клинического сТ-дескриптора патоморфологическому pT-дескриптору. Точность прогноза pT-дескриптора только по клиническому сТ-дескриптору составила 0,44 (95 %-й доверительный интервал (ДИ) от 0,41 до 0,47).

Таблица 3

Метрики классификации гистологического Т-дескриптора на основании клинического Т-дескриптора

Table 3

Classification metrics of pT-descriptor based on cT-descriptor

Метрики классификации *Classification metrics*	pT-дескриптор *pT-descriptor*			
	pT1	pT2	pT3	pT4
Чувствительность	0,224	0,483	0,614	0,467
Специфичность	0,998	0,722	0,617	0,982
Прогностическая ценность положительного результата	0,962	0,296	0,227	0,951
Прогностическая ценность отрицательного результата	0,824	0,853	0,897	0,706

На основании проделанного сравнения клинического и гистологического Т-дескрипторов можно сделать вывод о том, что клинически сложно определить категорию Т-дескриптора. Более 50 % случаев, определенных как сТ2, оказались случаями, классифицированными в категорию pT1, и более 50 % случаев, определенных как сТ3, оказались случаями, классифицированными в категорию pT4, т. е. в большинстве случаев стадия была завышена или занижена.

Представленные выше данные согласуются с данными литературы. В частности, сообщается, что частота совпадения клинической стадии РЖ при дооперационном стадировании и патоморфологической стадии, устанавливаемой при послеоперационном морфологическом исследовании, составляет 44 %, при этом отмечается как занижение, так и завышение стадии [26]. Следствием этого является неадекватный объем лечебных мероприятий, что согласно литературным данным сопровождается снижением показателей выживаемости. В частности, авторы [26] отметили статистически значимые различия в показателях пятилетней общей выживаемости в группе верно и неверно стадированных пациентов – 71,6 % против 41,8 % ($p < 0,001$). Особенно выраженными были различия в выживаемости при неточном определении категорий T4 и N.

Предположим, что pT-дескриптор является зависимой переменной, которую можно прогнозировать на основании комплексной оценки данных дооперационного обследования пациента, включая результаты лабораторных анализов. Дополнительные данные могут быть некоторым образом связаны с категорией T-дескриптора и при правильном подборе модели снизить ошибку в дооперационном определении категории T-дескриптора. По литературным данным и результатам собственных исследований были выделены предикторы, которые приведены в табл. 4.

Таблица 4
Предварительно исследованные предикторы

Table 4
Previously studied predictors

Фактор *Factor*	Описание *Description*	Тип переменной, уровни фактора *Type of variable, factor levels*
Возраст	Возраст пациента на момент операции (включены пациенты только 18+ лет)	Количественный, положительный
Макроскопическая форма роста опухоли	Макроскопическая форма роста первичной опухоли, устанавливается при эндоскопическом исследовании и интраоперационной ревизии	Бинарный; уровни: *инфильтративная* *неинфильтративная*
cT-дескриптор опухолевого процесса	Глубина инвазии первичной опухолью стенки желудка, устанавливается на основании результатов дооперационного обследования пациента, носит ориентировочный характер	Порядковый; уровни: *cT1* *cT2* *cT3* *cT4*
cN-дескриптор опухолевого процесса	Степень метастатического поражения регионарных лимфоузлов, устанавливается на основании результатов дооперационного обследования пациента, носит ориентировочный характер	Бинарный; уровни: *cN0* *cN+*
Степень дифференцировки аденокарциномы	Гистологическая градация степени дифференцировки опухоли, определяемой при дооперационной биопсии первичной опухоли	Бинарный; уровни: *некогезивная, high grade (GIII)* *когезивная, low grade (GI-II)*
Фибриноген до операции	Белок плазмы крови, показатель гемостаза, устанавливается на основании результатов лабораторных исследований	Бинарный; уровни: *норма* *выше нормы*
Тромбоциты до операции	Один из показателей сосудисто-тромбоцитарного компонента гемостаза, устанавливается на основании результатов лабораторных исследований	Количественный, положительный

5. Результаты моделирования. Для сравнительного моделирования использовались пять основных моделей (см. разд. 1). Зависимая переменная «pT-дескриптор» содержала четыре уровня: pT1, pT2, pT3, pT4.

Различные метрики сравнительной пригодности моделей приведены в табл. 5, метрики классификации – в табл. 6. Выбраны следующие метрики производительности модели: псевдо-R^2 МакФаддена [27], остаточная девиация, логарифм правдоподобия, BIC, AIC, критерий Хосмера – Лемешова. Для оценки классификационных возможностей моделей выбраны следующие метрики: точность, каппа Коэна, чувствительность, специфичность, ценность положительного прогноза, ценность отрицательного прогноза.

Таблица 5
Метрики производительности моделей для зависимой переменной «pT-дескриптор» с четырьмя категориями

Table 5
Model performance metrics for the dependent variable "pT-descriptor" with four categories

Метрики производительности модели *Model performance metrics*	Кумулятивная модель пропорциональных рисков *Proportional odds version of the cumulative logit model*	Последовательная модель *Continuation ratio model*	Модель смежных категорий *Adjacent category model*	Модель стереотипа *Stereotype model*	Мультиномиальная модель *Multinomial model*
R^2 МакФаддена	0,431	0,428	0,432	0,447	0,459
Остаточная девиация	1559,85	1568,63	1557,53	1514,97	1483,53
Логарифм правдоподобия	–779,93	–784,31	–778,76	–757,49	–741,76
BIC	1650,34	1659,11	1648,01	1619,38	1713,22
AIC	1585,85	1594,63	1583,53	1544,97	1549,53
Критерий Хосмера – Лемешова, p-value	0,014	<0,001	0,007	0,919	0,733

Таблица 6
Метрики классификации моделей

Table 6
Model classification metrics

Классификационные метрики модели *Model classification metrics*	pT-дескриптор *pT-descriptor*			
	pT1	pT2	pT3	pT4
Порядковая кумулятивная модель				
Точность (95 % ДИ)	0,671 (0,642, 0,699)			
Каппа Коэна	0,514			
Чувствительность	0,8553	0,3902	0,0982	0,9083
Специфичность	0,9298	0,8905	0,9327	0,7718
Ценность положительного прогноза	0,7708	0,4624	0,2105	0,7536
Ценность отрицательного прогноза	0,9588	0,8581	0,8497	0,9163
Последовательная порядковая модель				
Точность (95 % ДИ)	0,663 (0,634, 0,692)			
Каппа Коэна	0,501			
Чувствительность	0,8246	0,3854	0,0797	0,9148
Специфичность	0,9383	0,8881	0,9271	0,7584
Ценность положительного прогноза	0,7866	0,4540	0,1667	0,7442
Ценность отрицательного прогноза	0,9509	0,8568	0,8463	0,9206
Порядковая регрессионная модель смежных категорий				
Точность (95 % ДИ)	0,670 (0,641, 0,698)			
Каппа Коэна	0,510			
Чувствительность	0,8684	0,3756	0,0552	0,9214
Специфичность	0,9237	0,8916	0,9517	0,7483
Ценность положительного прогноза	0,7586	0,4556	0,1731	0,7378
Ценность отрицательного прогноза	0,9622	0,8554	0,8463	0,9253
Модель стереотипной порядковой регрессии				
Точность (95 % ДИ)	0,683 (0,654, 0,711)			
Каппа Коэна	0,525			
Чувствительность	0,8816	0,4439	0,0061	0,9323
Специфичность	0,9213	0,8775	0,9899	0,7383
Ценность положительного прогноза	0,7556	0,4667	0,1000	0,7324
Ценность отрицательного прогноза	0,9657	0,8673	0,8448	0,9342
Мультиномиальная модель				
Точность (95 % ДИ)	0,695 (0,666, 0,722)			
Каппа Коэна	0,546			
Чувствительность	0,8947	0,4146	0,1411	0,9170
Специфичность	0,9189	0,8963	0,9731	0,7601
Ценность положительного прогноза	0,7528	0,4913	0,4894	0,7460
Ценность отрицательного прогноза	0,9693	0,8638	0,8610	0,9226

В результате анализа выявлено, что порядковые регрессии, в основе которых лежат предположения о латентной переменной или о некоем непрерывном процессе, показали более низкую производительность, чем модель стереотипа и мультиномиальная модель. Более того, критерий Хосмера – Лемешова показал рассогласование порядковых моделей с имеющимися данными.

Самая большая ошибка классификации – это прогноз pT-дескриптора категории pT3 для любой из моделей. Также неудовлетворительно выглядит прогноз pT2. Однако общая точность моделей статистически одинакова. Это свидетельствует о том, что часть данных не может быть однозначно классифицирована в рамках предложенных четырех последовательных категорий pT-дескриптора согласно TNM-классификации.

Модель стереотипной порядковой регрессии, которая занимает промежуточное состояние между порядковыми регрессиями и мультиномиальной регрессионной моделью, показала, как изменяется параметр φ для каждой категории (рис. 3). Прогнозная категория pT3 отклоняется от линейного изменения при изменении ПП, что и отражается в классификационных метриках моделей. Видимо, в рамках предложенных предикторов четкая классификация категории pT3 крайне затруднительна.

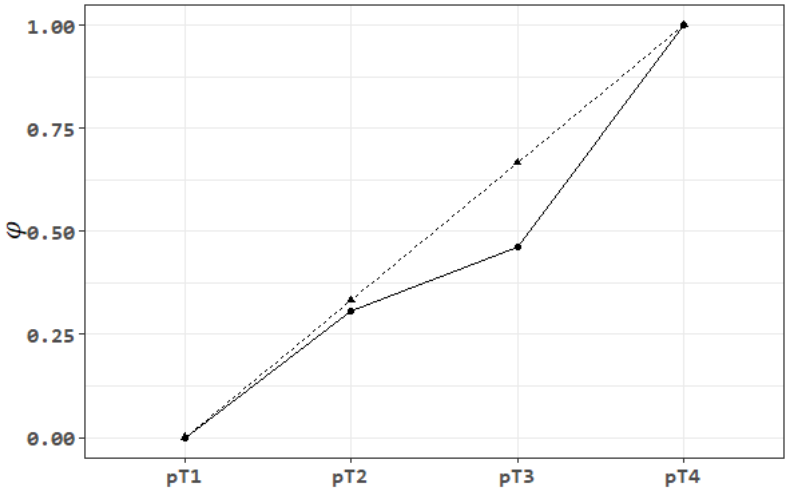


Рис. 3. Изменение параметра φ в модели стереотипа в зависимости от числового значения T-дескриптора

Fig. 3. Changing the φ parameter in a stereotype model depending on the T-descriptor category

Для понимания того, как предикторы различают категории pT-дескриптора, построена диаграмма изменения коэффициентов мультиномиальной модели, которая не связана с ограничениями моделей ПП (рис. 4).

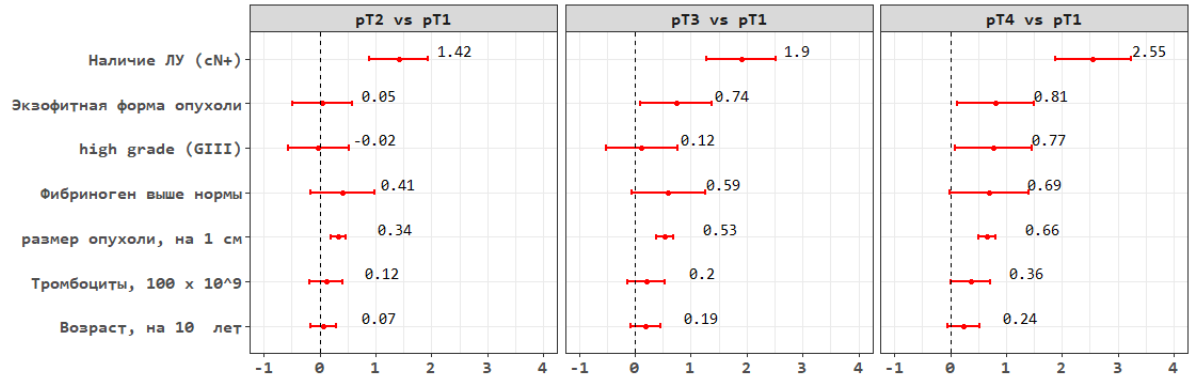


Рис. 4. Коэффициенты предикторов логистических регрессий, составляющих мультиномиальную модель

Fig. 4. Predictors coefficients of logistic regressions included in the multinomial model

На диаграмме видно, что некоторые предикторы, например размер опухоли, статистически значимы для любой категории, в то время как ряд других значимы только при более распространенном опухолевом процессе – pT3-4. Это свидетельствует о сложности и неоднородности процессов, лежащих в основе инвазии первичной опухолью стенки желудка, и необходимости комплексного учета нескольких признаков как при дооперационном Т-стадировании, так и при оценке возможных вариантов клинического течения РЖ на основании прогнозируемой местной распространенности опухолевого процесса (т. е. Т-стадирования РЖ).

На представленных предикторах невозможно с достаточной точностью классифицировать pT2- и pT3-дескрипторы. Для дальнейшего анализа категории pT2 и pT3 были объединены в одну категорию pT2-3. Соответственно, клинические категории cT2 и cT3 были также объединены в категорию cT2-3. Последнее логически обосновано и связано с ориентировочным характером разграничения Т2 и Т3 на этапе дооперационного обследования. При этом нетрудно представить, что это невозможно в принципе, поскольку инвазия в мышечную оболочку желудка или в субсерозный слой может быть определена и подтверждена только после постоперационного гистологического исследования удаленного желудка. Иными словами, разграничение категорий Т2 и Т3 на дооперационном этапе носит, как правило, субъективный характер за исключением случаев оценки глубины инвазии при использовании эндоультрасонографии. Хотя и в последнем случае также сообщается о высокой частоте неверного Т-стадирования [28].

Краткие результаты моделирования переменной «pT-дескриптор» с тремя категориями приведены в табл. 7.

Таблица 7
Метрики производительности и классификации моделей для зависимой переменной «pT-дескриптор» с тремя категориями

Table 7
Model performance and classification metrics for the dependent variable "pT-descriptor" with three categories

Метрики производительности модели *Model performance metrics*	Кумулятивная модель пропорциональных рисков *Proportional odds version of the cumulative logit model*	Последовательная модель *Continuation ratio model*	Модель смежных категорий *Adjacent category model*	Модель стереотипа *Stereotype model*	Мультиномиальная модель *Multinomial model*
R^2 МакФаддена	0,375	0,374	0,381	0,392	0,399
Остаточная девиация	1396,91	1400,38	1383,39	1359,82	1343,69
Логарифм правдоподобия	–698,46	–700,19	–691,70	–679,91	–671,84
BIC	1473,47	1476,94	1459,95	1443,35	1482,90
AIC	1418,91	1422,38	1405,39	1383,83	1383,69
Критерий Хосмера – Лемешова, p-value	0,006	<0,001	0,014	0,52	0,331
Точность (95 % ДИ)	0,689 (0,660, 0,717)	0,687 (0,658, 0,715)	0,687 (0,658, 0,715)	0,691 (0,662, 0,718)	0,700 (0,672, 0,728)
Каппа Коэна	0,514	0,508	0,510	0,522	0,535

Точность классификации и каппа Коэна в моделях с тремя категориями pT-дескриптора остались на уровне предыдущих моделей с четырьмя категориями pT-дескриптора, что свидетельствует об избыточности четырех категорий в рамках предложенного набора предикторов (см. табл. 4).

На диаграмме (рис. 5) показано, сколько процентов случаев в каждой категории зависимой переменной pT классифицированы корректно и ошибочно в мультиномиальной модели.

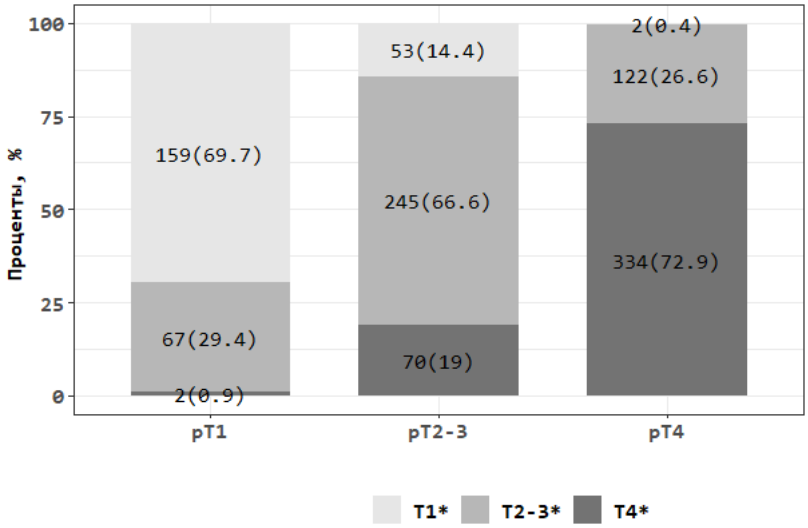


Рис. 5. Рассогласование гистологического cT-дескриптора и прогноза на базе мультиномиальной модели

Fig. 5. Misalignment between histological cT-descriptor and predicted T-descriptor based on the multinomial model

При дальнейших исследованиях данных по РЖ проводится дополнительная клиническая валидация моделей. Известно, что T-стадирование связано с показателями выживаемости, и авторы связали прогнозные категории T-дескриптора с выживаемостью пациентов после проведенного лечения.

Для каждой категории pT-дескриптора были построены кривые выживаемости в зависимости от прогнозных категорий T-дескриптора, полученных по модели. Так, для пациентов с pT1-дескриптором были выделены группы «прогноз T1» и «прогноз T2-4»; для пациентов с pT2-3-дескриптором – группы «прогноз T1», «прогноз T2-3» и «прогноз T4»; для группы с pT4-дескриптором – группы «прогноз T1-3» и «прогноз T4». Расчеты показали, что в рамках выделенных категорий прогнозные значения T-дескриптора по модели дают дополнительную информацию о выживаемости пациентов с РЖ. На рис. 6 показаны кривые выживаемости по каждой категории для мультиномиальной модели.

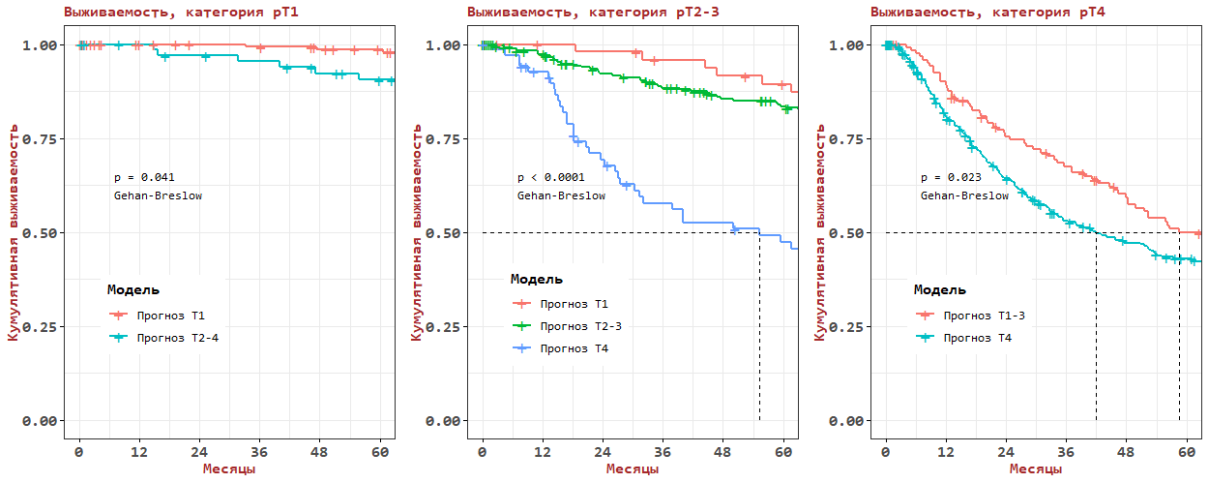


Рис. 6. Кумулятивная выживаемость по категориям cT с учетом модельного прогноза T-дескриптора для мультиномиальной модели

Fig. 6. Cumulative survival in cT categories grouped by predicted T-descriptors based on the multinomial model

С увеличением глубины инвазии первичной опухолью стенки желудка (от pT1 до pT4) пациенты ожидаемо продемонстрировали ухудшение показателей выживаемости, что может объясняться усилением процессов лимфогенного метастазирования и увеличением количества метастатически пораженных регионарных лимфоузлов [29]. Однако согласно данным литературы на практике прямой зависимости между глубиной инвазии стенки желудка и частотой прогрессирования опухолевого процесса после проведенного радикального хирургического лечения не существует. Ряд исследователей обращают внимание на то, что при одинаковой стадии зачастую наблюдается различное клиническое течение РЖ [31]. Этот факт свидетельствует о том, что помимо общеизвестных прогностических критериев согласно классификации TNM существуют и другие, выяснение роли которых в прогнозировании клинического течения РЖ нуждается в уточнении. Представленные на рис. 6 результаты являются подтверждением вышесказанному: учет дополнительных критериев при оценке клинического течения РЖ позволяет выделить в пределах каждой из когорт pT1, pT2-3, pT4 подгруппы со стандартным (т. е. соответствующим по клиническому течению рассматриваемой Т-стадии) и с неблагоприятным клиническим течением (т. е. соответствующим более распространенному опухолевому процессу с большей глубиной инвазии стенки желудка).

Таким образом, комплексный учет проанализированных в статье характеристик опухолевого процесса наряду с оценкой глубины инвазии первичной опухолью стенки при оценке прогноза клинического течения РЖ определил прогностическую неоднородность внутри каждой из когорт пациентов с идентичной глубиной инвазии стенки желудка (pT1, pT2-3, pT4), выделив (внутри каждой из упомянутых выше когорт) группы с различным клиническим течением РЖ, несмотря на радикальный характер выполненной операции и идентичное патоморфологическое Т-стадирование. Как следует из информации, представленной на рис. 6, комплексная оценка на предоперационном этапе ряда признаков, а именно возраста пациента, макроскопической формы роста первичной опухоли, ее гистологической структуры, размера, наряду с уровнем тромбоцитов и фибриногена до операции повышает точность традиционно используемой в онкологии классификации TNM при оценке клинического течения РЖ в отдаленные сроки после радикальной операции. Ранее рядом исследователей было отмечено, что при одинаковой TNM-стадии зачастую после радикального хирургического лечения наблюдаются различные клинические течения РЖ [30, 31]. Последнее лишний раз подчеркивает актуальность полученных при моделировании результатов.

Проведенные исследования позволили сделать следующие выводы:

1. Повышение точности дооперационного Т-стадирования возможно при комплексном учете факторов дооперационного обследования с использованием моделирования, которое позволило повысить точность прогноза глубины инвазии первичной опухолью стенки желудка (cT) в сравнении с традиционно используемым подходом, а также определить прогностическую неоднородность внутри каждой из когорт пациентов с идентичной глубиной инвазии стенки желудка (pT1, pT2-3, pT4).

2. Представляется актуальным дальнейший поиск маркеров, учет значений которых в рамках прогностических моделей может помочь в повышении точности определения дооперационного Т-дескриптора, а также в дифференцированном подходе к объему противоопухолевого лечения с учетом различных прогнозов клинического течения РЖ у пациентов с идентичной глубиной инвазии первичной опухолью его стенки.

3. В работе [15, с. 385] автор пишет: «Можно использовать упорядоченные уровни для полиномиальной модели, но информация, содержащаяся в порядке, теряется для понимания ответа» (One may use ordered levels for a multinomial model, but the information entailed in the order is lost to the understanding of the response). На основании проведенного анализа можно сказать, что сравнение результатов порядковых моделей и полиномиального аналога дает дополнительную информацию для понимания процессов, лежащих в основе данных реального мира.

4. Поведение латентной переменной, лежащей в основе процесса инвазии стенки желудка, обусловливается комплексными нарушениями гомеостаза организма, лежащими в основе прогрессирования опухолевого процесса, начиная с этапного увеличения размера первичной опу-

холи от инициальных неопластических изменений, описываемых как рак in situ, и заканчивая тотальной инвазией всех слоев стенки желудка.

Заключение. В настоящей работе продемонстрирован комплексный подход к моделированию ПП в прикладных онкологических задачах. В конечном счете принятие решения всегда остается за специалистом-онкологом, но современные математические модели позволяют уточнять комплексное влияние параметров организма и опухолевого процесса на исход заболевания и своевременно корректировать и оптимизировать противоопухолевое лечение.

Результаты представленного выше исследования наряду с данными литературы свидетельствуют о том, что помимо общеизвестных прогностических критериев согласно классификации TNM существуют и другие, выяснение роли которых в прогрессировании РЖ представляется целесообразным для персонифицированного подхода при планировании объема противоопухолевого лечения, направленного на повышение его результативности.

Проведенное исследование создает предпосылки к дифференцированному подходу к дооперационному планированию объема лечебных мероприятий у пациентов с одинаковой глубиной распространения опухоли в пределах стенки желудка в зависимости от прогнозируемых показателей выживаемости.

Вклад авторов. *О. В. Красько* – концепция и моделирование исследования, анализ данных, написание текста, редактирование; *М. Ю. Ревтович* – дизайн исследования, сбор материала, обработка, написание текста, редактирование; *А. И. Потейко* – обработка, написание текста.

References

1. Stevens S. S. On the theory of scales of measurement. *Science*, 1946, vol. 103, no. 2684, pp. 677–680.
2. Agresti A. Modelling ordered categorical data: recent advances and future challenges. *Statistics in Medicine*, 1999, vol. 18, no. 17–18, pp. 2191–2207.
3. McCullagh P. Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1980, vol. 42, no. 2, pp. 109–127.
4. Agresti A. *Analysis of Ordinal Categorical Data*, 2nd edition. John Wiley & Sons, 2010, 424 p.
5. McCullagh P. Proportional odds model: Theoretical background. *Wiley StatsRef: Statistics Reference Online*, 2014. Available at: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat05796 (accessed 27.02.2024).
6. Thompson Jr W. A. On the treatment of grouped observations in life studies. *Biometrics*, 1977, vol. 33, no. 3, pp. 463–470.
7. Cox D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1972, vol. 34, no. 2, pp. 187–202.
8. McCullagh P. Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1980, vol. 42, no. 2, pp. 109–127.
9. Fienberg S. E. *The Analysis of Cross-Classified Categorical Data*, 2nd edition. Springer Science & Business Media, 2007, 212 p.
10. Powers D., Xie Y. *Statistical Methods for Categorical Data Analysis*, 2nd edition. Emerald Group Publishing, 2008, 296 p.
11. Fullerton A. S., Xu J. Constrained and unconstrained partial adjacent category logit models for ordinal response variables. *Sociological Methods & Research*, 2018, vol. 47, no. 2, pp. 169–206.
12. Anderson J. A. Regression and ordered categorical variables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1984, vol. 46, no. 1, pp. 1–22.
13. Fernandez D., Liu I., Costilla R. A method for ordinal outcomes: The ordered stereotype model. *International Journal of Methods in Psychiatric Research*, 2019, vol. 28, no. 4, p. e1801.
14. Long J. S., Cheng S. Regression models for categorical outcomes. *Handbook of Data Analysis*, 2004, pp. 259–284.
15. Hilbe J. M. *Logistic Regression Models*, 1st edition. CRC press, 2009, 656 p.
16. Hauser R. M., Andrew M. 1. Another look at the stratification of educational transitions: the logistic response model with partial proportionality constraints. *Sociological Methodology*, 2006, vol. 36, no. 1, pp. 1–26.
17. Tutz G. Ordinal regression: A review arn a taxoromy of models. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2022, vol. 14, no. 2, p. e1545.

18. Hosmer D. W., Lemeshow S., Cook E. *Applied Logistic Regression*, 2nd edition. New York, John Wiley and Sons Inc., 2000, 376 p.

19. Lipsitz S. R., Fitzmaurice G. M., Molenberghs G. Goodness-of-fit tests for ordinal response regression models. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 1996, vol. 45, no. 2, pp. 175–190.

20. Pulkstenis E., Robinson T. J. Goodness-of-fit tests for ordinal response regression models. *Statistics in Medicine*, 2004, vol. 23, no. 6, pp. 999–1014.

21. Fagerland M. W., Hosmer D. W. Tests for goodness of fit in ordinal logistic regression models. *Journal of Statistical Computation and Simulation*, 2016, vol. 86, no. 17, pp. 3398–3418.

22. Tharwat A. Classification assessment methods. *Applied Computing and Informatics*, 2020, vol. 17, no. 1, pp. 168–192.

23. Japanese Gastric Cancer Association. Japanese Gastric Cancer Treatment Guidelines 2021. *Gastric Cancer: Official Journal of the International Gastric Cancer Association and the Japanese Gastric Cancer Association*, 2023, vol. 26, no. 1, pp. 1–25.

24. Götze T. O., Piso P., Lorenzen S., Bankstahl U. S., Pauligk C., …, Königsrainer A. Preventive HIPEC in combination with perioperative FLOT versus FLOT alone for resectable diffuse type gastric and gastroesophageal junction type II/III adenocarcinoma – the phase III "PREVENT"-(FLOT9) trial of the AIO/CAOGI/ACO. *BMC Cancer*, 2021, vol. 21, no. 1, pp. 1–9.

25. Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 1960, vol. 20, no. 1, pp. 37–46.

26. Lin M., Chen Q.-Y., Zheng C.-H., Li P., Xie J.-W., …, Huang C.-M. Effect of preoperative tumor under-staging on the long-term survival of patients undergoing radical gastrectomy for gastric cancer. *Cancer Research and Treatment: Official Journal of Korean Cancer Association*, 2021, vol. 53, no. 4, pp. 1123–1133.

27. Mcfadden D. Conditional logit analysis of qualitative choice behavior. *Frontier in Econometrics*, New York, Academic press, 1973, pp. 105–142.

28. Rosa F., Costamagna G., Doglietto G. B., Alfieri S. Classification of nodal stations in gastric cancer. *Translational Gastroenterology and Hepatology*, 2017, vol. 2. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5313280/pdf/tgh-02-2016.12.03.pdf (accessed 27.02.2024).

29. Deng J., Zhang R., Pan Y., Ding X., Cai M., …, Liang H. Tumor size as a recommendable variable for accuracy of the prognostic prediction of gastric cancer: a retrospective analysis of 1,521 patients. *Annals of Surgical Oncology*, 2015, vol. 22, pp. 565–572.

30. Dai N., Lu A.-P., Shou C.-C., Li J.-Y. Expression of phosphatase regenerating liver 3 is an independent prognostic indicator for gastric cancer. *World Journal of Gastroenterology*, 2009, vol. 15, no. 12, pp. 1499–1505.

31. Kim E. Y. Lee J. W., Yoo H. M., Park C. H., Song K. Y. The platelet-to-lymphocyte ratio versus neutrophil-to-lymphocyte ratio: which is better as a prognostic factor in gastric cancer? *Annals of Surgical Oncology*, 2015, vol. 22, pp. 4363–4370.

Информация об авторах

Красько Ольга Владимировна, кандидат технических наук, доцент, ведущий научный сотрудник, Объединенный институт проблем информатики Национальной академии наук Беларуси.
E-mail: krasko@newman.bas-net.by
https://orcid.org/0000-0002-4150-282X

Ревтович Михаил Юрьевич, доктор медицинских наук, доцент, декан лечебного факультета, Белорусский государственный медицинский университет.
E-mail: mihail_revtovich@yahoo.com
https://orcid.org/0000-0001-7202-6902

Потейко Александр Иванович, врач онколог-хирург, онкологическое отделение гастроэзофагеальной патологии, Республиканский научно-практический центр онкологии и медицинской радиологии им. Н. Н. Александрова.
E-mail: drpatseika@gmail.com
https://orcid.org/0009-0000-7271-3913

Information about the authors

Olga V. Krasko, Ph. D. (Eng.), Assoc. Prof., Leading Researcher, The United Institute of Informatics Problems of the National Academy of Sciences of Belarus.
E-mail: krasko@newman.bas-net.by
https://orcid.org/0000-0002-4150-282X

Mikhail Yu. Reutovich, D. Sc. (Med.), Assoc. Prof., Dean of the Faculty of General Medicine Belarusian State Medical University.
E-mail: mihail_revtovich@yahoo.com
https://orcid.org/0000-0001-7202-6902

Aliaksandr I. Patseika, Surgical Oncologist, Oncology Division of Gastroesophageal Abnormalities, N. N. Alexandrov National Cancer Centre of Belarus.
E-mail: drpatseika@gmail.com
https://orcid.org/0009-0000-7271-3913

ЛОГИЧЕСКОЕ ПРОЕКТИРОВАНИЕ
LOGICAL DESIGN

УДК 004.33.054
https://doi.org/10.37661/1816-0301-2024-21-2-54-72

Оригинальная статья
Original Article

Меры различия, основанные на применении расстояния Хэмминга, для генерирования управляемых вероятностных тестов

В. Н. Ярмолик[1]✉, В. В. Петровская[1], Н. А. Шевченко[2]

[1]*Белорусский государственный университет информатики и радиоэлектроники, ул. П. Бровки, 6, Минск, 220013, Беларусь*
✉*E-mail: yarmolik10ru@yahoo.com*

[2]*Дармштадтский технический университет, Каролиненплац, 5, Дармштадт, 64289, Германия*

Аннотация

Цели. Решается задача построения мер различия, основанных на применении расстояния Хэмминга, для генерирования управляемых вероятностных двоичных тестовых наборов. Целью настоящей статьи является развитие методов определения расстояния Хэмминга для нахождения различия между тестовыми наборами при их совпадении по оценкам других мер различия.

Методы. На базе расстояния Хэмминга, используемого в теории и практике формирования управляемых вероятностных тестов, предлагаются новые меры различия для сравнения двух двоичных n-разрядных тестовых наборов. Основой предлагаемых мер различия является формирование множества расстояний Хэмминга для исходных наборов, представляемых в виде последовательностей символов различных алфавитов.

Результаты. Показывается неразличимость пар двоичных тестовых наборов при использовании меры различия, основанной на применении расстояния Хэмминга. В этом случае отличающиеся пары наборов могут иметь совпадающие значения расстояния Хэмминга. Для построения новых мер различия исходные двоичные тестовые наборы представляются в виде последовательностей, состоящих из символов, принадлежащих различным алфавитам. Предлагаются различные стратегии применения новых мер различия, основанных на использовании одного из трех правил, при генерировании управляемых вероятностных тестов. Показано, что во всех трех случаях новых мер различия информативными являются только несколько первых их компонент, как правило, не более двух или трех. Соответственно, вычислительная сложность для всех трех вариантов сравнима и не превышает $3n$ операций сравнения. Проведенные экспериментальные исследования подтверждают эффективность предложенных мер различия и их невысокую вычислительную сложность.

Заключение. Предложенные меры различия расширяют возможности генерирования тестовых наборов при формировании управляемых вероятностных тестов. Показывается, что тестовые наборы, неразличимые при использовании в качестве меры различия расстояния Хэмминга, имеют отличающиеся значения предложенных мер различия. Это позволяет более точно классифицировать формируемые случайным образом наборы, которые являются кандидатами в тестовые наборы.

Ключевые слова: тестирование вычислительных систем, управляемые вероятностные тесты, двоичный тестовый набор, мера различия символьных наборов, расстояние Хэмминга

Для цитирования. Ярмолик, В. Н. Меры различия, основанные на применении расстояния Хэмминга, для генерирования управляемых вероятностных тестов / В. Н. Ярмолик, В. В. Петровская, Н. А. Шевченко // Информатика. – 2024. − Т. 21, № 2. – С. 54–72.
https://doi.org/10.37661/1816-0301-2024-21-2-54-72

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию | Received 25.03.2024
Подписана в печать | Accepted 15.04.2024
Опубликована | Published 28.06.2024

Dissimilarity measures based on the application of Hamming distance to generate controlled probabilistic tests

Vyacheslav N. Yarmolik[1]✉, Vita V. Petrovskaya[1], Nikolai A. Shevchenko[2]

[1]Belarusian State University of Informatics and Radioelectronics,
st. P. Brovki, 6, Minsk, 220013, Belarus
✉*E-mail: yarmolik10ru@yahoo.com*

[2]Technical University of Darmstadt,
Karolinenplatz, 5, Darmstadt, 64289, Germany

Abstract.
Objectives. The problem of constructing dissimilarity measures based on the application of the Hamming distance to generate controlled random binary test sets is solved. The main goal of this article is to develop methods for determining the Hamming distance for the achievability of finding the difference between test sets when they coincide according to estimates of other difference measures.
Methods. Based on the Hamming distance used in the theory and practice of generating controlled random tests, new dissimilarity measures are proposed for two binary test *n*-bit patterns. The basis of the proposed dissimilarity measures is the formation of sets of Hamming distances for initial sets, represented as sequences of characters from different alphabets.
Results. The indistinguishability of pairs of binary test sets T_i and T_k is shown using a dissimilarity measure based on the application of the Hamming distance. In this case, different pairs of sets may have identical Hamming distance values. To construct new measures of difference, the original binary test sequences are represented as sequences consisting of characters belonging to different alphabets. Various strategies are proposed for applying new measures of difference based on the use of one of three rules in generating controlled probability tests. It is shown that in all three cases of dissimilarity measures, only the first few of their components areinformative, as a rule, no more than two or three. Accordingly, the computational complexity for all three options is comparable and does not exceed 3*n* comparison operations. The experimental studies carried out confirm the effectiveness of the proposed dissimilarity measures and their low computational complexity.
Conclusion. The proposed dissimilarity measures expand the possibilities of generating test sets when forming controlled random tests. It is shown that test sets that are indistinguishable when using the Hamming distance as a dissimilarity measure have different values of the proposed dissimilarity measures, which makes it possible to more accurately classify randomly generated sets that are candidate test cases.

Keywords: computer systems testing, controlled random tests, binary test patterns, character patterns difference measure, Hamming distance

For citation. Yarmolik V. N., Petrovskaya V. V., Shevchenko N. A. *Dissimilarity measures based on the application of Hamming distance to generate controlled probabilistic tests*. Informatika [*Informatics*], 2024, vol. 21, no. 2, pp. 54−72 (In Russ.). https://doi.org/10.37661/1816-0301-2024-21-2-54-72

Conflict of interest. The authors declare of no conflict of interest.

Введение. Традиционно тестирование современных вычислительных систем заключается в использовании тестовых наборов, сформированных случайным образом из множества всех возможных входных данных. Подобная процедура тестирования систем называется вероятностным тестированием (Random Testing) [1–3]. Вероятностное тестирование является основополагающей технологией тестирования, основанной на методе черного ящика, которая, как правило, не учитывает особенности тестируемого объекта [3–5]. Вероятностный тест задается количеством q тестовых наборов $T_i = t_{i,0}, t_{i,1}, ..., t_{i,n-1}$, $i \in \{0, 1, \ldots, q-1\}$, данных $t_{i,j}$, $j \in \{0, 1, \ldots, n-1\}$, из заданного алфавита данных и их числом n в наборе. С целью увеличения эффективности вероятностных тестов используются их модификации, так называемые управляемые вероятностные тесты (Controlled Random Tests) [5–8]. Множество управляемых вероятностных тестов включает различные их разновидности, которые в англоязычной литературе встречаются под названиями быстрое антивероятностное тестирование (Fast Antirandom Testing, FAR), адаптивное вероятностное тестирование (Adaptive Random Testing), эволюционное вероятностное тестирование (Evolutionary Random Testing), эффективное вероятностное тестирование (Good Random Testing), ограниченное вероятностное тестирование (Restricted Random Testing), зеркальное вероятностное тестирование (Mirror Random Testing), упорядоченное вероятностное тестирование (Orderly Random Testing), гибридное адаптивное вероятностное тестирование (Hybrid Adaptive Random Testing), расширенное адаптивное вероятностное тестирование (Enhanced Adaptive Random Testing) и др. [8–10].

Процедура управляемого генерирования вероятностных тестовых наборов основывается на информации, которая извлекается в виде некоторых характеристик (мер) из ранее сгенерированных тестовых наборов и используется для формирования следующего набора [6]. Очередной тестовый набор T_i управляемого вероятностного теста формируется максимально удаленным (различным) от ранее сгенерированных наборов T_0, T_1,..., T_{i-1} в терминах заранее выбранных мер различия. Таким образом, принимается гипотеза, что для двух тестовых наборов T_i и T_k, имеющих максимальное отличие, количество обнаруживаемых неисправностей будет максимальным [5–10]. Главная проблема управляемого вероятностного тестирования состоит в нахождении численных значений мер различия для тестовых наборов T_i и T_k. Вычисление мер различия для символьных тестовых последовательностей, в свою очередь, сводится к задаче их сравнения [11].

На сегодня широко известны следующие меры сравнения для символьных последовательностей: расстояние Хэмминга [12], расстояние Левенштейна [13], расстояние Дамерау – Левенштейна [14], сходство Джаро – Винклера [15], метрика Нидлмана – Вунша [16], метрика Смита – Вотермана [17] и др. Отметим, что все известные авторам алгоритмы, применяемые для решения задачи определения расстояния между последовательностями символов, т. е. их отличия, так или иначе основаны именно на этих метриках. Расстояние Хэмминга является одной из универсальных мер близости последовательностей символов одинаковой размерности.

При формировании очередного тестового набора первоначально генерируется фиксированное количество кандидатов в тесты, представляющих собой, как правило, равномерно распределенные случайные наборы [6, 10, 18]. Для каждого из них вычисляются метрики различия, с учетом которых и выбирается наилучший из кандидатов в тесты в качестве очередного тестового набора T_i [6, 18–20].

Главная задача управляемого вероятностного тестирования состоит в нахождении меры различия для тестовых наборов T_i и T_k, которая максимально адекватно показывает их отличие и характеризуется невысокой вычислительной сложностью [5, 6, 19]. Вычисление мер различия тестовых наборов, в общем случае представляющих собой символьные последовательности, в свою очередь, сводится к задаче их сравнения. Большинство известных подходов генерирования управляемых вероятностных тестов основано на применении расстояния Хэмминга [4–8]. Это позволяет значительно уменьшить вычислительную сложность формирования тестов, что достигается за счет использования достаточно общей и зачастую неточной меры различия. Более точные меры различия символьных последовательностей имеют существенно большую вычислительную сложность [19–21].

Результаты, представленные в настоящей работе, направлены на решение задачи поиска новых эффективных мер различия тестовых наборов при формировании управляемых вероятностных тестов. Основываясь на применении в качестве меры различия расстояния Хэмминга, авторы предлагают ряд модификаций его определения в качестве мер различия, используемых для генерирования управляемых вероятностных тестов.

1. Мера различия для построения управляемых вероятностных тестов. Как уже отмечалось, формально мера различия последовательностей символов существует – это расстояние Хэмминга $D(T_i, T_k)$. Расстояние Хэмминга между последовательностями $T_i = t_{i,0}, t_{i,1}, ..., t_{i,n-1}$ и $T_k = t_{k,0}, t_{k,1}, ..., t_{k,n-1}$, включающими по n символов, равняется числу их несовпадающих значений $t_{i,j}$ и $t_{k,j}$ [11, 12]:

$$D(T_i, T_k) = \sum_{j=0}^{n-1} \mathrm{I}(t_{i,j} \neq t_{k,j}). \quad (1)$$

Выражение $\mathrm{I}(t_{i,j} \neq t_{k,j})$ представляет собой индикаторную функцию, равную единице при $t_{i,j} \neq t_{k,j}$ и нулю в противном случае. Минимальное значение $\min D(T_i, T_k)$ равняется нулю при совпадении всех символов последовательностей T_i и T_k, а максимальное $\max D(T_i, T_k) = n$ – при несовпадении всех n символов.

Расстояние Хэмминга как метрика малоэффективно, так как позволяет различать лишь полностью совпадающие последовательности при $D(T_i, T_k) = 0$ и все остальные несовпадающие [11]. Аргументом для подтверждения неразличимости несовпадающих последовательностей являются наборы двоичных символов T_i и $T_k = \overline{T_i}$, расстояние Хэмминга $D(T_i, \overline{T_i})$ для которых всегда неизменно и равняется n, например D(11111110, 00000001) = D(10101000, 01010111) = D(11001100, 00110011) = 8. Видно, что расстояние Хэмминга $D(T_i, T_k)$ во всех рассмотренных выше примерах равняется восьми. Это свидетельствует об одинаковом максимальном отличии T_i от $\overline{T_i}$ во всех рассмотренных парах наборов, хотя структуры пар последовательностей существенно отличаются. Еще большим отличием характеризуются пары последовательностей символов T_i, T_k: 00000000, 11110000; 11111111, 00001111 и 01011010, 11010100, для каждой из которых $D(T_i, T_k) = 4$. Приведенные примеры показывают необходимость использования новых более эффективных мер сравнения последовательностей символов, позволяющих более полно оценивать схожесть их структур.

Исследуем расширение возможностей применения расстояния Хэмминга для сравнения конечных последовательностей символов $T_i = t_{i,0}, t_{i,1}, ..., t_{i,n-1}$ и $T_k = t_{k,0}, t_{k,1}, ..., t_{k,n-1}$ как объектов, представляющих упорядоченные тестовые наборы T_i и T_k и состоящих из n символов (элементов) $t_{i,j}$ и $t_{k,j}$, где $j \in \{0, 1, \ldots, n-1\}$. Алфавит символов $t_{i,j}$ и $t_{k,j}$ может быть произвольным, так же как и их количество n в наборах T_i и T_k.

Не нарушая общности дальнейшего изложения, предположим, что тестовый набор T_i является двоичным, т. е. символы $t_{i,j} \in \{0, 1\}$ для $j \in \{0, 1, \ldots, n-1\}$, а $n = 2^w$, где w – целое. Отметим, что приведенные ограничения для T_i достаточно часто выполняются на практике при решении задач тестового диагностирования современных вычислительных систем. Такие ограничения позволяют сформировать $w + 1$ отображений исходной двоичной последовательности $T_i = t_{i,0}, t_{i,1}, ..., t_{i,n-1}$ в виде $T_i(0) = t_{i,0}(0), t_{i,1}(0), ..., t_{i,n-1}(0)$; $T_i(1) = t_{i,0}(1), t_{i,1}(1), ..., t_{i,n/2-1}(1)$; $T_i(2) = t_{i,0}(2), t_{i,1}(2), ..., t_{i,n/4-1}(2)$; $T_i(3) = t_{i,0}(3), t_{i,1}(3), ..., t_{i,n/8-1}(3)$; …; $T_i(w-1) = t_{i,0}(w-1), t_{i,1}(w-1)$ и $T_i(w) = t_{i,0}(w) = T_i$. В общем случае $T_i(v)$ состоит из 2^{w-v} символов, принадлежащих алфавиту, включающему (2^{2^v}) символов, каждый из которых получен на основании двух символов последовательности $T_i(v-1)$ для $v \in \{0, 1, 2, \ldots, w\}$. Соответственно, для $T_i(1)$ имеем $t_{i,0}(1) = t_{i,0}(0), t_{i,1}(0)$; $t_{i,1}(1) = t_{i,2}(0), t_{i,3}(0)$; …; $t_{i,n/2-1}(1) = t_{i,n-2}(0), t_{i,n-1}(0)$, а для остальных значений $T_i(v) - t_{i,j}(v) = t_{i,2j}(v-1), t_{i,2j+1}(v-1)$, где $j = 0, 1, 2, \ldots, n/2^v - 1$. Видно, что каждая из последовательностей $T_i(1)$, $T_i(2)$, $T_i(3)$, …, $T_i(w)$ состоит из символов, алфавиты которых различны. Так, алфавит символов последовательности $T_i(1)$ включает четыре (2^{2^1}) символа, а алфавит последовательности

$T_i(2)$ – 16 (2^{2^2}) символов и так далее вплоть до последовательности $T_i(w)$, которая состоит из одного символа, принадлежащего алфавиту, включающему (2^{2^w}) символов.

Пример 1. В качестве примера двоичных тестовых наборов рассмотрим T_i = 01100011 и T_k = 01011011, для которых выполняется условие $n = 8 = 2^3$. Для каждого набора двоичных символов $T_i = 01100011_{(2)}$ и $T_k = 01011011_{(2)}$ в соответствии с приведенными выше определениями существует $w + 1 = 4$ их представлений в виде последовательностей символов, принадлежащих различным алфавитам. Следовательно, $T_i(0) = 01100011_{(2)}$; $T_i(1) = 1203_{(4)}$; $T_i(2) = 63_{(16)}$; $T_i(3) = c_{(256)}$ и $T_k(0) = 01011011_{(2)}$; $T_k(1) = 1123_{(4)}$; $T_k(2) = 5B_{(16)}$; $T_k(3) = [_{(256)}$. Видно, что $T_i(0)$ и $T_k(0)$ представлены в виде наборов двоичных символов 0 и 1, а $T_i(1)$ и $T_k(1)$ состоят из символов четверичного алфавита, включающего символы 0, 1, 2, 3; $T_i(2)$ и $T_k(2)$ используют шестнадцатеричный алфавит; $T_i(3)$ и $T_k(3)$ включают по одному символу из алфавита, состоящего из 2^8 символов. В данном примере для представления $T_i(3)$ и $T_k(3)$ использованы символы кодов ASCII.

Следует отметить, что предложенная интерпретация последовательностей двоичных символов T_i и T_k в виде последовательностей $T_i(0)$, $T_i(1)$, $T_i(2)$, …, $T_i(w)$ и $T_k(0)$, $T_k(1)$, $T_k(2)$, …, $T_k(w)$ позволяет применять для их сравнения расстояние Хэмминга (1). При этом рассматривается пара символов из двух наборов одного и того же алфавита, причем каждый набор состоит из одинакового количества символов. Основываясь на предыдущем примере (T_i = 01100011 и T_k = 01011011) и используя (1), получим D(01100011, 01011011) = 3, D(1203, 1123) = 2, D(63, 5B) = 2 и D(c, [) = 1.

Приведенный пример определения расстояния Хэмминга показывает возможность получения на основании равенства (1) нескольких численных оценок соотношения исходных наборов двоичных символов. Определим новую меру различия двоичных тестовых наборов T_i и T_k, которая состоит из множества численных характеристик, представляющих собой расстояния Хэмминга.

Определение 1. Мера различия $HD(T_i, T_k)$ двоичных тестовых наборов $T_i = t_{i,0}, t_{i,1}, ..., t_{i,n-1}$ и $T_k = t_{k,0}, t_{k,1}, ..., t_{k,n-1}$, где $t_{i,j}, t_{k,j} \in \{0, 1\}$; $j \in \{0, 1, …, n – 1\}$ и $n = 2^w$, а w – целое, состоит из $(w + 1)$ значений расстояния Хэмминга (1) $HD_0, HD_1, ..., HD_v, …, HD_w$, полученных для T_i и T_k, которые представлены символами, задаваемыми их 2^v последовательными битами, где $v \in \{0, 1, …, w\}$.

Анализируемые символы $t_{i,j}$ и $t_{k,j}$ тестовых наборов T_i и T_k согласно определению 1 представляются одним, двумя, четырьмя и так далее вплоть до $n = 2^w$ двоичными битами. Соответственно, применяя равенство (1), формируем численные значения компонент $HD_0, HD_1, HD_2, …, HD_w$ меры различия $HD(T_i, T_k)$. В табл. 1 приведены примеры вычисления меры различия для различных пар тестовых наборов T_i и T_k в случае, когда $n = 2^3 = 8$.

Таблица 1
Численные значения компонент HD_0, HD_1, HD_2, HD_3 меры различия $HD(T_i, T_k)$ для тестовых наборов T_i и T_k

Table 1
Numerical values of the difference measure $HD(T_i, T_k)$ components HD_0, HD_1, HD_2, HD_3 for test patterns T_i and T_k

T_i, T_k		HD_0	T_i, T_k		HD_1	T_i, T_k		HD_2	T_i, T_k		HD_3
T_i	00000000	–	T_i	0000	–	T_i	00	–	T_i	0	–
T_k	11110000	4	T_k	3300	2	T_k	F0	1	T_k	p	1
	00110011	4		0303	2		33	2		3	1
	11100010	4		3202	3		E2	2		в	1
	01010110	4		1112	4		56	2		V	1

Например, для T_i = 00000000 и T_k = 00110011 значения компонент $HD(T_i, T_k)$ сформированы путем сравнения их различий согласно (1): сначала по одному биту, соответственно, имеем HD_0(00000000, 00110011) = 4, затем по двум битам – HD_1(0000, 0303) = 2, по четырем битам –

$HD_2(00, 33) = 2$, по восьми битам – $HD_3(0, 3) = 1$. Отметим, что для вычисления HD_0 использовался двоичный алфавит, для HD_1 и HD_2 – четверичный и шестнадцатеричный алфавиты, а для HD_3 – алфавит расширенных символов кодов ASCII.

Приведенные в табл. 1 примеры показывают неразличимость всех четырех наборов T_k по отношению к набору T_i при использовании классической меры различия – расстояния Хэмминга, так как во всех четырех случаях $HD_0 = 4$. В то же время новая мера различия (см. определение 1) показывает разную степень различия наборов T_k от T_i, выраженную в отличающихся величинах компонент HD_1 и HD_2 указанной меры (табл. 1).

Мера различия $HD(T_i, T_k)$ двоичных тестовых наборов T_i и T_k, соответствующая определению 1, имеет следующие очевидные свойства.

Свойство 1. *Минимальное численное значение компонент HD_0, HD_1, HD_2, …, HD_{w-2}, HD_{w-1}, HD_w меры различия $HD(T_i, T_k)$ равняется 0, а максимальные значения определяются количеством символов в сравниваемых наборах и равняются n, n/2, n/4, …, 4, 2, 1.*

Все компоненты HD_0, HD_1, HD_2, …, HD_w меры различия равняются 0 при совпадении тестовых наборов, т. е. когда $T_k = T_i$, а максимальные их значения достигаются для случаев, когда $T_k = \overline{T_i}$. Отметим, что $n/2^w = n/n = 1$. Соответственно, компонента HD_w принимает только два значения, а именно 0 при $T_k = T_i$ и 1 при $T_k \neq T_i$.

Свойство 2. *Численные значения компонент меры различия $HD(T_i, T_k)$ связаны соотношениями $HD_0 \geq HD_1 \geq \ldots \geq HD_v \geq \ldots \geq HD_w$.*

Выполнение данного свойства объясняется тем, что при вычислении HD_{v+1} используются символы наборов $T_i(v+1)$ и $T_k(v+1)$, каждый из которых состоит их двух символов наборов $T_i(v)$ и $T_k(v)$, т. е. $t_{i,j}(v+1) = t_{i,2j}(v)$, $t_{i,2j+1}(v)$ и $t_{k,j}(v+1) = t_{k,2j}(v)$, $t_{k,2j+1}(v)$, где $j = 0, 1, 2, \ldots, n/2^v - 1$. Таким образом, результатом сравнения $I(t_{i,j}(v+1) \neq t_{k,j}(v+1))$ символов наборов $T_i(v+1)$ и $T_k(v+1)$ могут быть только значения 0 или 1, используемые для получения HD_{v+1} (1), а результатом сравнения двух пар символов последовательностей $T_i(v)$ и $T_k(v)$ могут быть 0, 1 и 2. Следует отметить, что при выполнении неравенства $t_{i,j}(v+1) \neq t_{k,j}(v+1)$ значение HD_{v+1} (1) увеличится только на 1, а HD_v может быть увеличено как на 1, так и на 2, но как минимум на 1.

Из свойства 2 следует, что, как уже отмечалось ранее, при выполнении равенства $HD_0 = 0$ все остальные компоненты меры различия HD_1, HD_2, …, HD_w также равняются 0. Кроме того, при $HD_0 \neq 0$ значения всех остальных компонент меры различия также принимают ненулевые значения. Например, это видно в случае $T_i = 00000000$ и $T_k = 10000000$, для которых $HD_0 = 1$, $HD_1 = 1$, $HD_2 = 1$ и $HD_3 = 1$.

Как упоминалось в работах [5, 6], идея управляемых вероятностных тестов заключается в том, что очередной тестовый набор T_i формируется максимально отличным (удаленным) от ранее сгенерированных наборов $T_0, T_1, \ldots, T_{i-1}$ в терминах заранее определенных мер различия. Для этого на каждом шаге формирования очередного тестового набора осуществляется его выбор из множества кандидатов в тестовые наборы [5, 6, 10, 18]. Основная операция процедуры выбора заключается в определении численного значения используемой меры различия между двумя наборами T_i и T_k, один из которых, например первый, является тестовым набором, а другой – одним из кандидатов в тесты. В результате в качестве очередного тестового набора выбирается тот кандидат в тестовый набор, для которого величины мер (меры) различия принимают максимальные значения.

Поясним процедуру генерирования управляемого вероятностного теста на простейшем примере, представленном в табл. 1, для случая использования расстояния Хэмминга в качестве меры различия.

Предположив, что первым набором управляемого вероятностного теста является $T_i = 00000000$, случайным образом генерируются четыре кандидата в тестовые наборы $T_k = 11110000$, $T_k = 00110011$, $T_k = 11100010$ и $T_k = 01010110$. Затем для каждого кандидата в тесты T_k вычисляется значение меры различия (1) по отношению к тестовому набору T_i. Как видно из табл. 1, значения HD_0 во всех четырех случаях равняются четырем. Классическая методика формирования управляемых вероятностных тестов предполагает использование любого из четырех кандидатов в тесты в качестве следующего тестового набора.

В случае получения максимального значения HD_0 для нескольких кандидатов в тесты введенная авторами новая мера различия $HD(T_i, T_k)$ (см. определение 1) позволяет более полно учитывать различия кандидатов в тесты T_k по отношению к тестовому набору T_i. Для этого необходимо проанализировать значения следующей компоненты HD_1 предложенной авторами меры различия. Из рассматриваемого примера видно, что максимальное значение $HD_1 = 4$ достигается для набора $T_k = 01010110$, который в дальнейшем можно использовать как тестовый набор управляемого вероятностного теста.

В качестве интегральной меры различия двоичных тестовых наборов T_i и T_k размерностью $n = 2^w$ бит можно применить арифметическую сумму компонент HD_0, HD_1, ..., HD_v, …, HD_w предложенной меры:

$$HD_{Total}(T_i, T_k) = \sum_{v=0}^{w} HD_v(T_i, T_j),\ w = \log_2 n. \tag{2}$$

Критерием включения набора в тест может быть максимальное значение $HD_{Total}(T_i, T_k)$. Для примера, рассмотренного выше (см. табл. 1), с четырьмя кандидатами в тесты HD_{Total}(00000000, 11110000) = 8, HD_{Total}(00000000, 00110011) = 9, HD_{Total}(00000000, 11100010) = 10, HD_{Total}(00000000, 01010110) = 11. Максимальное значение интегральной меры различия достигается для набора $T_k = 01010110$.

2. Анализ меры различия. Рассматриваемая мера различия $HD(T_i, T_k)$, ориентированная на применение в управляемых вероятностных тестах, может быть использована и для сравнения двух символьных наборов T_i и T_k любой природы. Как указывалось ранее, новая мера различия основана на сравнении наборов T_i и T_k, представленных символами из различных алфавитов. Основная операция при сравнении наборов символов состоит в определении значения индикаторной функции $I(t_{i,j} \neq t_{k,j})$ как результата сравнения двух символов $t_{i,j}$ и $t_{k,j}$ (1). При вычислении HD_v, $v \in \{0, 1, \ldots, w\}$, согласно равенству (1), используется алфавит символов, каждый из которых состоит из 2^{2^v} значений. Принимая во внимание, что в случае генерирования управляемых вероятностных тестов тестовые наборы и кандидаты в тестовые наборы представляют собой случайные и независимые последовательности символов, можно оценить вероятность получения единичного значения индикаторной функции $I(t_{i,j} \neq t_{k,j})$.

Учитывая, что символы $t_{i,j}$ и $t_{k,i}$ набора $T_i(v)$ формируются равновероятно из их алфавита, состоящего из 2^{2^v} символов, вероятность выполнения равенства $I(t_{i,j} \neq t_{k,j}) = 1$ определяется как $P(t_{i,j} \neq t_{k,j}) = (1 - 1/2^{2^v})$, где выражение $1/2^{2^v}$ определяет вероятность выполнения равенства $t_{i,j} = t_{k,j}$.

Например, для двоичного случая ($v = 0$) $P(t_{i,j} \neq t_{k,j}) = (1 - 1/2)$, а для случая, когда T_i и T_k представлены одним символом из алфавита, включающего $2^{2^w} = 2^n$ символов ($v = w$), $P(t_{i,j} \neq t_{k,j}) = (1 - 1/2^n)$. Для произвольного значения $HD_0(T_i, T_k) = h$ (1), где $h \in \{0, 1, \ldots, n\}$, а $n = 2^w$, вероятность равенства расстояния Хэмминга для компонент новой метрики, а именно HD_0, HD_1, HD_2, …, величине h определяется выражением

$$P(HD_v = 0) = (1/2^{2^v})^{2^{w-v}};$$
$$P(HD_v = h) = \binom{n/2^v}{h} \cdot \left[(1/2^{2^v})^{2^{w-v}-h} \cdot (1-1/2^{2^v})^h \right]; v = 1, \ldots, \left\lfloor \log_2(2^w/h) \right\rfloor; 0 < h \le 2^{w-v}. \tag{3}$$

Численные значения вероятностей $P(HD_v = h)$ для $n = 2^w = 2^3$ и различных значений $h \in \{0, 1, 2, \ldots, 8\}$ приведены в табл. 2.

Таблица 2
Значения вероятностей компонент HD_0, HD_1, HD_2, HD_3 меры различия двоичных тестовых наборов

Table 2
Probability values of components HD_0, HD_1, HD_2, HD_3 of the difference measure for binary test patterns

h	$P(HD_0 = h)$	$P(HD_1 = h)$	$P(HD_2 = h)$	$P(HD_3 = h)$
0	0,00391	0,00391	0,00391	0,00391
1	0,03125	0,04688	0,11718	0,99609
2	0,10937	0,21094	0,87891	–
3	0,21875	0,42188	–	–
4	0,27344	0,31640	–	–
5	0,21875	–	–	–
6	0,10937	–	–	–
7	0,03125	–	–	–
8	0,00391	–	–	–

Указанные значения вероятностей подтверждают данные примера, представленного в табл. 1. Действительно, например, HD_3 принимает только значения 0 и 1, причем вероятность $P(HD_3 = 1)$ практически равняется единице, что в том числе определяется и свойством 1.

Согласно свойству 2 соотношение значений $HD_0 \geq HD_1 \geq \ldots \geq HD_v \geq \ldots \geq HD_w$ компонент меры различия $HD(T_i, T_k)$ и их максимальных величин указывает на зависимость компонент меры различия двоичных тестовых наборов. Определяющей является наиболее информативная компонента HD_0, значение которой накладывает ограничения на величины остальных компонент.

Все множество ненулевых значений $HD_0 \in \{1, 2, \ldots, n\}$, $n = 2^w$, представим в виде диапазонов, каждый из которых определяется w представлениями T_i и T_k в виде последовательностей символов $T_i(v)$ и $T_k(v)$ различных алфавитов. Соответственно, в качестве минимальной границы $\min HD_0$ используем значения 1, 3, 5, 9, …, $2^{w-1}+1$, а в качестве максимальной $\max HD_0$ – значения 2, 4, 8, …, 2^w. Диапазоны значений компонент HD_0, HD_1, HD_2, …, HD_w меры различия $HD(T_i, T_k)$ для случая $n = 64$ ($w = 6$) представлены в табл. 3.

Таблица 3
Диапазоны значений HD_0, HD_1, HD_2, HD_3, HD_4, HD_5 и HD_6 меры различия двоичных тестовых наборов

Table 3
Value ranges HD_0, HD_1, HD_2, HD_3, HD_4, HD_5 and HD_6 of the difference measures between binary test patterns

HD_0		HD_1		HD_2		HD_3		HD_4		HD_5		HD_6	
min	max	min	max	min	max	min	max	min	max	min	max	min	max
1	2	1	2	1	2	1	2	1	2	1	2	1	1
3	4	2	4	1	4	1	4	1	4	1	2	1	1
5	8	3	8	2	8	1	8	1	4	1	2	1	1
9	16	5	16	3	16	2	8	1	4	1	2	1	1
17	32	9	32	5	16	3	8	2	4	1	2	1	1
33	64	17	32	9	16	5	8	3	4	2	2	1	1

В первых двух столбцах табл. 3 приведены значения $\min HD_0$ и $\max HD_0$, определяющие границы диапазонов величины HD_0 для случая $n = 64$. Численные значения границ диапазонов для остальных компонент HD_1, HD_2, …, HD_6 меры различия, приведенные в табл. 3, получены на основании величин $\min HD_0$ и $\max HD_0$. Например, для $3 \leq HD_0 \leq 4$ значение $\min HD_1$ не может быть меньше двух в связи с тем, что три отличающихся бита в наборах T_i и T_k приведут к отличию как минимум двух символов в наборах $T_i(1)$ и $T_k(1)$, так как каждый символ в этих наборах представлен двумя битами. В то же время для тех же значений $3 \leq HD_0 \leq 4$ величина $\min HD_2 = 1$ в силу того, что три отличающихся бита в наборах T_i и T_k могут оказаться в четырехбитном коде одного и того же символа наборов $T_i(2)$ и $T_k(2)$ (см. свойство 2). Значения $\max HD_v$, приве-

денные в табл. 3, повторяют величину $\max HD_0$ для всех диапазонов, для которых $\max HD_0 \le 2^{w-v}$, а в остальных случаях равняются 2^{w-v}.

В общем случае минимальное значение $\min HD_v$ компоненты HD_v для T_i и T_k, состоящих из $n = 2^w$ двоичных символов, а также максимальная ее величина $\max HD_v$ определяются в соответствии со следующими выражениями:

$$\min HD_v = \left\lceil \min HD_0 / 2^v \right\rceil, v = 0,1,...,w, \text{ для } \min HD_0 = 0,1,..., 2^w; \tag{4}$$

$$\begin{aligned} &\max HD_v = \max HD_0, v = 0,1,...,w - \left\lceil \log_2(\max HD_0) \right\rceil, \text{ для } \max HD_0 = 1,2,...,2^w, \\ &\max HD_v = 2^{w-v}, v = w - \left\lceil \log_2(\max HD_0) \right\rceil + 1,...,w-1,w, \text{ для } \max HD_0 = 1,2,...,2^w. \end{aligned} \tag{5}$$

Как видно из табл. 3 и соотношений (4) и (5), с ростом $v \in \{0, 1, \ldots, w\}$ множество значений HD_v заметно уменьшается. Для $n = 2^3 = 8$ численные величины компонент HD_0, HD_1, HD_2 и HD_3 принадлежат диапазонам [0 ÷ 8], [0 ÷ 4], [0 ÷ 2] и [0 ÷ 1], что подтверждается данными табл. 1.

Анализ приведенных выше соотношений значений компонент HD_0, HD_1, HD_2, …, HD_w меры различия $HD(T_i, T_k)$ позволяет сформулировать три основные стратегии применения данной меры различия для генерирования управляемых вероятностных тестов. Напомним, что критерием для выбора одного из кандидатов в тесты в качестве следующего тестового набора является его максимальное отличие от ранее сгенерированных наборов. Рассмотрим предлагаемые стратегии на примере сравнения тестового набора T_i с кандидатами в тесты T_k, множество которых может составлять от десятков наборов до их тысяч [10, 18]. Для двоичных наборов T_i и T_k, состоящих из n бит, выполняются ранее введенные ограничения, т. е. $n = 2^w$, где w – целое. Для случая классической стратегии в качестве тестового набора случайным образом выбирается один из кандидатов в тесты, имеющий максимальное значение HD_0, полученное согласно равенству (1). Допускается, что максимальное значение HD_0 могут иметь несколько кандидатов в тесты. Каждая из стратегий в рамках предложенной меры различия $HD(T_i, T_k)$ основана на применении правила, определяющего критерий различия.

Правило 1. В качестве тестового набора T_k выбирается кандидат в тесты, который только один из всего их множества имеет максимальное значение HD_v для минимального $v \in \{0, 1, \ldots, w\}$ меры различия $HD(T_i, T_k)$.

Как уже отмечалось, для примера, представленного в табл. 1, согласно данной стратегии, из четырех кандидатов в тесты выбирается T_k = 01010110, который для $v = 1$ имеет максимальное значение HD_1, равное четырем, при этом остальные кандидаты в тесты имеют меньшие значения HD_1. Заметим, что уже для $v = 2$ и $v = 3$ кандидаты в тесты T_k = 01010110 и T_k = 11100010 неразличимы (см. табл. 1).

Правило 2. В качестве критерия различия двоичных тестовых наборов T_i и T_k размерностью $n = 2^w$ используется их интегральная мера различия $HD_{Total}(T_i, T_k)$ (2), согласно которой в качестве тестового набора выбирается один из кандидатов в тесты T_k, имеющий ее максимальное значение.

В соответствии с данным правилом, как показывалось ранее, в случае примера, представленного в табл. 1, также будет выбран набор T_k = 01010110.

Для формулировки правила 3 обратим внимание на максимальные значения $\max HD_v$ (5) компонент новой меры различия и вероятность равенства расстояния Хэмминга этим величинам для произвольных двоичных тестовых наборов T_i и T_k (3). В случае когда для T_i и T_k выполняется неравенство $HD_w \ne \max HD_w$, имеем полное совпадение анализируемых наборов. При выполнении неравенства $HD_{w-1} \ne \max HD_{w-1}$ наборы T_i и T_k либо полностью совпадают, либо совпадают их первые $n/2$ бит или последующие $n/2$ бит. Как видно из примера, представленного в табл. 1, HD_{w-1}(00000000, 11110000) = HD_2(00000000, 11110000) = $1 \ne \max HD_{w-1} = \max HD_2(T_i, T_k) = 2$. Соответственно, имеем $n/2 = 4$ совпадающих бит в указанных наборах. Ранее уже показывалось, что максимальные значения для всех компонент HD_v, $v \in \{0, 1, \ldots, w\}$,

которые интерпретируются как максимальные отличия, достигаются для двоичного случая, когда $T_k = \overline{t_{i,0}}, \overline{t_{i,1}}, ..., \overline{t_{i,n-1}}$.

Правило 3. В качестве тестового набора T_k выбирается один из кандидатов в тесты, для которого выполняются равенства $HD_w = \max HD_w$, …, $HD_{v+1} = \max HD_{v+1}$, $HD_v = \max HD_v$ при минимальном значении $v \in \{0, 1, \ldots, w\}$.

Применяя данное правило для того же примера, получаем, что только один кандидат в тесты, а именно T_k = 01010110, соответствует условию правила 3, так как только для этого набора из четырех рассматриваемых $HD_3 = \max HD_3 = 1$, $HD_2 = \max HD_2 = 2$ и $HD_1 = \max HD_1 = 4$ для минимального $v = 1$ (см. табл. 1). Если рассматривать только три кандидата в тесты, а именно 11110000, 00110011 и 11100010, то для двух из них (00110011 и 11100010) будут выполняться равенства $HD_3 = \max HD_3 = 1$ и $HD_2 = \max HD_2 = 2$ при минимальном значении $v = 2$. Соответственно, один из указанных наборов может быть использован в качестве очередного тестового набора.

Необходимо отметить уточнение правила 3 в части принятия к рассмотрению расстояния HD_{v-1} и расстояний HD_{v-2}, HD_{v-3}, … в случае, когда условия этого правила выполняются более чем для одного кандидата в тесты. Для трех кандидатов в тесты предыдущего примера в качестве тестового набора будет выбран кандидат в тесты 11100010, так как для него HD_{v-1}(00000000, 11100010) = HD_1(000000000, 11100010) = 3, а для 00110011, соответственно, HD_1(00000000, 00110011) = 2. Такое решение принимается на базе основного свойства расстояния Хэмминга, заключающегося в том, что чем больше значение этого расстояния, тем больше различий между базовым набором T_i и набором T_k, для которого получено это значение.

В результате при использовании правил 1 и 3 на примере двоичных наборов, приведенных в табл. 1, в обоих случаях был выбран один и тот же кандидат в тесты T_k = 01010110 в качестве тестового набора. Этот факт свидетельствует о близости, но неэквивалентности правил 1 и 3 по результату их применения для генерирования управляемых вероятностных тестов. Их неэквивалентность подтверждается следующим примером.

Пример 2. В качестве примера двоичных тестовых наборов рассмотрим T_i = 0000000000000000 как базовый набор и два кандидата в тесты: T_k = 0101010111110000 и T_k = 0011010111001100, для которых выполняется условие $n = 16 = 2^4$. Для обоих кандидатов в тесты значение $HD_0 = 8$, что предопределяет использование остальных компонент новой меры различия HD_1, HD_2, HD_3, HD_4 и одного из правил 1, 2 или 3. Для первого кандидата в тесты T_k = 0101010111110000 компоненты меры различия принимают значения $HD_1 = 6$, $HD_2 = 3$, $HD_3 = 2$, $HD_4 = 1$, а для второго кандидата T_k = 0011010111001100 значения $HD_1 = 5$, $HD_2 = 4$, $HD_3 = 2$, $HD_4 = 1$. Применяя первое правило, в качестве тестового набора будет выбран первый набор, так как для него значение $HD_1 = 6$ больше чем для второго кандидата в тесты, для которого $HD_1 = 5$. Третье правило определит в качестве тестового набора второго кандидата в тесты, так как для него $HD_2 = \max HD_2 = 4$ для минимального $v = 2$. Оба кандидата в тесты являются неразличимыми при применении правила 2, так как в обоих случаях $HD_{Total}(T_i, T_k)$ (2) равняется 12.

3. Практические модификации меры различия $HD(T_i, T_k)$. Первоначально рассмотрим ограничение, в рамках которого была определена новая мера различия $HD(T_i, T_k)$, соответствующая определению 1. Требование к размерности n двоичного набора T_i о том, что $n = 2^w$, где w – целое, может не всегда выполняться на практике. Соответственно, для $n \neq 2^w$ при отображении исходного набора T_i в последовательности $T_i(1)$, $T_i(2)$, $T_i(3)$, …, $T_i(w)$ для последнего символа последовательности $T_i(v)$, $v \in \{1, 2, \ldots, w\}$, может отсутствовать необходимое количество бит, равное 2^v. Например, в силу того что для набора $T_i = 01100_{(2)}$ ($n = 5$) $w = \lceil \log_2 n \rceil = 3$, возможно его представление в виде последовательностей $T_i(1)$, $T_i(2)$, $T_i(3)$. Однако во всех трех случаях, а именно $T_i(1)$, $T_i(2)$ и $T_i(3)$, для последнего символа соответствующего алфавита отсутствует необходимое количество бит, для $T_i(1)$ не хватает одного бита, а для $T_i(2)$ и $T_i(3)$ – трех бит. Очевидным решением для устранения данного ограничения является циклическая интерпретация исходного набора $T_i = t_{i,0}, t_{i,1}, \ldots, t_{i,n-1}$. Подобная интерпретация предполагает, что следующим битом набора T_i за последним битом $t_{i,n-1}$ будет первый его бит $t_{i,0}$. Таким обра-

зом, для получения необходимого количества бит для последнего символа набора $T_i(v)$ используются первые биты набора. В случае набора $T_i = 01100_{(2)}$ подобная интерпретация позволяет получить $T_i(1) = 01100\underline{\mathbf{0}}_{(2)} = 120_{(4)}$, $T_i(2) = 01100\underline{\mathbf{011}}_{(2)} = 63_{(16)}$ и $T_i(3) = 01100\underline{\mathbf{011}} = c_{(256)}$. В приведенном примере выделены последние биты набора T_i, повторяющие его первые биты, которые использовались для расширения T_i и получения $T_i(1)$, $T_i(2)$ и $T_i(3)$. Возможны и другие подходы доопределения исходного набора T_i для получения необходимого количества бит, представляющих последний символ набора $T_i(v)$. Число дополнительных бит определяется соотношением величины n и числом бит 2^v, необходимых для представления символа. Для заданного n число недостающих бит для последнего символа набора $T_i(v)$ определяется как $2^v - n \bmod 2^v$. Эти биты необязательно доопределяются рассмотренным выше способом. Например, они могут быть определены постоянными нулевыми или единичными значениями либо случайными двоичными величинами.

Снятие ограничения на размерность n двоичного набора T_i путем его доопределения до нужного количества бит позволяет расширить количество алфавитов для иных отображений исходного набора. Напомним, что согласно определению 1 для вычисления компонент HD_0, HD_1, ..., HD_v, …, HD_w используются следующие алфавиты: двоичный, четверичный, шестнадцатеричный, …, (2^{2^v})-й, …, (2^{2^w})-й. Общее их количество равняется величине $w+1=\lceil \log_2 n \rceil+1$. Очевидно, что с учетом возможности расширения исходного двоичного набора до нужного количества бит число алфавитов может быть увеличено до n. Эти алфавиты состоят из символов, задаваемых одним, двумя, тремя, четырьмя битами и т. д., вплоть до алфавита, в котором каждый символ определяется n последовательными битами. Рассматривая пример исходного набора $T_i = 01100_{(2)}$ и его циклические расширения, представим его в виде последовательностей, полученных для $n = 5$ различных алфавитов. Соответственно, $T_i(0)= 01100_{(2)}$, $T_i(1) = 01100\underline{\mathbf{0}}_{(2)} = 120_{(4)}$, $T_i(2) = 01100\underline{\mathbf{0}}_{(2)} = 30_{(8)}$, $T_i(3) = 01100\underline{\mathbf{011}}_{(2)} = 63_{(16)}$, $T_i(4) = 01100_{(2)} = C_{(32)}$. Отметим, что в зависимости от соотношения величины n и количества бит, используемых для задания символа в заданном алфавите, исходный набор не всегда требует своего расширения (см., например, $T_i(4)$).

Определим значения компонент модифицированной меры различия $MD(T_i, T_k)$ как MD_0, MD_1, ..., MD_v, …, MD_{n-1} для исходных двоичных наборов T_i и T_k, представленных в виде последовательностей $T_i(0)$, $T_i(1)$, …, $T_i(n-1)$ и $T_k(0)$, $T_k(1)$, …, $T_k(n-1)$. При отображении наборов T_i и T_k в последовательности $T_i(1)$, $T_i(2)$, …, $T_i(n-1)$ и $T_k(1)$, $T_k(2)$, …, $T_k(n-1)$ будем использовать их циклическую интерпретацию. Значения MD_v вычисляются согласно соотношению (1). Отметим, что каждая из последовательностей будет состоять из фиксированного количества символов, определяемого числом бит для представления символа из заданного алфавита. Для восьмиразрядных двоичных наборов T_i и T_k, используемых в табл. 1, $T_i(0)$ и $T_k(0)$ представляются восемью двоичными символами, $T_i(1)$ и $T_k(1)$ – четырьмя четверичными символами, $T_i(2)$ и $T_k(2)$ – тремя восьмеричными символами; $T_i(3)$, $T_i(4)$, $T_i(5)$ и $T_i(6)$, а также $T_k(3)$, $T_k(4)$, $T_k(5)$ и $T_k(6)$ включают по два символа из соответствующего алфавита, $T_i(7)$ и $T_k(7)$ – по одному символу.

В свою очередь, число символов в последовательностях $T_i(v)$ и $T_k(v)$, $v \in \{0, 1, …, n-1\}$, определяет максимальные значения компонент MD_v меры различия $MD(T_i, T_k)$, вычисляемые согласно равенству (1):

$$\max MD_v = \lceil n/(v+1) \rceil, v = 0,1,...,n-1. \quad (6)$$

В табл. 4 представлены результаты вычислений на основании (1) компонент MD_0, MD_1, MD_2, MD_3 и MD_4 меры различия $MD(T_i, T_k)$ для двоичных наборов T_i и T_k, приведенных в табл. 1. Значение $MD_2(T_i, T_k) = MD_2(00000000, 11100010) = 2$ (табл. 4) получено как результат циклического расширения восьмибитовых наборов $T_i = 00000000$ и $T_k = 11100010$ до наборов, состоящих из девяти бит $T_i = 00000000\underline{\mathbf{0}}$ и $T_k = 11100010\underline{\mathbf{1}}$. Значение $MD_2(\underline{\mathbf{000}}000\underline{\mathbf{000}}$,

111000**101**) = 2 равняется числу несовпадающих восьмеричных символов (см. формулу (1)), выделенных подчеркиванием.

Таблица 4
Численные значения компонент MD_0, MD_1, MD_2, MD_3 и MD_4 меры различия $MD(T_i, T_k)$ для наборов T_i и T_k

Table 4
Numerical values for the difference measure $MD(T_i, T_k)$ components MD_0, MD_1, MD_2, MD_3 and MD_4 for patterns T_i and T_k

		MD_0	MD_1	MD_2	MD_3	MD_4
T_i	00000000	–	–	–	–	–
T_k	11110000	4	2	3	1	2
	00110011	4	2	3	2	2
	11100010	4	3	2	2	2
	01010110	4	4	3	2	2

Остальные значения компонент MD_5, MD_6 и MD_7 модифицированной меры различия $MD(T_i, T_k)$, так же как и MD_4, равняются их максимальным значениям.

Приведенный анализ модифицированной меры различия показывает, что по мере увеличения v значимость компоненты MD_v, так же как и HD_v, существенно уменьшается. Это объясняется тем, что для $v > n/2$ все компоненты MD_v, кроме последней MD_{n-1}, принимают только три возможных значения, а именно значение 0, если $T_i = T_k$, и значение 1 или 2, если $T_i \neq T_k$. Если $HD_v = \max HD_v$, то все последующие компоненты HD_{v+1}, HD_{v+2}, …, HD_{n-1} также принимают максимальные значения, как это видно из примера, приведенного в табл. 3. В обоих случаях новой меры различия, а именно $HD(T_i, T_k)$, описанной в определении 1, и ее модификации $MD(T_i, T_k)$, представленной в настоящем разделе, значимость компонент HD_v и MD_v с ростом v уменьшается из-за уменьшения количества символов при их вычислении согласно равенству (1).

Для более полного сопоставления сравниваемых последовательностей T_i и T_k, состоящих из n бит, расширим диапазон значений компонент новой меры различия за счет представления их отображений n символами в иных алфавитах. В этом случае исходный двоичный набор T_i представляется последовательностями $T_i(1)$, $T_i(2)$, $T_i(3)$, …, $T_i(n-1)$, каждая из которых будет состоять из n символов соответствующего алфавита. При построении $T_i(v)$, $v \in \{1, 2,\ldots, n-1\}$, для $T_i = t_{i,0}, t_{i,1}, \ldots, t_{i,n-1}$ используем его циклическую интерпретацию, в которой каждый бит $t_{i,j}$, $j \in \{0, 1,\ldots, n-1\}$, набора T_i является первым битом двоичного кода символа, состоящего из $v + 1$ бит. В качестве примера подобного отображения рассмотрим получение $T_i(2)$ для $T_i = 01100_{(2)}$. Каждый из $n = 5$ символов $T_i(2)$ состоит из $2 + 1 = 3$ бит T_i. Процедура получения символов восьмеричного алфавита последовательности $T_i(2)$ для исходного двоичного набора $T_i = 01100_{(2)}$ показана на рисунке. Выделены биты исходного набора $T_i = 01100$, используемые для каждого из пяти восьмеричных символов его отображения в последовательность $T_i(2) = 36401$.

01100 0**110**0 01**100** **0**11**00** **01**10**0**
3 6 4 0 1

Формирование символов $T_i(2)$
Character formation of $T_i(2)$

Двоичный тестовый набор $T_i = 000\ldots0$, состоящий из n нулей, в любом алфавите будет иметь вид, аналогичный $T_i = 000\ldots0$, а набор $T_i = 111\ldots1$, состоящий из n единиц, будет состоять из n старших символов алфавита. Например, для восьмеричного алфавита имеем $777\ldots7$, а шестнадцатеричного – FFF…F.

С учетом рассмотренных модификаций и уточнений новая мера различия $MHD(T_i, T_k)$ между двоичными тестовыми наборами T_i и T_k соответствует следующему определению.

Определение 2. Мера различия $MHD(T_i, T_k)$ двоичных тестовых наборов $T_i = t_{i,0}, t_{i,1}, ..., t_{i,n-1}$ и $T_k = t_{k,0}, t_{k,1}, ..., t_{k,n-1}$, где $t_{i,j}, t_{k,j} \in \{0, 1\}, j \in \{0, 1, \ldots, n-1\}$, а n – произвольное целое, состоит из n компонент $MHD_0, MHD_1, \ldots, MHD_{n-1}$, полученных согласно соотношению

$$MHD_v(T_i, T_k) = \sum_{j=0}^{n-1} \mathrm{I}(t_{i,j}\, t_{i,(j+1) \bmod n} \ldots t_{i,(j+v) \bmod n} \neq t_{k,j}\, t_{k,(j+1) \bmod n} \ldots t_{k,(j+v) \bmod n}),\ v = \overline{0,1,\ldots,(n-1)}. \quad (7)$$

Подобная мера различия, основанная на использовании нескольких значений расстояния Хэмминга, выполняемых для $v = 2$, 4 и 8, применяется в работе [22] для анализа близости анализируемой последовательности к идеальному белому шуму. Анализ расстояния Хэмминга для восьмибитных фрагментов двоичной последовательности с эталонными последовательностями с целью анализа их качества как криптографических паролей приведен в статье [23].

С помощью соотношения (7) формируются численные значения MHD_0, MHD_1, MHD_2, …, MHD_{n-1}. В табл. 5 приведены примеры вычисления меры различия $MHD(T_i, T_k)$ для разных пар тестовых наборов T_i и T_k для случая, когда $n = 8$.

Таблица 5

Численные значения компонент меры различия MHD_0, MHD_1, MHD_2, MHD_3 для тестовых наборов T_i и T_k

Table 5

Numerical values of the difference measure components MHD_0, MHD_1, MHD_2, MHD_3 for test patterns T_i and T_k

T_i, T_k		MHD_0	T_i, T_k		MHD_1	T_i, T_k		MHD_2	T_i, T_k		MHD_3
T_i	00000000	–	T_i	00000000	–	T_i	00000000	–	T_i	00000000	–
T_k	11110000	4	T_k	33320001	5	T_k	77640013	6	T_k	FEC80137	7
	00110011	4		01320132	6		13641364	8		36C936C9	8
	11100010	4		33200121	6		76401253	7		EC8125B7	8
	01010110	4		12121320	7		25253641	8		5A5B6C92	8

Из приведенных данных следует, что в качестве следующего тестового набора будет выбран $T_k = 01010110$ в результате применения любого из трех ранее рассмотренных правил. Действительно, согласно правилу 1 $T_k = 01010110$ имеет максимальное значение MHD_1, равное семи, при этом остальные кандидаты в тесты имеют меньшие значения MHD_1. В соответствии с правилом 2 аналогично уравнению (2) определяется $MHD_{Total}(T_i, T_k)$ как сумма всех компонент $MHD(T_i, T_k)$. В результате получим, что наибольшее значение $MHD_{Total}(00000000, 01010110) = 27$ достигается для набора $T_k = 01010110$. И правило 3 предопределяет выбор в тестовые наборы того же кандидата в тесты. Согласно этому правилу два кандидата в тесты 00110011 и 01010110 имеют максимальное значение MHD_2, равное восьми, однако следующая компонента MHD_1 имеет большее значение, равное семи, для набора 01010110 (см. табл. 5).

Для случая, когда $T_i = 000\ldots0$, наличие нулевых символов в представлении T_k в последовательности, представленной в ином алфавите, позволяет существенно упростить процедуру вычисления компонент рассмотренных мер различия, включая последнюю $MHD(T_i, T_k)$. Действительно, результатом вычислений согласно (6) значения MHD_v является разность величины n и количества нулевых символов в представлении $T_i(v)$, $v \in \{0, 1, 2, \ldots, n-1\}$. Например, в последовательности четверичных символов $T_k(1) = 33320001$ (см. табл. 5) число нулевых символов равняется трем. Соответственно, $MHD_1 = 8 - 3 = 5$.

Для определения вычислительной сложности трех вариантов, а именно $HD(T_i, T_k)$, $MD(T_i, T_k)$ и $MHD(T_i, T_k)$, рассмотренной меры различия в качестве базовой используем операцию сравнения двух символов исходных наборов T_i и T_k, которые представлены в виде последовательностей $T_i(0)$, $T_i(1)$, …, $T_i(n-1)$ и $T_k(0)$, $T_k(1)$, …, $T_k(n-1)$. Анализ ранее рассмотренных примеров показывает, что определение расстояния Хэмминга можно свести к подсчету количества s нулевых двоичных символов поразрядной суммы по модулю два T_s исходных наборов T_i и T_k, где $t_{s,j} = t_{i,j} \oplus t_{k,j}$, $j \in \{0, 1, 2, \ldots, n-1\}$. Тогда, например, HD_0, MD_0 и MHD_0 будут определяться величиной $n - s$. Аналогичным образом путем подсчета нулевых символов для других алфавитов рассчитываются остальные компоненты $HD(T_i, T_k)$, $MD(T_i, T_k)$ и $MHD(T_i, T_k)$.

Для $HD(T_i, T_k) = \{HD_0, HD_1, \ldots, HD_v, \ldots, HD_w\}$ количество операций сравнения символов определяется как $O(HD) = n + n/2 + n/4 + \ldots + 1 = 2n - 1$, где $n = 2^w$, а w – целое. В случае $MD(T_i, T_k) = \{MD_0, MD_1, \ldots, MD_v, \ldots, MD_{n-1}\}$ число операций сравнения двух символов определяется выражением $O(MD) = n + \lceil n/2 \rceil + \lceil n/3 \rceil + \lceil n/(n-1) \rceil$.

Наибольшей вычислительной сложностью характеризуется мера различия $MHD(T_i, T_k) = \{MHD_0, MHD_1, \ldots, MHD_v, \ldots, MHD_{n-1}\}$, для которой $O(MHD) = n + n + \ldots + n = n^2$.

Как показывалось ранее, во всех трех случаях информативными являются только несколько первых компонент мер различия, как правило не более двух или трех. Соответственно, вычислительная сложность для всех трех вариантов сравнима и не превышает $3n$.

4. Экспериментальные оценки меры различия. Рассмотренные меры позволяют оценить степень различия двух тестовых наборов T_i и T_k, которые могут быть неразличимыми при использовании других мер различия, например расстояния Хэмминга. Для подтверждения факта неразличимости тестовых наборов был реализован следующий эксперимент. Для заданного тестового набора T_i, полученного случайным образом, формировалось множество из 1000 кандидатов в тесты T_k, которые также формировались случайным образом по равномерному закону распределения. Затем рассчитывались расстояния Хэмминга $D = HD_0$ между T_i и всеми остальными двоичными тестовыми наборами из списка кандидатов T_k и определялось подмножество кандидатов в тесты, которые имели максимальное значение HD_0. Отметим, что HD_0 является первой компонентой всех трех предложенных мер различия $HD(T_i, T_k)$, $MD(T_i, T_k)$ и $MHD(T_i, T_k)$. Эксперименты проводились для различных величин разрядности n наборов T_i и T_k. В качестве примера в табл. 6 приведены результаты вычисления трех компонент HD_0, HD_1 и HD_2 меры различия $HD(T_i, T_k)$ для $n = 16$. Для каждого из семи приведенных в табл. 6 наборов T_i формировались 1000 кандидатов в тесты T_k, и среди них находились наборы с максимальным значением HD_0.

Таблица 6
Результаты вычисления меры различия $HD(T_i, T_k)$ для $n = 16$

Table 6
Results of calculating the difference measure $HD(T_i, T_k)$ for $n = 16$

Номер эксперимента *Experiment number*	T_i	$\max HD_0(T_i, T_k)$, T_k	$\max HD_1(T_i, T_k)$, T_k	$\max HD_2(T_i, T_k)$, T_k
1	1001100010001110	$\max HD_0 = 15$ 0110011001110001	–	–
2	0101000000010010	$\max HD_0 = 14$ 1000101111101101 1000111111100101	$\max HD_1 = 8$ 1000101111101101 1000111111100101	$\max HD_2 = 4$ 1000101111101101 1000111111100101
3	0000111111000001	$\max HD_0 = 14$ 1111000001110110 1111000000010110	$\max HD_1 = 8$ 1111000001110110 1111000000010110	$\max HD_2 = 4$ 1111000001110110 1111000000010110
4	0101011100111111	$\max HD_0 = 14$ 1010010011000000 1010100010100000	$\max HD_1 = 8$ 1010100010100000	–
5	1100110110001110	$\max HD_0 = 14$ 0011001011110000	–	–
6	1110101000000110	$\max HD_0 = 13$ 0001011111011011 0001011010111001 0001010110100001 0001000110111101 1001111111111001 0011011110111001 1001010101110001 1001100111111001 0001010101110011	$\max HD_1 = 8$ 0001011111011011 0001010110100001 0001000110111101 1001111111111001 0011011110111001 1001010101110001 0001010101110011	$\max HD_2 = 4$ 0001011111011011 0001010110100001 0001000110111101 1001111111111001 0011011110111001 1001010101110001 0001010101110011
7	0100000010100001	$\max HD_0 = 14$ 1011110101010110	–	–

Представленные в табл. 6 результаты подтверждают выводы, которые следуют из проведенных авторами экспериментальных исследований:

1. Результаты 1, 5 и 7 доказывают эффективность расстояния Хэмминга, позволившего выбрать единственного из 1000 кандидатов в тесты, максимально отличающегося от T_i.

2. Данные экспериментов 2–4 и 6 показывают, что больше одного из 1000 кандидатов в тесты имеют максимальное значение расстояния Хэмминга. Например, в эксперименте 2 $\max HD_0 = 14$ имеют два кандидата в тесты, а именно 1000101111101101 и 1000111111100101.

3. Эксперимент 4 иллюстрирует работоспособность новой меры $HD(T_i, T_k)$, позволившей выбрать один набор T_k, максимально отличный от T_i. Эксперимент 6 также свидетельствует о целесообразности применения данной меры различия, так как за счет использования компоненты HD_1 множество потенциальных кандидатов в тесты уменьшилось с 9 до 7.

4. Как видно из всех представленных в табл. 6 результатов экспериментов, достаточным оказалось вычисление только трех компонент HD_0, HD_1 и HD_2, что свидетельствует о невысокой вычислительной сложности определения $HD(T_i, T_k)$, превышающей сложность вычисления расстояния Хэмминга не более чем в три раза.

5. Данные экспериментов 2 и 3 показывают неразличимость в обоих случаях двух кандидатов в тесты, имеющих одинаковые значения как максимального расстояния Хэмминга, так и компонент меры $HD(T_i, T_k)$. Данный вывод констатирует тот факт, что, так же как и в случае расстояния Хэмминга, мера $HD(T_i, T_k)$ не всегда позволяет определить единственного кандидата в тесты T_k, максимально отличного от T_i. Однако подмножество T_k максимально отличных от T_i наборов в случае $HD(T_i, T_k)$ меньше либо равно подмножеству, полученному на основании расстояния Хэмминга.

Суммарные значения количеств кандидатов в тесты T_k из 1000, 100 и 10, сгенерированных случайным образом и имеющих максимальное значение HD_0, приведены в табл. 7. Эти значения получены для трех величин общего количества кандидатов в тесты, из которых и выбирается лучший.

Таблица 7
Суммарные значения количеств кандидатов в тесты T_k с максимальным значением HD_0 для $n = 16$

Table 7
Total values of the numbers of test candidates T_k with maximum value of HD_0 for $n = 16$

Общее количество T_k *Total value of T_k*	Количество T_k с максимальным значением HD_0 *Number of T_k with maximum HD_0 value*															
1000	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	424	227	141	67	27	13	25	14	13	19	14	6	3	3	1	1
100	1	2	3	4	5	6	7	8	9	10	11					
	535	219	111	63	42	12	10	3	2	2	1					
20	1	2	3	4	5	6	7	8								
	623	229	99	23	17	4	4	1								

Данные табл. 7 во всех случаях получены на основании 1000 повторяющихся экспериментов. Например, для общего количества T_k, равного 1000, значения 1 и 424 в первом столбце означают, что в 424 экспериментах был найден только один кандидат с максимальным значением HD_0. Следующий столбец показывает, что в 227 экспериментах получено по два набора, имеющих максимальное HD_0. Во всех экспериментах значения T_k и T_i формировались случайным образом по равномерному закону распределения. Данные, представленные в табл. 7, получены для $n = 16$.

Анализ данных табл. 7 позволяет сделать вывод о часто встречающейся неразличимости кандидатов в тесты, имеющих максимальное значение расстояния Хэмминга. Действительно, в среднем в половине случаев более чем один кандидат в тесты T_k имеет максимальное значение расстояния Хэмминга по отношению к T_i (табл. 7). Эта тенденция сохраняется как для различных разрядностей n тестовых наборов, так и для разного числа кандидатов в тесты, из которых выбирается очередной тестовый набор. Действительно, выбрав предельно малое значение общего количества кандидатов в тесты, равное пяти, и проведя 1000 аналогичных эксперимен-

тов, были получены 270 результатов с двумя и более неразличимыми наборами T_k на основании HD_0. Данный вывод обосновывает необходимость применения других мер различия, позволяющих более эффективно и однозначно выбирать очередной тестовый набор из списка кандидатов в тесты.

При неразличимости нескольких кандидатов в тесты на основании расстояния Хэмминга можно применять предложенные авторами новые меры различия. При этом важным является количество компонент, вычисляемых для каждой из таких ситуаций при выборе одного из кандидатов. Это количество зависит от состава множества кандидатов в тесты T_k и набора T_i. При тех же условиях, что и в предыдущих экспериментальных исследованиях, была проведена оценка количества компонент новых мер различия, необходимых для выбора одного кандидата. В табл. 8 представлено количество экспериментов с определенным числом итераций расчета компонент $HD(T_i, T_k)$ и $MHD(T_i, T_k)$, которые были проведены для обнаружения наиболее удаленного от T_i тестового набора T_k. Рассматривался случай 32-битных наборов T_i и T_k (списки из 10 кандидатов в тесты), который в целом повторяет результаты авторов, полученные для других разрядностей тестовых наборов и разных значений размерности списка кандидатов в тесты.

В первом и четвертом столбцах табл. 8 перечислены компоненты мер различия $HD(T_i, T_k)$ и $MHD(T_i, T_k)$, которые вычислялись для выбора одного кандидата. В столбцах *Количество* приведено число случаев из 1000 экспериментов, для которых вычислялась более чем одна компонента. Например, в первой строке и первых трех столбцах показано, что в 175 экспериментах из 1000 вычислялись две компоненты – HD_0 и HD_1. В процентном исчислении это составляет 69,4 % от всех случаев вычислений более одной компоненты.

Таблица 8
Количество экспериментов с определенным числом итераций вычисления компонент мер отличия для $n = 32$

Table 8
Number of experiments with a certain number of iterations of calculating the components of difference measures for n = 32

$HD(T_i, T_k)$			$MHD(T_i, T_k)$		
Компоненты *Components*	Количество *Quantity*	%	Компоненты *Components*	Количество *Quantity*	%
HD_0, HD_1	175	69,4	MHD_0, MHD_1	220	78,3
HD_0, HD_1, HD_2	75	29,8	MHD_0, MHD_1, MHD_2	47	16,7
HD_0, HD_1, HD_2, HD_3	2	0,8	MHD_0, MHD_1, MHD_2, MHD_3	12	4,3
–	–	–	MHD_0, MHD_1, MHD_2, MHD_3, MHD_4	2	0,7

Анализ данных табл. 8 подтверждает невысокую вычислительную сложность новых мер различия. В подавляющем числе случаев для $HD(T_i, T_k)$ и $MHD(T_i, T_k)$ необходимо вычисление только двух компонент, заметно реже – трех и в исключительно редких случаях – больше трех.

Заключение. В работе рассмотрена мера различия, основанная на применении модификаций определения расстояния Хэмминга и отображении двоичных тестовых наборов в виде последовательностей символов, представленных в различных алфавитах. Предложенная мера различия расширяет возможности определения тестовых последовательностей при генерировании управляемых вероятностных тестов. В среднем в половине случаев возникает вопрос о выборе единственного тестового набора из множества наборов, для которых расстояние Хэмминга принимает максимальное значение относительно предыдущего тестового набора. Показано, что тестовые наборы, неразличимые при использовании в качестве меры различия расстояния Хэмминга, имеют отличающиеся значения компонент мер различия $HD(T_i, T_k)$, $MD(T_i, T_k)$ и $MHD(T_i, T_k)$. Это позволяет более точно классифицировать формируемые случайным образом наборы, которые являются кандидатами в тесты. Введенные меры не всегда дают возможность определить единственного кандидата в тесты, однако подмножество потенциальных наборов после применения мер различия меньше исходного, полученного на основании расстояния Хэмминга. В редких случаях подмножество не изменяется. Вычислительные сложности предложенных модификаций определения расстояния Хэмминга для всех трех вариантов сравнимы

и превышают сложность вычисления расстояния Хэмминга не более чем в три раза. Проведенные экспериментальные исследования показали невысокую временную сложность вычисления предложенных мер различия. Дальнейшие исследования целесообразно расширить в части исследования свойств новой меры различия и ее применимости для различных прикладных задач.

Вклад авторов. *В. Н. Ярмолик* предложил меру различия для тестовых наборов, основанную на применении расстояния Хэмминга. *В. В. Петровская* провела экспериментальные исследования. *Н. А. Шевченко* принял участие в обобщении и анализе полученных результатов.

Список использованных источников

1. Duran, J. W. An evaluation of random testing / J. W. Duran, S. C. Ntafos // IEEE Transactions on Software Engineering. – 1984. – Vol. SE-10, no. 4. – P. 438–444.

2. Arcuri, A. Random testing: Theoretical results and practical implications / A. Arcuri, M. Z. Iqbal, L. Briand // IEEE Transactions on Software Engineering. – 2011. – Vol. 38, no. 2. – P. 258–277.

3. An orchestrated survey on automated software test case generation / S. Anand [et al.] // J. of Systems and Software. – 2014. – Vol. C-39, no. 4. – P. 582–586.

4. An empirical comparison of combinatorial testing, random testing and adaptive random testing / H. Wu [et al.] // IEEE Transactions on Software Engineering. – 2020. – Vol. 46, no. 3. – P. 302–320.

5. Ярмолик, В. Н. Контроль и диагностика вычислительных систем / В. Н. Ярмолик. – Минск : Бестпринт, 2019. – 387 с.

6. A survey on adaptive random testing / R. Huang [et al.] // IEEE Transactions on Software Engineering. – 2021. – Vol. 47, no. 10. – P. 2052–2083.

7. A preliminary study of adaptive random testing techniques / M. S. Roslina [et al.] // Intern. J. of Information Technology & Computer Science. – 2015. – Vol. 19, no. 1. – P. 116–127.

8. Ярмолик, С. В. Управляемое случайное тестирование / С. В. Ярмолик, В. Н. Ярмолик // Информатика. – 2011. – № 1(29). – С. 79–88.

9. Nikravan, E. Hybrid adaptive random testing / E. Nikravan, S. Parsa // Intern. J. of Computing Science & Mathematics. – 2020. – Vol. 11, no. 3. – P. 209.

10. Zhibo, Li. An enhanced adaptive random testing by dividing dimensions independently / Li. Zhibo, Li. Qingbao, Yu Lei // Mathematical Problems in Engineering. – 2019. – Vol. 2019. – P. 1–15.

11. Садовский, М. Г. О сравнении символьных последовательностей / М. Г. Садовский // Вычислительные технологии. – 2005. – № 3(10). – С. 106–116.

12. Hamming, R. W. Error detecting and error correcting codes / R. W. Hamming // The Bell System Technical J. – 1950. – Vol. 29, no. 2. – P. 147–160.

13. Левенштейн, В. И. Двоичные коды с исправлением выпадений, вставок и замещений символов / В. И. Левенштейн // Доклады Академии наук СССР. – 1965. – Т. 163, № 4. – С. 845–848.

14. Алгоритмы: построение и анализ : пер. с англ. / Т. Кормен [и др.]. – 3-е изд. – М. : Изд. дом «Вильямс», 2013. – 1328 с.

15. Tannga, M. J. Comparative analysis of Levenshnein distance algorithm and Jaro Winkler for text plagiarism detection application / M. J. Tannga, S. Rahman, Hasniati // J. of Technology Research in Information System and Engineering. – 2017. – Vol. 4, no. 2. – P. 44–54.

16. Needleman, S. A general method applicable to the search for similarities in the amino acid sequence of two proteins / S. Needleman, C. Wunsch // J. of Molecular Biology. – 1970. – Vol. 48, no. 3. – P. 443–453.

17. Smith, T. F. Identification of common molecular subsequences / T. F. Smith, M. S. Waterman // J. of Molecular Biology. – 1981. – Vol. 147. – P. 195–197.

18. Candidate test set reduction for adaptive random testing: An overheads reduction technique / R. Huang [et al.] // J. of Molecular Biology. – 2021. – Vol. 214, no. C. – P. 102730.

19. Ярмолик, В. Н. Мера различия для тестовых наборов при генерировании управляемых вероятностных тестов / В. Н. Ярмолик, В. В. Петровская, И. Мрозек // Информатика. – 2022. – Т. 19, № 4. – С. 7–26.

20. Ярмолик, В. Н. Мера различия для управляемых вероятностных тестов / В. Н. Ярмолик, Н. А. Шевченко, В. В. Петровская // Доклады БГУИР. – 2022. – Т. 20, № 6. – С. 52–60.

21. Неразрушающие тесты с четным повторением адресов для запоминающих устройств / В. Н. Ярмолик [и др.] // Информатика. – 2021. – Т. 18, № 3. – С. 18–35.

22. Многомерный портрет цифровых последовательностей идеального «белого шума» в свертках Хэмминга / В. И. Волчихин [и др.] // Известия высших учебных заведений. Поволжский регион. Технические науки. – 2017. – № 4(44). – С. 4–13.

23. Условия корректного вычисления энтропии осмысленных длинных паролей в пространстве сверток Хэмминга с эталонными текстами на русском и английском языках / В. И. Волчихин [и др.] // Измерение. Мониторинг. Управление. Контроль. – 2019. – № 3(29). – С. 33–38.

References

1. Duran J. W., Ntafos S. C. An evaluation of random testing. *IEEE Transactions on Software Engineering*, 1984, vol. SE-10, no. 4, pp. 438–444.

2. Arcuri A., Iqbal M. Z., Briand L. Random testing: Theoretical results and practical implications. *IEEE Transactions on Software Engineering*, 2011, vol. 38, no. 2, pp. 258–277.

3. Anand S., Burke E. K., Chen T. Y., Clark J., Cohen M. B., …, Zhu H. An orchestrated survey on automate software test case generation. *Journal of Systems and Software*, 2014, vol. C-39, no. 4, pp. 582–586.

4. Wu H., Nie C., Petke J., Jia Y., Harman M. An empirical comparison of combinatorial testing, random testing and adaptive random testing. *IEEE Transactions on Software Engineering*, 2020, vol. 46, no. 3, pp. 302–320.

5. Yarmolik V. N. Control' i diagnostika vuchislitel'nuch system. *Computer Systems Testing and Diagnoses.* Minsk, Bestprint, 2019, 387 p. (In Russ.).

6. Huang R., Sun W., Xu Y., Chen H., Towey D., Xia X. A survey on adaptive random testing. *IEEE Transactions on Software Engineering*, 2021, vol. 47, no. 10, pp. 2052–2083.

7. Roslina M. S., Ghani A. A. A., Baharom S., Zulzazil H. A preliminary study of adaptive random testing techniques. *International Journal of Information Technology & Computer Science*, 2015, vol. 19, no. 1, pp. 116–127.

8. Yarmolik S. V., Yarmolik V. N. *Controlled random testing*. Informatika [*Informatics*], 2011, no. 1(29), pp. 79–88 (In Russ.).

9. Nikravan E., Parsa S. Hybrid adaptive random testing. *International Journal of Computing Science and Mathematics*, 2020, vol. 11, no. 3, p. 209.

10. Zhibo Li., Qingbao Li., Lei Yu. An enhanced adaptive random testing by dividing dimensions independently. *Mathematical Problems in Engineering*, 2019, vol. 2019, pp. 1–15.

11. Sadovskii M. G. *About symbolical sequences comparigion*. Vuchislitel'nue technologii [*Computational Technologise*], 2005, no. 3(10), pp. 106–116 (In Russ.).

12. Hamming R. W. Error detecting and error correcting codes. *The Bell System Technical Journal*, 1950, vol. 29, no. 2, pp. 147–160.

13. Levenshtein V. I. *Binary codes with correction of deletions, insertions and substitutions of characters*. Doklady Akademii nauk SSSR [*Proceedings of the USSR Academy of Sciences*], 1965, vol. 163, no. 4, pp. 845–848 (In Russ.).

14. Cormen T. H., Leiserson C. E., Rivest R. L., Stein C. *Introduction to Algorithm*, 3rd edition. MIT Press, 2009, 1292 p.

15. Tannga M. J., Rahman S., Hasniati. Comparative analysis of Levenshnein distance algorithm and Jaro Winkler for text plagiarism detection application. *Journal of Technology Research in Information System and Engineering*, 2017, vol. 4, no. 2, pp. 44–54.

16. Needleman S., Wunsch C. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 1970, vol. 48, no. 3, pp. 443–453.

17. Smith T. F., Waterman S. Identification of common molecular subsequences. *Journal of Molecular Biology*, 1981, vol. 147, pp. 195–197.

18. Huang R., Chen H., Sun W., Towey D. Candidate test set reduction for adaptive random testing: An overheads reduction technique. *Journal of Molecular Biology*, 2021, vol. 214, no. C, p. 102730.

19. Yarmolik V. N., Petrovskaya V. V., Mrozek I. *Distance measure for controlled random tests*. Informatika [*Informatics*], 2022, vol. 19, no. 4, pp. 7–26 (In Russ.).

20. Yarmolik V. N., Shauchenka M. A., Petrovskaya V. V. *Distance measure for controlled random tests*. Doklady Belorusskogo gosudarstvennogo universiteta informatiki i radioèlektroniki [*Doklady BGUIR*], 2022, vol. 20, no. 6, pp. 52–60 (In Russ.).

21. Yarmolik V. N., Mrozek I., Levantsevich V. A., Demenkovets D. V. *Transparent memory tests with even repeating addresses for storage devices*. Informatika [*Informatics*], 2021, vol. 18, no. 3, pp. 18–35 (In Russ.).

22. Volchikhin V. I., Ivanov A. I., Yunin A. P., Malygina E. A. *A multidimensional picture of numerical sequences of the ideal "white noise" in Hamming convolutions*. Izvestija vysshih uchebnyh zavedenij. Povolzhskij region. Tehnicheskie nauki [*University Proceedings. Volga Region. Engineering Sciences*], 2017, no. 4(44), pp. 4–13 (In Russ.).

23. Volchikhin V. I., Ivanov A. I., Karpov A. P., Yunin A. P. Conditions for the correct calculation of the entropy of meaningful long passwords in the Hamming convolution space with reference texts in Russian and English. Izmerenie. Monitoring. Upravlenie. Kontrol' [*Measuring. Monitoring. Management. Control*], 2019, no. 3(29), pp. 33–38 (In Russ.).

Информация об авторах

Ярмолик Вячеслав Николаевич, доктор технических наук, профессор, Белорусский государственный университет информатики и радиоэлектроники.
E-mail: yarmolik10ru@yahoo.com

Петровская Вита Владленовна, магистр технических наук, Белорусский государственный университет информатики и радиоэлектроники.
E-mail: vita.petrovskaya@gmail.com

Шевченко Николай Алексеевич, студент, Дармштадтский технический университет.
E-mail: nik.sh.de@gmail.com

Information about the authors

Vyacheslav N. Yarmolik, D. Sc. (Eng.), Prof., Belarusian State University of Informatics and Radioelectronics.
E-mail: yarmolik10ru@yahoo.com

Vita V. Petrovskaya, M. Sc. (Eng.), Belarusian State University of Informatics and Radioelectronics.
E-mail: vita.petrovskaya@gmail.com

Nikolai A. Shevchenko, Student, Darmstadt Technical University.
E-mail: nik.sh.de@gmail.com

ОБРАБОТКА СИГНАЛОВ, ИЗОБРАЖЕНИЙ, РЕЧИ, ТЕКСТА И РАСПОЗНАВАНИЕ ОБРАЗОВ

SIGNAL, IMAGE, SPEECH, TEXT PROCESSING AND PATTERN RECOGNITION

УДК 004.93'1; 004.932
https://doi.org/10.37661/1816-0301-2024-21-2-73-85

Оригинальная статья
Original Article

Сравнительный анализ производительности одноплатных компьютеров для разработки микроархитектурного вычислительного комплекса обнаружения возгораний

Д. А. Павленко

Объединенный институт проблем информатики Национальной академии наук Беларуси, ул. Сурганова, 6, Минск, 220012, Беларусь
E-mail: dmitri.pavlenko@gmail.com

Аннотация

Цели. Целью работы является выбор базовой вычислительной микроплатформы бортового микроархитектурного вычислительного комплекса для обнаружения аномальных ситуаций на территории Республики Беларусь из космоса на основе методов искусственного интеллекта.

Методы. Для выбора вычислительного комплекса используется метод сравнительного анализа. К выбранному оборудованию применяется серия тестов производительности и проводится сопоставительный анализ (бенчмаркинг). Сравнительный и сопоставительный анализы осуществляются в соответствии с требованиями технического задания на текущий проект.

Результаты. Проведены сравнительный анализ и тестирование производительности одноплатных компьютеров Raspberry Pi 4 Model B и Cool Pi 4 Model B, а также ИИ-ускорителя Google Coral USB Accelerator с Google Edge TPU. Сравнительный анализ показал, что Raspberry Pi 4 Model B и Cool Pi 4 Model B полностью соответствуют требованиям технического задания на разработку бортового микроархитектурного вычислительного комплекса обнаружения аномальных ситуаций. При этом Cool Pi 4 Model B хорошо справляется с нейросетевыми вычислениями, но в четыре раза медленнее, чем Google Coral USB Accelerator. Нейросетевые вычисления на Raspberry Pi 4 Model B в 22 раза медленнее, чем аналогичные вычисления на Google Coral USB Accelerator. Cool Pi 4 Model B опережает Raspberry Pi 4 Model B примерно в два-три раза при решении задач копирования и сжатия данных и почти в шесть раз при нейросетевых вычислениях.

Заключение. Несмотря на то что Raspberry Pi 4 Model B подходит под требования технического задания в качестве вычислительной основы, при разработке бортового микроархитектурного вычислительного комплекса обнаружения аномальных ситуаций стоит использовать более мощные альтернативы со встроенным ускорителем нейронных сетей (например, Radxa Rock 5 Model A) либо с дополнительным внешним ИИ-ускорителем (например, сочетание Cool Pi 4 Model B и Google Coral USB Accelerator). Использование Raspberry Pi 4 Model B с дополнительным ИИ-ускорителем также приемлемо и увеличит

скорость вычислений в десятки раз. ИИ-ускорители обеспечивают самые быстрые нейросетевые вычисления, но есть нюансы, связанные с новизной технологий, которые будут исследоваться при дальнейшей разработке.

Ключевые слова: одноплатные компьютеры, нейронные процессоры, ускорители нейронных сетей, ИИ-ускорители, тестирование производительности, сопоставительный анализ

Благодарности. Работа выполнена при финансовой поддержке научно-технической программы Союзного государства «Комплекс-СГ» в рамках НИР № 9СГ2.1-225 от 24.02.2023. Выражается благодарность сотрудникам ОИПИ НАН Беларуси Эдуарду Витальевичу Снежко и Дмитрию Васильевичу Морозову за ценные замечания, сделанные в ходе исследовательской работы.

Для цитирования. Павленко, Д. А. Сравнительный анализ производительности одноплатных компьютеров для разработки микроархитектурного вычислительного комплекса обнаружения возгораний / Д. А. Павленко // Информатика. – 2024. − Т. 21, № 2. – С. 73–85. https://doi.org/10.37661/1816-0301-2024-21-2-73-85

Конфликт интересов. Автор заявляет об отсутствии конфликта интересов.

Поступила в редакцию | Received 05.04.2024
Подписана в печать | Accepted 26.04.2024
Опубликована | Published 28.06.2024

Comparative analysis of single-board computers for the development of a microarchitectural computing system for fire detection

Dzmitry A. Paulenka

The United Institute of Informatics Problems of the National Academy of Sciences of Belarus, st. Surganova, 6, Minsk, 220012, Belarus
E-mail: dmitri.pavlenko@gmail.com

Abstract

Objectives. The purpose of the work is to select the basic computing microplatform of the onboard microarchitectural computing complex for the detection of anomalous situations in the territory of the Republic of Belarus from space on the basis of artificial intelligence methods.

Methods. The method of comparative analysis is used to select a computing platform. A series of performance tests and comparative analysis (benchmarking) are performed on the selected equipment. The methods of comparative and benchmarking analysis are performed in accordance with the terms of reference to the current project.

Results. A comparative analysis and performance testing of Raspberry Pi 4 Model B and Cool Pi 4 Model B single-board computers, as well as AI-accelerator Google Coral USB Accelerator with Google Edge TPU have been performed. The comparative analysis showed that Raspberry Pi 4 Model B and Cool Pi 4 Model B fully meet the terms of reference to the current project. At the same time Cool Pi 4 Model B handles neural network calculations well, but four times slower than similar calculations on Google Coral USB Accelerator. Neural network computations on the Raspberry Pi 4 Model B are 22 times slower than similar computations on the Google Coral USB Accelerator. Cool Pi 4 Model B outperforms Raspberry Pi 4 Model B by the factor of two to three for data copying and compression and almost six times faster for neural network computations.

Conclusion. Despite the fact that Raspberry Pi 4 Model B meets the terms of reference to the project as a computational basis, when developing an on-board microarchitectural computing system for detecting anomalous situations, it is worth using more powerful alternatives with built-in AI-accelerators (e.g., Radxa Rock 5 Model A) or with an additional external AI-accelerator (e.g., a combination of Cool Pi 4 Model B and Google Coral USB

Accelerator). Using a Raspberry Pi 4 Model B with an additional AI-accelerator is also acceptable and will speed up computations by several dozen times. AI-accelerators provide the fastest neural network computations, but there are features related to the novelty of the technology that will be explored in further development.

Keywords: single-board computers, neural processors, neural network accelerators, AI-accelerators, performance testing, comparative analysis

Acknowledgements. The work was carried out with the financial support of the scientific and technical Union State Program "Complex-SG". Special thanks to Eduard Vitalievich Snezhko and Dmitry Vasilievich Morozov, employees of the UIIP NAS of Belarus, for valuable comments made during the research work.

For citation. Paulenka D. A. *Comparative analysis of single-board computers for the development of a microarchitectural computing system for fire detection*. Informatika [*Informatics*], 2024, vol. 21, no. 2, pp. 73−85 (In Russ.). https://doi.org/10.37661/1816-0301-2024-21-2-73-85

Conflict of interest. The author declares of no conflict of interest.

Введение. Современные методы обработки изображений должны использовать нейросетевые методы искусственного интеллекта (ИИ), чтобы быть инновационными и конкурентоспособными. Быстрое увеличение рынка устройств, которые используют специализированные микропроцессоры для эффективного обучения и (или) работы нейронной сети, требует новых способов практической реализации нейросетевых методов на этих устройствах.

Объектом данного исследования являются доступные на рынке микрокомпьютерные решения и одноплатные компьютеры с поддержкой параллельных нейросетевых вычислений. Цель работы – выбор базовой вычислительной микроплатформы бортового микроархитектурного вычислительного комплекса для обнаружения аномальных ситуаций на территории Республики Беларусь из космоса на основе методов ИИ (далее МВК «БортВК»).

В предыдущей научно-технической программе Союзного государства «Технология-СГ» в рамках НИОКР № 3.2.4.1/111/34 от 07.09.2016 был проведен сравнительный анализ одноплатных компьютеров и их аналогов для классификации подстилающих поверхностей Земли [1, 2] и выполнена такая классификация с помощью специализированной сверточной нейронной сети на одноплатном компьютере Raspberry Pi Zero Wireless [3–5]. В результате был разработан микромодуль оперативного распознавания, отбора и сжатия видеоинформации на борту малых космических аппаратов (КА), который является автономным аппаратно-программным комплексом для автоматического обнаружения и классификации изображений подстилающей поверхности Земли[1] (рис. 1).

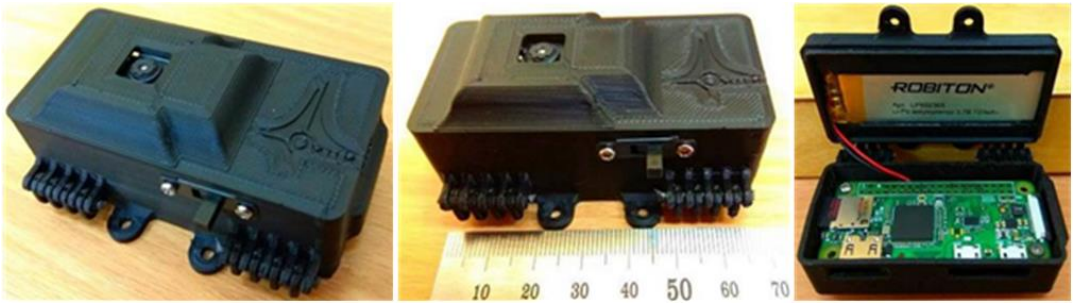


Рис. 1. Опытный образец микромодуля оперативного распознавания, отбора и сжатия видеоинформации на борту малых КА (полетная часть)

Fig. 1. Prototype of a micromodule for operational recognition, selection and compression of video information on board small spacecraft (flight part)

[1]Разработать микромодуль оперативного распознавания, отбора и сжатия видеоинформации на борту малых космических аппаратов : отчет о НИР (заключ.) / ОИПИ НАН Беларуси ; рук. В. А. Ковалев ; исполн.: Д. А. Павленко [и др.]. – Минск, 2020. – 79 с. – № ГР 20164285. – Инв. № 90757.

При дальнейшей разработке автономного аппаратно-программного комплекса данного типа осуществляется переход от задачи классификации подстилающих поверхностей к более сложной задаче обнаружения аномальных ситуаций на спутниковых снимках. После исследований были определены для обнаружения следующие аномальные ситуации:

– последствия пожаров и буреломов в природных экосистемах;

– последствия затопления территорий.

Также предполагается оценка возможности разрабатываемых алгоритмов для обнаружения засухи.

Задача обнаружения аномальных ситуаций использует значительно больше вычислительных мощностей, поэтому для ее решения недостаточно вычислительной мощности одноплатного компьютера Raspberry Pi Zero Wireless.

Общий обзор микрокомпьютерных решений. Областью применения МВК «БортВК» являются поиск и обнаружение аномальных ситуаций и нетипичных визуально детектируемых изменений на спутниковых изображениях на основе методов машинного обучения, ИИ и многоядерного параллелизма. МВК «БортВК» будет выполнен на базе формфактора 1U CubeSat для малых КА. Общая схема работы МВК «БортВК» изображена на рис. 2.

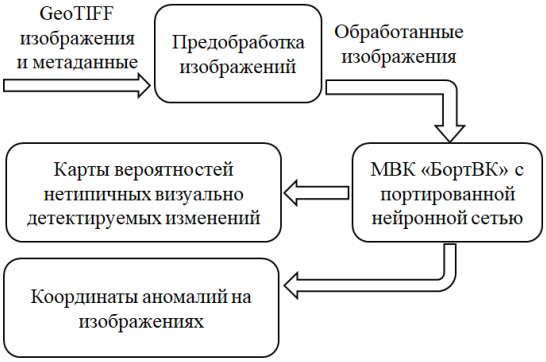


Рис. 2. Общая схема работы МВК «БортВК» без его наземного обеспечения

Fig. 2. General operation scheme of the on-board microcomputing complex MCC "BoardCC" without its ground support

В настоящей статье не рассматривается схема работы наземного программного обеспечения (ПО) для обучения специализированных нейронных сетей и для испытаний функциональных характеристик МВК «БортВК».

Для поддержки параллельных нейросетевых вычислений существуют различные акселераторы нейросетей (AI accelerators), которые отражают эксперименты компаний и стартапов в области аппаратного обеспечения для ИИ. Общее название таких нейросетевых ускорителей по аналогии с центральным процессором (CPU, central processing unit) и графическим процессором (GPU, graphical processing unit) – NPU (neural processing unit), или нейронный процессор.

NPU значительно ускоряет расчеты с использованием нейронных сетей, но есть различные нюансы, связанные с новизной технологий. Важно отметить, что нейронный процессор, в отличие от центрального процессора, нельзя использовать для вычислений общего назначения.

На сегодняшний день существуют следующие разновидности NPU: Tensor Processing Unit (TPU), Neural Network Processor (NNP), Intelligence Processing Unit (IPU), Dataflow Processing Unit (DPU), Vision Processing Unit (VPU), Analog Deep Neural Network (Analog DNN), Associative Processing Unit (APU), AWS Trainium и AWS Inferentia от компании Amazon, Neuromorphic Chip, Quantum Processing Unit (QPU), Photonic Integrated Circuit (PIC) и др.[2, 3].

[2]Аппаратное ускорение глубоких нейросетей: GPU, FPGA, ASIC, TPU, VPU, IPU, DPU, NPU, RPU, NNP и другие буквы [Электронный ресурс]. – Режим доступа: https://habr.com/ru/articles/455353. – Дата доступа: 04.04.2024.

[3]Processing Units – CPU, GPU, APU, TPU, VPU, FPGA, QPU [Electronic resource]. – Mode of access: https://primo.ai/index.php?title=Processing_Units_-_CPU,_GPU,_APU,_TPU,_VPU,_FPGA,_QPU. – Date of access: 04.04.2024.

Некоторые нейронные процессоры (Google TPU, Intel Movidius, Nvidia NVDLA, Amazon AWS) являются составной частью компьютеров в виде интегральных схем либо поставляются отдельно как ИИ-ускорители: платы расширения, USB-ускорители, дополнительное оборудование для масштабирования. Примеры USB ИИ-ускорителей: Google Coral USB Accelerator с вычислительной мощностью 4 TOPS (trillions operations per second), RK1808 NPU с вычислительной мощностью 3 TOPS, Orange Pi AI Stick Lite с вычислительной мощностью 2,8 TOPS, Intel Neural Stick 2 с вычислительной мощностью 1 TOPS и др. Современные встроенные NPU одноплатных компьютеров имеют вычислительную мощность 6 TOPS.

Сравнительный анализ, тестирование производительности и выбор компьютерных решений для МВК «БортВК» (см. рис. 2) осуществляются в соответствии с требованиями технического задания (ТЗ) (табл. 1).

Таблица 1
Требования к составу и параметрам бортовой части МВК «БортВК»

Table 1
Requirements to the composition and parameters of the on-board part of the MCC "BoardCC"

Пункт ТЗ *ToR item*	Наименование требования *Requirement name*	Значение *Value*
4.1.4	Базовое ПО	TensorFlow, PyTorch, Keras, Python, OpenCV, TensorFlow Lite, TensorRT, NumPy
4.1.5	Архитектуры нейронных сетей, которые используют особенности современных нейронных сетей	MobileNet, EfficientNet, DeepLab, SSD
4.4.1	Масса изделия, г, *не более*	400
4.4.2	Мощность энергопотребления, *Вт, не более*	20
4.4.4	Размеры экспериментального образца в конструктиве наноспутника со сторонами, *см, не более*	10×10×10
4.4.5	МВК «БортВК» должен функционировать под управлением операционной системы (ОС) семейства Linux x86-64 и иметь характеристики *не хуже*	ОС Linux x86-64, CPU ARM v8 64-bit, RAM 4 Гб, память 64 Гб, разъем USB 2.0, сетевой адаптер wi-fi
4.5.1	На вход экспериментального образца МВК «БортВК» подаются изображения подстилающей поверхности Земли в формате GeoTIFF и вспомогательные файлы с метаданными. На выходе после обработки изображений формируются карты вероятности нетипичных визуально детектируемых изменений и соответствующие относительные координаты этих изменений на изображениях	*Входные данные*: снимки GeoTIFF и вспомогательные файлы с метаданными *Выходные данные*: карты вероятности нетипичных визуально детектируемых изменений и соответствующие относительные координаты этих изменений на изображениях

Примечание. В таблице приведены только требования к бортовой части МВК «БортВК» без пояснений, которые используются для выбора микроархитектурного решения. Остальные требования ТЗ не приводятся.

Note. The table shows only the requirements for the onboard part of the MCC "BoardCC" without explanations, which are used to select a microarchitectural solution. The remaining requirements of the technical specifications are not given.

При выборе вычислительного комплекса учитывалась его доступность на коммерческом рынке. В силу существующих ограничений на поставку ряда категорий вычислительной техники в Республике Беларусь их приобретение оказалось затруднительным. В итоге для тестов производительности были отобраны и куплены два одноплатных компьютера и один нейронный процессор (рис. 3):

– одноплатный компьютер Raspberry Pi 4 Model B (далее Pi4)[4];
– одноплатный компьютер Cool Pi 4 Model B (далее CoolPi)[5];
– ИИ-ускоритель Google Coral USB Accelerator с Google Edge TPU (далее Coral)[6].

[4]Raspberry Pi 4 [Electronic resource]. – Mode of access: https://www.raspberrypi.com/products/raspberry-pi-4-model-b. – Date of access: 05.04.2024.

[5]Cool Pi 4 Model B – A much faster alternative to Raspberry Pi 4 SBC [Electronic resource]. – Mode of access: https://www.cnx-software.com/2022/12/04/cool-pi-4-model-b-powerful-raspberry-pi-4-alternative. – Date of access: 05.04.2024.

[6]Coral USB Accelerator [Electronic resource]. – Mode of access: https://coral.ai/products/accelerator. – Date of access: 05.04.2024.

a) *b)* *c)*

Рис. 3. Raspberry Pi 4 Model B (*a*), Cool Pi 4 Model B (*b*) и Google Coral USB Accelerator (*c*)
Fig. 3. Raspberry Pi 4 Model B (a), Cool Pi 4 Model B (b) and Google Coral USB Accelerator (c)

Среди одноплатных компьютеров, которые доступны экономически и поддерживают параллельные нейросетевые вычисления, можно также выделить одноплатные компьютеры Orange Pi 5 Plus[7] и Radxa Rock 5 Model A[8].

Orange Pi 5 Plus идентичен CoolPi. У них одинаковые модели чипсетов и GPU, быстрая память типа eMMC, достаточное количество разъемов для масштабирования. В отличие от CoolPi Orange Pi 5 Plus имеет встроенный NPU мощностью 6 TOPS, что на 2 TOPS (или на 33 %) больше, чем у Coral. Операционная система OrangePi OS похожа на оперативную систему Debian и Ubuntu. Продукты компании Orange хорошо задокументированы и поддерживаются сообществом разработчиков. Предполагается, что Orange Pi 5 Plus должен работать быстрее, чем сочетание CoolPi и Coral.

Очевидным недостатком Orange Pi 5 Plus является размер 7,5×10,0 см, в результате чего могут возникнуть трудности при установке его в корпус малого КА размерами 10×10×10 см (см. табл. 1, п. 4.4.4 ТЗ). Адаптер питания к этому одноплатному компьютеру имеет мощность 5 В · 4 А = 20 Вт, что граничит с техническим требованием к энергопотреблению (см. табл. 1, п. 4.4.2 ТЗ).

Одноплатный компьютер Radxa Rock 5 Model A имеет встроенный NPU 6 TOPS, меньшие размеры 8,5×5,6×1,7 см, идентичен Orange Pi 5 Plus (одинаковый чипсет, но меньше разъемов) и экономически доступен. Он менее популярен и менее задокументирован, чем Raspberry Pi и Orange Pi. Имеет модификацию Radxa Rock 5 Model B на том же чипсете RK3588, но с бóльшим количеством разъемов и размерами 10,0×7,5×2,0 см (формфактор Pico-ITX).

В настоящее время нет необходимости приобретать и тестировать Radxa Rock 5 Model A и Orange Pi 5 Plus. Имеющееся в наличии сочетание CoolPi и Coral будет работать на 30 % медленнее при расчетах на NPU и приблизительно одинаково при расчетах на CPU.

Pi4 и CoolPi удовлетворяют всем требованиям ТЗ. Pi4 и его более новый аналог Raspberry Pi 5 (далее Pi5) являются самыми популярными одноплатными компьютерами, но отстают в вычислительной мощности от более современных моделей одноплатных компьютеров. Чипсеты Broadcom BCM2711 у Pi4 и Broadcom BCM2712 у Pi5 не самые мощные на сегодняшний день, но их требования к энергопотреблению значительно ниже, чем у чипсета Rockchip RK3588s компьютера CoolPi.

CoolPi хуже задокументирован и протестирован, чем Pi4 и Pi5. Один пример из многих, с которыми пришлось столкнуться: включение и работа CoolPi под ОС Ubuntu 22.04 выявили частые зависания системы. Информации о причине зависаний на специализированных сайтах найдено не было. На устранение проблемы с зависаниями было потрачено много времени. Выяснилось, что зависания системы происходят из-за графического интерфейса пользователя (GUI) GNOME. Выполнение ТЗ и функционирование МВК «БортВК» не требуют GUI на одноплатном компьютере. Однако разработка и тестирование будут проходить значительно легче

[7]Orange Pi 5 Plus (32GB) [Electronic resource]. – Mode of access: http://www.orangepi.org/html/hardWare/computerAndMicrocontrollers/details/Orange-Pi-5-plus-32GB.html. – Date of access: 05.04.2024.

[8]Radxa ROCK 5A [Electronic resource]. – Mode of access: http://radxa.com/products/rock5/5a. – Date of access: 05.04.2024.

при использовании какого-либо интерфейса пользователя. Удаление сервисов GNOME и замена их на аналогичные менеджеры входа в систему LXDM и среду рабочего стола LXDE устранили зависания системы для CoolPi.

Coral не является самостоятельным компьютером, а представляет собой ускоритель нейросетевых вычислений. Он подключается к компьютеру через разъем USB 3.0 Type-C и используется как дополнительный ИИ-ускоритель при вычислениях через библиотеку машинного обучения TensorFlow Lite. Из недостатков Coral стоит отметить, что данный TPU поддерживает только библиотеку TensorFlow Lite и не поддерживает другие библиотеки машинного обучения.

Технические характеристики тестируемых устройств. Сравнение технических характеристик выполнено для Pi4, CoolPi, Coral и настольного персонального компьютера (ПК) (далее Desktop). Тесты для ПК Desktop в НИР не требуются и выполнены исключительно для наглядности и сравнения. Основные технические характеристики тестируемых устройств приведены в табл. 2.

Таблица 2
Основные технические характеристики тестируемых устройств

Table 2
Main technical specifications of the tested devices

Техническая характеристика *Technical specification*	Pi4	CoolPi	Coral	Desktop
Дата выпуска, месяц/год	06/2019	12/2022	01/2020	06/2013
Цена на сайте AliExpress от 24.03.2024, бел. руб.	263	544	309	Примерно 1300
Компания производитель, страна	Raspberry Pi Foundation, Великобритания	Shenzhen Yanyi Technology Co. Ltd, КНР	Google LLC, США	Intel Corporation, США
Физические размеры, см	8,6×5,7×1,7	8,8×5,7×1,7	6,5×3,0×0,8	41,6×41,0×18,0
Масса, г	46,00	54,00	19,53 (с кабелем 36,37)	Более 6000
ОС	Debian GNU / Linux 11 (bullseye)	Ubuntu Linux 22.04.3 LTS (Jammy Jellyfish)	Нет	Microsoft Windows 10
NPU, TOPS	Нет	Нет	4	Нет
Чипсет	Broadcom BCM2711	Rockchip RK3588s	Google Edge TPU and PMIC	Intel Lynx Point Z87, Intel Haswell
Модель CPU	Cortex-A72	Cortex-A76 и Cortex-A55	Нет	Intel i7-4770
Частота CPU, ГГц	1,8	2,4 и 1,8	Нет	3,6
Кол-во ядер CPU, шт.	4	4 и 4	Нет	4
Модель GPU	Broadcom VideoCore VI	Arm Mali-G610	Нет	Nvidia GeForce GTX 760
Размер RAM, ГБ	3,71	3,63	Нет	15,8
Тип памяти, размер	SanDisk SC64G UHS DDR50 SDXC card, 59,5 ГБ	AT2S9C HS400 eMMC 5.1 card, 58,2 ГБ	Нет	Toshiba THNSNJ512GCST SATA-3, 475 ГБ
Разъемы USB, шт.×тип	2×USB3.0, 2×USB2.0, USB-C OTG	2×USB3.0, 2×USB2.0, USB-C OTG	USB-C	4×USB3.0, 4×USB2.0
Сетевой адаптер wi-fi	Есть	Есть	Нет	Есть

Технические характеристики для табл. 2 взяты из спецификаций продукции на сайтах производителей и других источников. Чипсет Rockchip RK3588s обладает четырьмя ядрами Arm Cortex-A76 и еще четырьмя ядрами Arm Cortex-A55. Всего у него восемь ядер двух типов.

Реальные размеры оперативной памяти (RAM) и памяти хранения немного меньше заявленных, однако это обычная практика изготовителей.

Потребляемая электроэнергия. В соответствии с п. 4.4.2 ТЗ (см. табл. 1) мощность энергопотребления вычислительного устройства не должна превышать 20 Вт. По этой причине были проведены измерения электропитания для одноплатных компьютеров Pi4 и CoolPi.

Измерения проводились на сертифицированном источнике питания МНИПИ Б5-84. Измерялись электрическое напряжение в вольтах (В) и сила электрического тока в амперах (А). Мощность энергопотребления в ваттах (Вт) получена перемножением напряжения и силы тока: Вт = В · А. Результаты измерений приведены в табл. 3.

Таблица 3
Потребляемая одноплатными компьютерами электроэнергия
Table 3
Electricity consumption by single board computers

Измерение, Вт *Measurement, Wt*	Pi4	CoolPi
Пиковое электропотребление при нагрузке	5,12 В · 1,20 А = 6,14	5,00 В · 2,06 А = 10,30
Электропотребление при частичной нагрузке	5,00 В · 1,15 А = 5,75	5,00 В · 1,83 А = 9,15
Пиковое электропотребление при включении	Измерение не проводилось	5,00 В · 1,68 А = 8,40
Энергопотребление в выключенном состоянии	5,00 В · 0,32 А = 1,60	Измерение не проводилось

Максимальная и частичная нагрузки на одноплатные компьютеры производились с помощью ПО stress через следующую команду с различными опциями:

```
# sudo apt install stress  # install stress software
stress --cpu 8 --io 4 --vm 4 --vm-bytes 256M --hdd 4 --hdd-bytes 1024M --timeout 10s
```

CoolPi способен работать при напряжениях больше 5 В, но измерения силы тока при напряжении 12 В и выше не проводились, потому что напряжение бортовой (внешней) системы питания наноспутника CubeSat будет не более 5 В.

В соответствии с официальной документацией[9] максимальная мощность энергопотребления Coral составляет 2 Вт. Одновременное измерение энергопотребления одноплатного компьютера и Coral не проводилось. Поэтому при использовании Coral необходимо прибавить еще 2 Вт к пиковому электропотреблению одноплатного компьютера при нагрузке.

Таким образом, мощность энергопотребления Pi4 и CoolPi с дополнительным ИИ-ускорителем Coral не превышает 13 Вт, что соответствует требованию ТЗ (не более 20 Вт).

Тестирование производительности. В ТЗ нет требований к скорости обработки данных. Однако чем быстрее будут обрабатываться данные, тем лучше.

Тесты производительности (бенчмарки) ПК Desktop проведены для наглядности и сравнения, чтобы показать, что современные одноплатные компьютеры не уступают, а иногда даже превосходят настольные компьютеры с датой выпуска более пяти лет. Для ОС Windows ПК Desktop не удалось найти бесплатное кроссплатформенное ПО, поэтому для тестов производительности на ОС Windows были использованы контейнеры Docker: phoronix/pts[10] и ubuntu[11].

В табл. 4–6 приведены результаты тестов производительности для трех имеющихся в наличии микроплатформ (Pi4, CoolPi, Coral) и ПК Desktop.

Таблица 4
Тестирование производительности тензорного сопроцессора Coral
Table 4
Performance testing of the Coral tensor coprocessor

Тест, мс *Test, ms*	Pi4	CoolPi	Desktop
TPU+TFLite	5,2	4,8	4,7
CPU+TF	117,8	19,7	215,5
CPU+TFLite	140,8	23,9	6420,3

Примечание. TPU может работать только с библиотекой TFLite, поэтому четвертое сочетание TPU+TF не применяется.

Note. TPU can only work with the TFLite library, so the fourth combination TPU+TF does not apply.

[9]Coral USB Accelerator datasheet [Electronic resource]. – Mode of access: https://coral.ai/static/files/Coral-USB-Accelerator-datasheet.pdf. – Date of access: 05.04.2024.

[10]phoronix/pts [Electronic resource]. – Mode of access: https://hub.docker.com/r/phoronix/pts. – Date of access: 05.04.2024.

[11]ubuntu [Electronic resource]. – Mode of access: https://hub.docker.com/_/ubuntu. – Date of access: 05.04.2024.

Таблица 5
Тестирование производительности с помощью скрипта sbc-bench.sh

Table 5
Performance testing with the use of sbc-bench.sh script

Тест *Test*	Pi4	CoolPi	Desktop
memcpy, Мбайт/с	2469	7829	7407
memset, Мбайт/с	3077	24 766	14 722
7-zip, MIPS	5720	15 100	16 870
AES-256, 16кБ, мегахеш	36	1091	584
Троттлинг	Нет	Да	Да

Таблица 6
Тестирование производительности с помощью ПО Phoronix Test Suite, тест stress-ng

Table 6
Performance testing with the Phoronix Test Suite software, stress-ng test

Тест, Bogo Ops/s *Test, Bogo Ops/s*	Pi4	CoolPi	Desktop
CPU Stress	392	852	6806
Matrix Math	1342	14 838	18 742
Memory Copying	395	719	1260

Тесты производительности в табл. 4 проведены для обученной модели нейронной сети MobileNet v3, которая находится в файлах репозитория Coral[12]: tf2_mobilenet_v3_edgetpu_1.0_224_ptq_edgetpu.tflite для TPU и tf2_mobilenet_v3_edgetpu_1.0_224_ptq.tflite для CPU. ПО для проведения тестов производительности Coral TPU разработано специально для данного испытания и находится в файлах tf_lite_benchmarks.py и tf_lite.py репозитория автора исследования[13].

В первом ряду табл. 4 TPU+TFLite приведены результаты тестов производительности для TPU Coral и библиотеки TensorFlow Lite (TFLite), во втором ряду CPU+TF – тесты производительности для CPU и библиотеки TensorFlow (TF), в третьем ряду CPU+TFLite – тесты производительности для CPU и библиотеки TensorFlow Lite (TFLite).

Для тестов производительности в табл. 5 использовано ПО Томаса Кайзера sbc-bench.sh[14], которое разработано специально для тестов различных моделей одноплатных компьютеров.

В табл. 5 первый ряд memcpy – это скорость копирования данных из одного блока памяти в другой с помощью функции memcpy языка программирования C/C++, второй ряд memset – скорость инициализации блока памяти с помощью функции memset языка программирования C/C++, третий ряд 7-zip – скорость сжатия данных с помощью ПО 7-zip, четвертый ряд AES-256 – скорость шифрования данных с помощью алгоритма AES-256, пятый ряд Троттлинг – наличие или отсутствие троттлинга при большой нагрузке на CPU в течение пяти минут. Троттлинг отсутствует только для Pi4, что связано с плохой системой охлаждения для CoolPi и Desktop.

Для тестов производительности в табл. 6 использовано ПО Phoronix Test Suite[15]. Оно включает в себя сотни различных тестов производительности, из которых был выбран тест stress-ng. С помощью стресс-теста stress-ng проведены тесты производительности CPU (ряд CPU Stress), матричной математики (ряд Matrix Math) и копирования блоков памяти (ряд Memory Copying).

[12]google-coral/test_data [Electronic resource]. – Mode of access: https://github.com/google-coral/test_data. – Date of access: 27.03.2024.

[13]Paulenka, D. A. Coral TPU project [Electronic resource] / D. A. Paulenka. – Mode of access: https://github.com/foobar167/junkyard/tree/master/coral_tpu. – Date of access: 05.04.2024.

[14]ThomasKaiser/sbc-bench [Electronic resource]. – Mode of access: https://github.com/ThomasKaiser/sbc-bench. – Date of access: 05.04.2024.

[15]Phoronix Test Suite [Electronic resource]. – Mode of access: https://www.phoronix-test-suite.com. – Date of access: 05.04.2024.

В ПО Phoronix Test Suite производительность измеряется величиной Bogo Ops/s (bogus operations per second), которая является способом измерения скорости исполнения инструкций на компьютере в ядре Линукс.

Журнальные файлы с результатами тестирования находятся в файлах репозитория автора исследования[16]:

– data/coral-tpu-benchmark-results.txt для табл. 4;

– data/Thomas-Kaiser-sbc-bench-results.txt для табл. 5;

– data/phoronix-stress-ng-results.txt для табл. 6.

Для наглядности данные из табл. 4–6 показаны в виде диаграмм на рис. 4–6. На рис. 4 чем меньше столбик диаграммы, тем лучше. На рис. 5 и 6 чем выше столбик диаграммы, тем лучше.

Для наглядности на рис. 4 шкала времени представлена в логарифмическом масштабе. Видно, что скорости расчетов на Coral TPU через библиотеку TFLite (TPU + TFLite) отличаются незначительно для различных устройств и значительно опережают расчеты на CPU. Расчеты на CPU для одной и той же модели нейросети MobileNet v3 через библиотеку TensorFlow (CPU + TF) производятся немного быстрее, чем расчеты через библиотеку TensorFlow Lite (CPU + TFLite), кроме ПК Desktop.

Возможно, для ПК Desktop значительное отличие почти в 30 раз между скоростью расчетов CPU + TF (215,5 мс) и CPU + TFLite (6420,3 мс) связано с какими-то внутренними оптимизациями библиотек TF и TFLite.

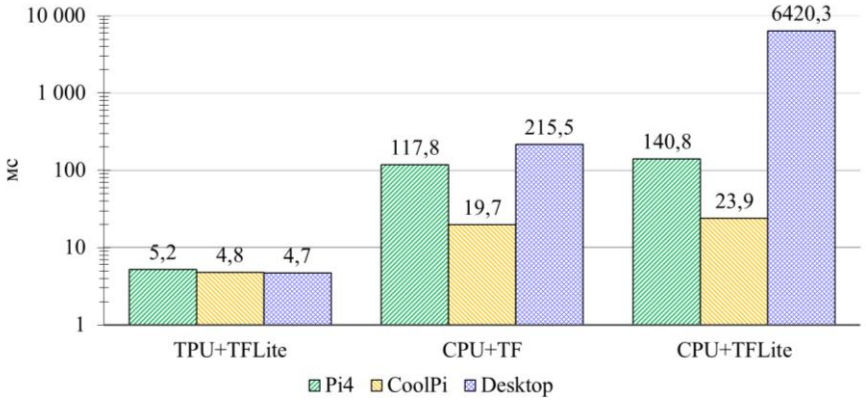


Рис. 4. Тестирование производительности тензорного сопроцессора Coral TPU

Fig. 4. Performance testing of the Coral TPU tensor coprocessor

На рис. 4 видно, что восемь ядер одноплатного компьютера CoolPi хорошо справляются с параллельными нейросетевыми вычислениями, но эти расчеты в четыре раза медленнее, чем аналогичные расчеты на Coral. Вычисления на процессоре Pi4 в 22 раза медленнее, чем аналогичные расчеты на Coral, и в шесть раз медленнее, чем расчеты на CoolPi. Более современные одноплатные компьютеры CoolPi и Pi4 значительно опережают по нейросетевым вычислениям устаревший процессор Intel i7-4770 ПК Desktop.

На рис. 5 и 6 видно, что более быстрый одноплатный компьютер CoolPi опережает Pi4 примерно в два-три раза для задач копирования (memcpy, Memory Copying), сжатия данных (7-zip) и нагрузки на центральный процессор (CPU Stress). CoolPi опережает Pi4 в восемь раз для задачи инициализации (выделения) блока памяти (memset) и в 11 раз для матричных операций (Matrix Math). CoolPi в 30 раз быстрее шифрует данные с помощью алгоритма AES-256, чем Pi4, но этот показатель важен при защищенной передаче данных и не важен для нейросетевых расчетов. Скорость матричных операций восьмиядерного процессора CoolPi сравнима со скоростью восьмиядерного процессора Intel i7-4770 ПК Desktop.

[16]URL: https://github.com/foobar167/junkyard/tree/master/coral_tpu

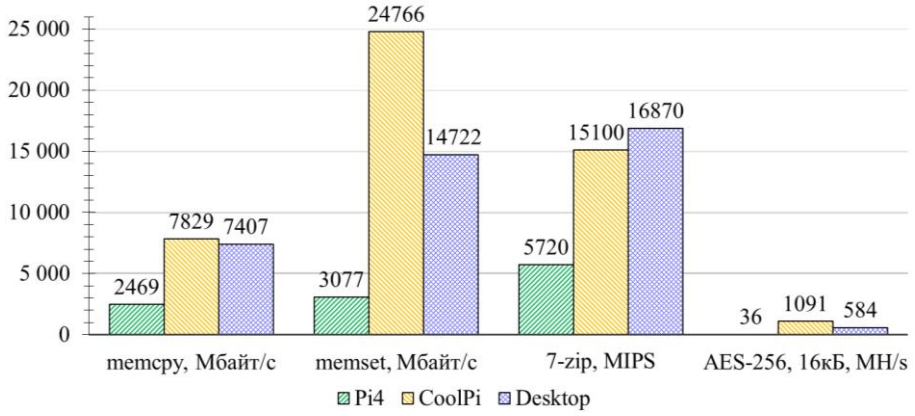


Рис. 5. Тестирование производительности с помощью скрипта sbc-bench.sh Томаса Кайзера

Fig. 5. Performance testing with the use of Thomas Kaiser's sbc-bench.sh script

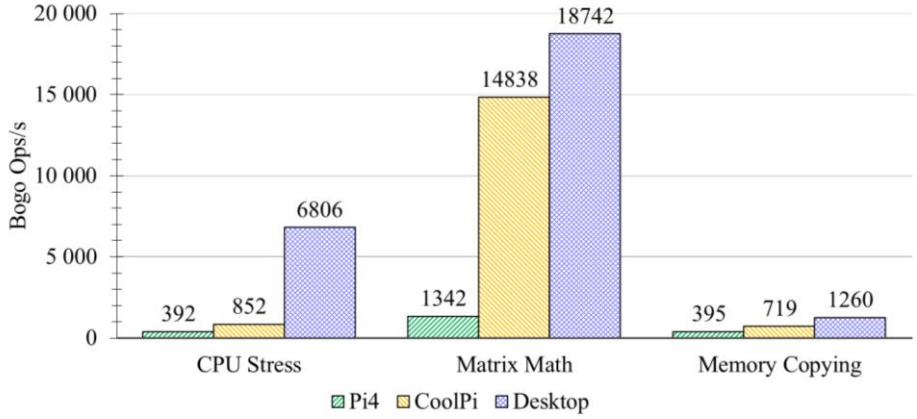


Рис. 6. Тестирование производительности с помощью ПО Phoronix Test Suite, тест stress-ng

Fig. 6. Performance testing with the Phoronix Test Suite software, stress-ng test

Несмотря на то что Pi4 подходит под требования ТЗ в качестве вычислительной основы, при разработке МВК «БортВК» стоит использовать более мощные альтернативы со встроенным NPU (например, Radxa Rock 5 Model A) либо без встроенного NPU, но с дополнительным ИИ-ускорителем (сочетание CoolPi и Coral TPU). Использование Pi4 с дополнительным ИИ-ускорителем (сочетание Pi4 и Coral TPU) также приемлемо и увеличит скорость вычислений в десятки раз.

Coral дает значительное ускорение нейросетевых расчетов, однако есть особенности, исследование которых будет проводиться на следующих этапах НИР:

– Coral работает только с библиотекой TensorFlow Lite;

– для использования Coral TPU нейронная сеть должна быть предварительно квантована (quantization), когда все веса нейронной сети переводятся из чисел с плавающей запятой (float или double) в целочисленные значения (uint8 или int8);

– если по какой-либо причине не все веса нейронной сети преобразовались из float в uint8, например какой-то слой нейросети не поддерживает квантование, то такая частично квантованная сеть будет выполняться не на Coral TPU, а на CPU;

– тесты производительности Coral в табл. 4 и на рис. 4 выполнены при помощи готовых квантованных нейронных сетей, но, чтобы выполнить задачу адаптации (портирования) нейросети для одноплатного компьютера, необходимо научиться самостоятельно квантовать веса нейронной сети.

Как было показано выше, акселераторы нейросетей обеспечивают самые быстрые нейросетевые вычисления, однако есть нюансы, связанные с новизной технологий: отсутствие подробной документации, малое количество примеров, слабая поддержка пользователя, редкое обновление аппаратного и программного обеспечения и др. Со временем ситуация в области разработки NPU будет улучшаться, а использование дополнительных ИИ-ускорителей упрощаться и лучше документироваться.

Заключение. Обзор доступных на рынке микрокомпьютерных решений, поддерживающих параллельные нейросетевые вычисления, показал, что на данный момент оптимальной вычислительной платформой для разработки МВК «БортВК» в соответствии с ТЗ является одноплатный компьютер Cool Pi 4 Model B с подключенным к нему тензорным сопроцессором Google Coral USB Accelerator для ускорения параллельных нейросетевых вычислений. Альтернативой ему служит менее мощный, но более задокументированный и протестированный одноплатный компьютер Raspberry Pi 4 Model B или его аналог Raspberry Pi 5 с подключенным к нему тензорным сопроцессором Google Coral USB Accelerator.

Список использованных источников

1. Сравнительный анализ вычислительных платформ для бортового микромодуля предварительного распознавания изображений / В. А. Ковалев [и др.] // Информатика. – 2018. − Т. 15, № 3. – С. 7–21.

2. Comparative analysis of budget computing platforms for a portable micromodule of on-board image classification / V. A. Kovalev [et al.] // BIG DATA and Advanced Analytics : Collection of Materials of the Fourth Intern. Scientific and Practical Conf., Minsk, Belarus, 3–4 May 2018 / editorial board: M. Batura [et al.]. – Minsk, BSUIR, 2018. – P. 31–42.

3. Распознавание подстилающей поверхности Земли с помощью сверточной нейронной сети на одноплатном микрокомпьютере / Д. А. Павленко [и др.] // Информатика. – 2020. − Т. 17, № 3. – С. 36–43. https://doi.org/10.37661/1816-0301-2020-17-3-36-43

4. Интеллектуальная технология распознавания подстилающей поверхности Земли / С. В. Кругликов [и др.] // Радиоэлектронные технологии. – 2019. – № 1. – С. 90–94.

5. Recognition of underlying surface using a convolutional neural network on a single-board computer / D. A. Paulenka [et al.] // BIG DATA and Advanced Analytics : сб. материалов VI Междунар. науч.-практ. конф., Минск, Беларусь, 20–21 мая 2020 г. : в 3 ч. Ч. 1 / редкол.: В. А. Богуш [и др.]. – Минск : Бестпринт, 2020. – С. 71–77.

References

1. Kovalev V. A., Paulenka D. A., Snezhko E. V., Liauchuk V. A., Kalinovski A. A. *Comparative analysis of computing platforms for onboard micromodule of provisional image recognition*. Informatika [*Informatics*], 2018, vol. 15, no. 3, pp. 7–21 (In Russ.).

2. Kovalev V. A., Paulenka D. A., Snezhko E. V., Liauchuk V. A. Comparative analysis of budget computing platforms for a portable micromodule of on-board image classification. *BIG DATA and Advanced Analytics : Collection of Materials of the Fourth International Scientific and Practical Conference, Minsk, Belarus, 3–4 May 2018*. Editorial board: M. Batura [et al.]. Minsk, Belorusskij gosudarstvennyj universitet informatiki i radiojelektroniki, 2018, pp. 31–42.

3. Paulenka D. A., Kovalev V. A., Snezhko E. V., Liauchuk V. A., Pechkovsky E. I. *Recognition of the Earth's underlying surface using a convolutional neural network on a single-board microcomputer*. Informatika [*Informatics*], 2020, vol. 17, no. 3, pp. 36–43 (In Russ.). https://doi.org/10.37661/1816-0301-2020-17-3-36-43

4. Kruglikov S. V., Kovalev V. A., Paulenka D. A., Snezhko E. V., Liauchuk V. A. *Intelligent technology for recognizing the underlying surface of the Earth*. Radiojelektronnye tehnologii [*Radioelectronic Technology*], 2019, no. 1, pp. 90–94 (In Russ.).

5. Paulenka D. A., Kovalev V. A., Snezhko E. V., Liauchuk V. A., Pechkovsky E. I. *Recognition of underlying surface using a convolutional neural network on a single-board computer*. BIG DATA and Advanced Analytics : sbornik materialov VI Mezhdunarodnoj nauchno-prakticheskoj konferencii, Minsk, Belarus', 20–21 maja 2020 goda : v 3 chastjah. Chast' 1 [*BIG DATA and Advanced Analytics : Collection of Materials of the VI International Scientific and Practical Conference, Minsk, Belarus, 20–21 May 2020) : in 3 Parts. Part 1*]. Editorial board: V. A. Bogush [et al.]. Minsk, Bestprint, 2020, pp. 71–77.

Информация об авторе

Павленко Дмитрий Анатольевич, ведущий инженер-программист, лаборатория анализа биомедицинских изображений, Объединенный институт проблем информатики Национальной академии наук Беларуси.
E-mail: dmitri.pavlenko@gmail.com
https://www.researchgate.net/profile/Dzmitry-Paulenka
https://scholar.google.com/citations?user=2AX0it0AAAAJ
https://orcid.org/0009-0007-9911-4356

Information about the author

Dzmitry A. Paulenka, Lead Software Engineer, Laboratory of Biomedical Images Analysis, The United Institute of Informatics Problems of the National Academy of Sciences of Belarus.
E-mail: dmitri.pavlenko@gmail.com
https://www.researchgate.net/profile/Dzmitry-Paulenka
https://scholar.google.com/citations?user=2AX0it0AAAAJ
https://orcid.org/0009-0007-9911-4356

УДК 004
https://doi.org/10.37661/1816-0301-2024-21-2-86-93

Оригинальная статья
Original Article

Интерактивная сегментация изображений на основе их кластеризации

Б. А. Залесский

Объединенный институт проблем информатики
Национальной академии наук Беларуси,
ул. Сурганова, 6, Минск, 220012, Беларусь
E-mail: zalesky@newman.bas-net.by

Аннотация

Цели. Рассматривается задача сегментации цветных изображений без использования предварительного обучения. Она возникает, например, когда необходимо выполнить сегментацию изображений с неизвестными заранее семантическими и цветовыми свойствами непосредственно после их получения или когда набор изображений, предназначенных для сегментации, слишком мал, а также при выполнении предварительного «разведочного» анализа изображений. В таких случаях невозможно использование мощных нейросетевых и других средств сегментации, требующих глубокого обучения.

Методы. Предлагается алгоритм интерактивной сегментации изображений, основанный на анализе цветов областей, выделенных в интерактивном режиме. Вначале в интерактивном режиме выделяются весьма приближенно области изображения, принадлежащие объектам, а затем – принадлежащие фону. На следующем шаге множество цветов выделенных областей объектов и множество цветов выделенных областей фона кластеризуются по отдельности одним из алгоритмов кластеризации, например *k*-средних, нечетких *c*-средних, или предложенным автором алгоритмом многоуровневой кластеризации. После этого из множества центров кластеров, описывающих объект, и множества кластеров, описывающих фон, удаляются неинформативные элементы. Преобразованные множества центров кластеров объектов и фона используются для сегментации изображения.

Результаты. Построенный алгоритм позволяет выделить на цветном изображении требуемые объекты в случае, когда их цвет отличается от цвета фона. Интерактивное выделение областей объектов и областей фона не требует аккуратности и больших усилий и обычно занимает от нескольких десятков секунд до минуты. Для выделения достаточно использовать прямоугольные области, лежащие целиком внутри изображений объектов, и прямоугольные области, лежащие целиком внутри фона. Приводятся пример интерактивного выделения областей и результаты сегментирования цветных изображений.

Заключение. Проведенные эксперименты показали эффективность предложенного подхода сегментирования цветных изображений. Его можно применять в случаях, когда заранее неизвестны семантические и цветовые свойства изображений, и в случаях, когда использование более мощных методов глубокого обучения, включая нейронные сети, слишком затратно или невозможно.

Ключевые слова: цветные изображения, сегментация по цвету, кластеризация, метод *k*-средних, метод нечетких *c*-средних

Для цитирования. Залесский, Б. А. Интерактивная сегментация изображений на основе их кластеризации / Б. А. Залесский // Информатика. – 2024. − Т. 21, № 2. – С. 86–93.
https://doi.org/10.37661/1816-0301-2024-21-2-86-93

Конфликт интересов. Автор заявляет об отсутствии конфликта интересов.

Поступила в редакцию | Received 10.05.2024
Подписана в печать | Accepted 24.05.2024
Опубликована | Published 28.06.2024

Clustering-based interactive image segmentation

Boris A. Zalesky

The United Institute of Informatics Problems
of the National Academy of Sciences of Belarus,
st. Surganova, 6, Minsk, 220012, Belarus
E-mail: zalesky@newman.bas-net.by

Abstract
Objectives. The task of color image segmentation without the use of preliminary training is considered. It arises, for example, when it is necessary to perform image segmentation with semantic and color properties unknown in advance immediately after their acquisition, or when the set of images intended for segmentation is too small, as well as when performing preliminary "exploratory" analysis of images. In such cases, powerful neural network and other segmentation tools that require deep learning can not be used.
Methods. An algorithm for interactive image segmentation is proposed, based on the analysis of the colors of areas selected interactively. First, in interactive mode, the image areas belonging to the objects are selected very approximately, and then regions belonging to the background are chosen. In the next step, the set of colors of the selected object areas and the set of colors of the selected background areas are clustered separately by one of the clustering algorithms, for example, *k*-means, fuzzy *c*-means, or the multi-level clustering algorithm proposed by the author. After this, non-informative elements are removed from the set of cluster centers describing the objects and the set of clusters presenting the background. The modified sets of object and background cluster centers are used for image segmentation.
Results. The constructed algorithm allows selection of the required objects in color images if the colors of the objects and the background are different. Interactive selection of object areas and background areas does not require accuracy or much effort and usually takes several tens of seconds. For selection, rectangular areas that lie entirely inside the object images, and rectangular areas that belong completely to the background can be used. Below an example of interactive regions selection and color image segmentation is shown.
Conclusion. The experiments performed showed the effectiveness of the proposed approach to segmenting color images. It can be used in cases where the semantic and color properties of images are not known in advance, and in cases where the use of more powerful deep learning methods, including neural networks, is too expensive or impossible.

Keywords: color images, color segmentation, clustering, method *k*-means, method fuzzy *c*-means

For citation. Zalesky B. A. *Clustering-based interactive image segmentation*. Informatika [*Informatics*], 2024, vol. 21, no. 2, pp. 86–93 (In Russ.). https://doi.org/10.37661/1816-0301-2024-21-2-86-93

Conflict of interest. The author declares of no conflict of interest.

Введение. В настоящее время существует большое число алгоритмов сегментации изображений. Во многих книгах по компьютерному зрению присутствуют разделы, посвященные известным алгоритмам сегментации изображений [1–4]. В некоторых из них описаны алгоритмы сегментации изображений с помощью кластеризации, например алгоритм *k*-means и его многочисленные нечеткие версии [5], алгоритм наращивания областей [1] и др.

В последние годы появилось значительное число нейросетевых алгоритмов сегментации изображений. Это алгоритмы, основанные на применении сетей-персептронов, сверточных сетей U-Nets, DeepLab, Mask R-CNNs, более современных трансформеров, использующих механизм внимания, сетей VIT, созданный в 2023 г. и наиболее часто используемый в 2024 г. Yolov8[1] и др. Большое число статей, посвященных решению задачи семантической сегментации изображений с помощью упомянутых нейронных сетей, и программных реализаций нейронных сетей приведено на сайте[2].

[1]Ultralytics YOLOv8 Docs [Electronic resource]. – Mode of access: https://docs.ultralytics.com/ru. – Date of access: 07.05.2024.

[2]Image Segmentation [Electronic resource]. – Mode of access: https://paperswithcode.com/task/image-segmentation. – Date of access: 07.05.2024.

Однако в некоторых случаях применение мощных нейросетевых алгоритмов сегментации не представляется возможным (или слишком затратным по времени и ресурсам), так как они требуют предварительного обучения. В ряде задач свойства изображений заранее неизвестны или сети оказываются не обученными для применения на конкретном типе изображений и при этом отсутствуют размеченные обучающие наборы данных. Например, для быстрого анализа одного или нескольких изображений клеток человека или животного бывает необходимо оценить их количество или форму и размер. Если под рукой нет алгоритмов, обученных для сегментации имеющихся изображений, то можно применить предлагаемый алгоритм сегментации через кластеризацию, которому не нужно предварительного обучения. Он менее вычислительно затратен и не требует использования современной видеокарты или мощного многоядерного процессора.

Алгоритм может применяться при сопровождении объекта, наблюдаемого видеокамерой (без интерактивного выделения областей этого объекта и фона на каждом кадре), в случаях, когда на начальном кадре сопровождаемый объект выделяется каким-либо средством, затем на следующем кадре алгоритм сопровождения находит ограничивающий объект прямоугольник, тогда цветовые характеристики объекта берутся с предыдущего кадра, а цветовые характеристики фона выбираются из небольшой окрестности прямоугольника, ограничивающего найденный объект. После этого используется предложенный алгоритм для сегментации объекта интереса на текущем кадре [6].

Алгоритм сегментации цветных изображений на основе кластеризации ISBC. Название алгоритма является сокращением слов *Interactive Segmentation by Clustering*. Для описания ISBC обозначим через $\mathbf{I}$ RGB-изображение, определенное на множестве пикселов $S = \{(x, y)\}$, $x = 0,..., w-1,\ y = 0,..., h-1$, и принимающее значения $I(\mathbf{p}) = (I_R(\mathbf{p}), I_G(\mathbf{p}), I_B(\mathbf{p}))$, $\mathbf{p} = (x, y) \in S$.

Под фоновой частью изображения будем понимать множество пикселов, не принадлежащих ни одному из объектов, предназначенных для сегментации.

Интерактивная часть алгоритма заключается в выделении одной или нескольких частей изображения, целиком принадлежащих объектам, которые предназначены для сегментации, и отдельно – областей изображения, целиком принадлежащих фону. При этом достаточно выделить лишь некоторое количество пикселов (не все), принадлежащих объектам, и некоторое количество пикселов, принадлежащих фону. Пример выделения наборов пикселов, принадлежащих объектам и фону, достаточных для успешной работы алгоритма, приведен на рис. 1. Красными прямоугольниками выделяются области объектов, а синими – области фона.

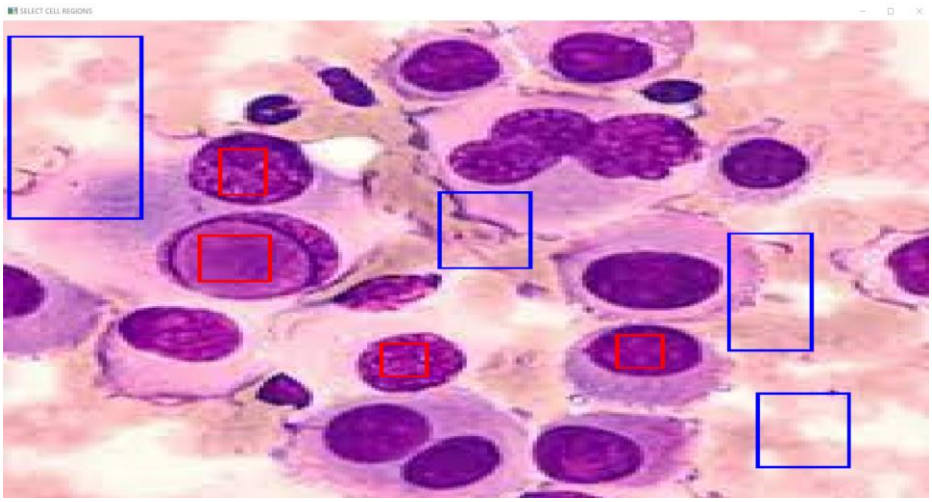


Рис. 1. Пример интерактивного выделения пикселов объектов и фона

Fig. 1. An example of interactive pixels object and background selection

Обозначим значения цветов выделенных пикселов, принадлежащих объектам, через $\mathbf{v}(j)=(v_R(j),v_G(j),v_B(j))$, а множество цветов $\mathbf{v}(j)$ – через A_1. Аналогично обозначим цвета выделенных пикселов фона через $\mathbf{u}(j)=(u_R(j),u_G(j),u_B(j))$, а множество цветов $\mathbf{u}(j)$ – через A_2.

Алгоритм ISBC можно представить в виде следующих последовательно выполняющихся блоков:

– кластеризации множества A_1 цветов выделенных областей объектов и отдельно – множества A_2 цветов выделенных областей фона одним из известных алгоритмов кластеризации, например *k*-средних, нечетких *c*-средних, или предложенным автором алгоритмом многоуровневой кластеризации [7] с сохранением центров кластеров объектов $C=\{\mathbf{c}_1,\mathbf{c}_2,...,\mathbf{c}_m\}$ и центров кластеров фона $B=\{\mathbf{b}_1,\mathbf{b}_2,...,\mathbf{b}_n\}$, представляющих собой *3d*-векторы цвета;

– сегментации множеств A_1 и A_2 на основе вычисления минимального расстояния векторов этих множеств до векторов общего множества $D=B\bigcup C$ всех центров кластеров;

– построения множества $\tilde{C}$ путем удаления из множества C центров кластеров $\mathbf{c}_k$, для которых каждый вектор $\mathbf{v}$ из A_1 расположен ближе к какому-либо центру из C, отличному от $\mathbf{c}_k$, и построения множества $\tilde{B}$ путем удаления из множества B центров кластеров $\mathbf{b}_\ell$, для которых каждый вектор $\mathbf{u}$ из A_2 расположен ближе к какому-либо центру из B, отличному от $\mathbf{b}_\ell$, а также удаления из C и B центров, дающих слишком большое число ошибок при сегментации A_1 и A_2;

– сегментации всего изображения на основе вычисления минимального расстояния до элементов общего множества центров кластеров $\tilde{D}=\tilde{B}\bigcup\tilde{C}$. Если цвет $I(\mathbf{p})=(I_R(\mathbf{p}),I_G(\mathbf{p}),I_B(\mathbf{p}))$ пиксела $\mathbf{p}$ расположен ближе всего к какому-либо элементу из множества $\tilde{C}$, этот пиксел считается принадлежащим области объектов, а если ближе всего к какому-либо элементу из множества $\tilde{B}$, то принадлежащим фону.

Входными параметрами алгоритма помимо самого RGB-изображения являются: количество кластеров n_{cell}, используемых для кластеризации выделенных в интерактивном режиме областей объектов; количество кластеров n_{bgrnd}, используемых для кластеризации выделенных в интерактивном режиме областей фона; вещественный параметр τ $(0<\tau\le 1)$, задающий максимально допустимую относительную погрешность классификации на основе кластеризации выделенных в интерактивном режиме областей A_1 и A_2.

Приведем подробное описание алгоритма:

Шаг 0. Чтение и предобработка RGB-изображения I, выбранного для сегментации.

Шаг 1. Выделение на I в интерактивном режиме нескольких областей, например прямоугольных, полностью принадлежащих изображениям объектов, которые предназначены для сегментации. Формирование массива A_1 цветов пикселов выделенных областей (см. рис. 1).

Шаг 2. Выделение на I в интерактивном режиме нескольких областей, например прямоугольных, полностью принадлежащих фону (не принадлежащих ни одному из объектов, предназначенных для сегментации). Формирование массива A_2 цветов пикселов выделенных областей (см. рис. 1).

Шаг 3. Кластеризация массива A_1 например, методом *k*-средних (или методом нечетких *c*-средних или алгоритмом многоуровневой кластеризации [7]). Число кластеров n_{cell} должно быть параметром алгоритма. Сохранение векторов $\mathbf{c}_j$ – центров кластеров – в виде множества C.

Шаг 4. Вычисление и занесение в элементы t_j^c вектора $\mathbf{t}^c$ размерностью n_{cell} количества векторов цвета $\mathbf{v} \in A_1$, которые из всех элементов множества C находятся ближе всего к $\mathbf{c}_j$. Иными словами, занесение в элементы t_j^c количества векторов $\mathbf{v} \in A_1$, для которых выполняется равенство

$$\min_{\mathbf{c} \in C} \|\mathbf{v} - \mathbf{c}\| = \|\mathbf{v} - \mathbf{c}_j\|.$$

Шаг 5. Кластеризация массива A_2. Число кластеров n_{bgrnd} должно быть параметром алгоритма. Сохранение векторов $\mathbf{b}_j$ – центров кластеров – в виде множества B.

Шаг 6. Вычисление и занесение в элементы t_j^b вектора $\mathbf{t}^b$ размерностью n_{bgrnd} количества векторов цвета $\mathbf{u} \in A_2$, которые из всех элементов множества B находятся ближе всего к $\mathbf{b}_j$. Иными словами, занесение в элемент t_j^b количества векторов $\mathbf{u} \in A_2$, удовлетворяющих условию

$$\min_{b \in B} \|\mathbf{u} - \mathbf{b}\| = \|\mathbf{u} - \mathbf{b}_j\|.$$

Шаг 7. Формирование множества $\tilde{C}$, в которое включаются лишь те центры кластеров $\mathbf{c}_j$, для которых $t_j^c > 0$ (в множество $\tilde{C}$ включаются лишь те $\mathbf{c}_j$ из C, для которых найдется хотя бы один вектор цвета $\mathbf{v} \in A_1$, находящийся ближе всего к нему по норме). Формирование вектора $\tilde{\mathbf{t}}^c$ путем выбора из вектора $\mathbf{t}^c$ координат, удовлетворяющих условию $t_j^c > 0$.

Шаг 8. Формирование множества $\tilde{B}$, в которое включаются лишь те центры кластеров $\mathbf{b}_j$, для которых $t_j^b > 0$ (в множество $\tilde{B}$ включаются лишь те $\mathbf{b}_j$, для которых найдется хотя бы один вектор цвета $\mathbf{u} \in A_2$, находящийся ближе всего к нему по норме). Формирование вектора $\tilde{\mathbf{t}}^b$ путем выбора из вектора $\mathbf{t}^b$ координат, удовлетворяющих условию $t_j^b > 0$.

Шаг 9. Формирование вектора $\tilde{\mathbf{e}}^b$ с элементами e_j^b, равными количеству векторов $\mathbf{u} \in A_2$, для которых верно равенство

$$\min_{\mathbf{w} \in \tilde{C} \cup \tilde{B}} \|\mathbf{w} - \mathbf{u}\| = \|\mathbf{c}_j - \mathbf{u}\|, \text{ где } \mathbf{c}_j \in \tilde{C}.$$

Элементы e_j^b вектора $\tilde{\mathbf{e}}^b$ равны количеству векторов цвета $\mathbf{u}$ фона, которые при кластеризации множества векторов A_2 с помощью центров всех кластеров $\tilde{C} \cup \tilde{B}$ будут ошибочно распознаны как цвета объектов.

Шаг 10. Формирование вектора $\tilde{\mathbf{e}}^c$ с элементами e_j^c, равными количеству векторов $\mathbf{v} \in A_1$, для которых выполняется условие

$$\min_{\mathbf{w} \in \tilde{C} \cup \tilde{B}} \|\mathbf{w} - \mathbf{v}\| = \|\mathbf{b}_j - \mathbf{v}\|.$$

Элементы e_j^c вектора $\tilde{\mathbf{e}}^c$ равны количеству векторов цвета $\mathbf{v}$ объектов, которые при кластеризации множества векторов A_1 с помощью центров всех кластеров $\tilde{C} \cup \tilde{B}$ будут ошибочно распознаны как цвета фона.

Шаг 11. Формирование подмножества центров кластеров $\hat{C} \subset \tilde{C}$, которое будет использоваться для классификации всего изображения путем поэлементного просмотра $\tilde{C}$, и включение в $\hat{C}$ только $\mathbf{c}_j \in \tilde{C}$, удовлетворяющих условию $e_j^c / t_j^c \le \tau$, где e_j^c – координаты вектора $\tilde{\mathbf{e}}^c$, а t_j^c – вектора $\tilde{\mathbf{t}}^c$. Иными словами, образование множества

$$\hat{C} = \left\{ \mathbf{c}_j \middle| \mathbf{c}_j \in \tilde{C},\ e_j^c / t_j^c \le \tau \right\}.$$

Шаг 12. Формирование подмножества центров кластеров $\hat{B} \subset \tilde{B}$, которое будет использоваться для классификации всего изображения, путем поэлементного просмотра $\tilde{B}$ и включение в $\hat{B}$ только $\mathbf{b}_j \in \tilde{B}$, удовлетворяющих условию $e_j^b / t_j^b \le \tau$, где e_j^b – координаты вектора $\tilde{\mathbf{e}}^b$, а t_j^b – вектора $\tilde{\mathbf{t}}^b$. Иными словами, образование множества

$$\hat{B} = \left\{ \mathbf{b}_j \middle| \mathbf{b}_j \in \tilde{B},\ e_j^b / t_j^b \le \tau \right\}.$$

Шаг 13. Классификация исходного изображения путем его кластеризации по двум множествам кластеров – $\hat{C}$ и $\hat{B}$. Построение 2*d*-бинарной маски $J = J(p)$, $p \in S$, размер которой совпадает с размером исходного изображения I, следующим образом:

$J(p) = 1$, если $\min\limits_{\mathbf{w} \in \hat{C} \cup \hat{B}} \left\| \mathbf{w} - I(p) \right\| = \left\| \mathbf{c}_j - I(p) \right\|$ для некоторого элемента $\mathbf{c}_j \in \hat{C}$,

и

$J(p) = 0$, если $\min\limits_{\mathbf{w} \in \hat{C} \cup \hat{B}} \left\| \mathbf{w} - I(p) \right\| = \left\| \mathbf{b}_k - I(p) \right\|$ для некоторого элемента $\mathbf{b}_k \in \hat{B}$.

Шаг 14. Построение результирующего RGB-изображения $R = (R_R, R_G, R_B)$, на котором область, соответствующая объектам, будет иметь оригинальные цвета, а область, соответствующая фону, – черный цвет, по формуле $R = (JI_R, JI_G, JI_B)$.

Результат сегментации изображения клеток (см. рис. 1) показан на рис. 2. Изображение не подвергалось постобработке.

Результаты экспериментов и их обсуждение. Для исследования характеристик разработанного алгоритма ISBC было проведено 148 вычислительных экспериментов на 74 изображениях. Для тестирования выбраны изображения, имеющие отличающиеся цветовые, текстурные и семантические характеристики. Среди них 52 микроскопических изображения (104 эксперимента) различных типов: клеток и ткани человека, а также животных; 10 изображений (20 экспериментов) лиц людей; 10 изображений (20 экспериментов) цветов (растений) и два аэрофотоснимка (четыре эксперимента) городской застройки, на которых выделялись крыши домов. Были использованы микроскопические изображения клеток и тканей различных типов из наборов данных IDR[3] и Bialystok Outline Annotated Cervical Cytology Dataset[4].

Каждое изображение применялось для тестирования первый раз с параметром $\tau = 100$ или $\tau = 200$, а второй раз – с параметром $\tau = 0$. В 103 экспериментах из 104 результат сегментации с ненулевыми параметрами $\tau = 100$ или $\tau = 200$ был лучше, чем с параметром $\tau = 0$.

[3]Image Data Resource [Electronic resource]. – Mode of access: https://idr.openmicroscopy.org/. – Date of access: 07.05.2024.

[4]Nalecz Institute of Biocybernetics and Biomedical Engineering Polish Academy of Sciences [Electronic resource]. – Mode of access: https://ibib.pl/en/211%20boacd/1243%20boacdeng%20#%20:~:text%20=The%20Bialystok%20Dataset%20%%2020contains%%2020162,WSI)%20%20of%20routine%20cervical%20smears. – Date of access: 07.05.2024.

Количество кластеров $\hat{C}$ и $\hat{B}$, использованных для сегментации при ненулевых значениях параметра τ, было в среднем на 12 % меньше, чем в случае $\tau = 0$.

Все микроскопические изображения клеточных структур были отсегментированы корректно – на них были выделены от 95 до 100 % клеток и тканей, запланированных для сегментации по цвету.

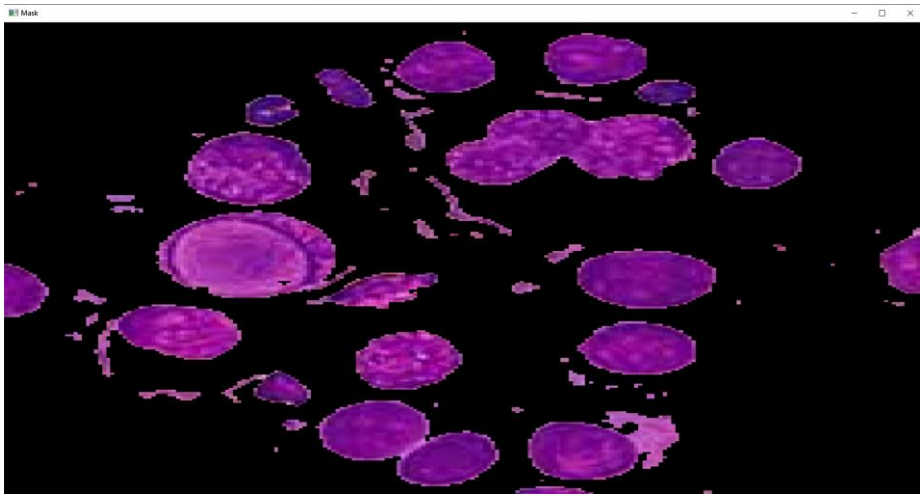


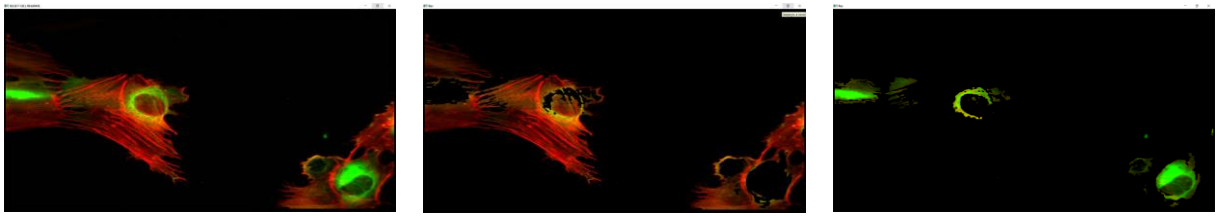
Рис. 2. Результат сегментации ядер клеток алгоритмом ISBC с использованием интерактивного выделения областей, показанного на рис. 1

Fig. 2. The result of the cell nuclei segmentation by the ISBC algorithm using interactive areas selection shown in Fig. 1

На изображениях лиц людей и цветов результаты были ожидаемо хуже в случаях, когда на них присутствовали области, принадлежащие объектам и фону и имеющие совпадающие или близкие цветовые характеристики. Из 10 изображений лиц пять были отсегментированы с ошибками первого и второго рода, не превосходящими 5 %, на четырех – каждая из ошибок первого и второго рода не превосходила 10 %, на одном – упомянутые ошибки достигали 40 %. На восьми из 10 изображений цветов ошибки первого и второго рода не превосходили 5 %, на двух остальных не превосходили 10 %. На двух цветных аэрофотоснимках городской застройки корректно были отсегментированы по цвету крыши зданий.

Эксперименты подтверждают теоретическое предположение о возможности интерактивной сегментации изображений по цвету алгоритмом ISBC в случае, если цветовые характеристики объектов, предназначенных для сегментации, и фона различны. Иными словами, если множество векторов цвета объектов, предназначенных для сегментации, и множество векторов цвета фона в цветовом RGB-кубе не пересекаются.

На рис. 3 показано изображение и выделенные на нем алгоритмом ISTC объекты различных типов.



a) *b)* *c)*

Рис. 3. Исходное изображение (*a*); клеточная структура красных оттенков, выделенная предложенным алгоритмом (*b*); части клеток, окрашенные зеленым цветом, отсегментированные ISTC (*c*)

Fig. 3. The original image (a); a red cellular structure, selected by the proposed algorithm (b); the green parts of cells segmented by the ISBC (c)

Заключение. В статье предложен алгоритм ISBC интерактивной сегментации цветных изображений, преимуществом которого является отсутствие необходимости его предварительного обучения. Это дает возможность, во-первых, сегментировать объекты, имеющие заранее неизвестные цвета, а во-вторых, в течение нескольких минут выделять на одном изображении несколько типов объектов в случае, если они имеют отличающиеся цветовые характеристики.

В дальнейшем планируется усовершенствовать алгоритм так, чтобы при сегментации изображений учитывались не только цветовые, но и пространственные характеристики объектов.

Список использованных источников

1. Гонсалес, Р. Цифровая обработка изображений : пер. с англ. / Р. Гонсалес, Р. Вудс. – М. : Техносфера, 2005. – 1075 с.

2. Шапиро, Л. Компьютерное зрение : пер. с англ. / Л. Шапиро, Дж. Стокман. – М. : Бином, 2006. – 752 с.

3. Селянкин, В. В. Компьютерное зрение. Анализ и обработка изображений / В. В. Селянкин. – СПб. : Лань, 2019. – 152 с.

4. Snyder, W. E. Fundamentals of Computer Vision / W. E. Snyder, H. Qi. – Cambridge : Cambridge University Press, 2017. – 386 p.

5. Bezdek, J. C. Pattern Recognition with Fuzzy Objective Function Algorithms / J. C. Bezdek. – N. Y. : Springer New York, 1981. – 272 p.

6. Залесский, Б. А. Алгоритм отслеживания объектов движущейся видеокамерой / Б. А. Залесский // Докл. Нац. акад. наук Беларуси. – 2020. – Т. 64, № 2. – С. 144–149.

7. Залесский, Б. А. Многоуровневый алгоритм цветовой кластеризации изображений / Б. А. Залесский // Докл. Нац. акад. наук Беларуси. – 2021. – Т. 65, № 3. – С. 209–274.

References

1. Gonzales R. C., Woods R. E. *Digital Image Processing*. Upper Saddle River, New Jersey, Prentice Hall, 2002, 814 p.

2. Shapiro L. S., Stockman G. C. *Computer Vision*. Upper Saddle River, New Jersey, Prentice Hall, 2001, 608 p.

3. Selyankin V. V. Komp'juternoe zrenie. Analiz i obrabotka izobrazhenij. *Computer Vision. Image Analysis and Processing*. Saint Petersburg, Lan', 2019, 152 p. (In Russ.).

4. Snyder W. E., Qi H. *Fundamentals of Computer Vision*. Cambridge, Cambridge University Press, 2017, 386 p.

5. Bezdek J. C. *Pattern Recognition with Fuzzy Objective Function Algorithms*. New York, Springer New York, 1981, 272 p.

6. Zalesky B. A. *Object tracking algorithm by moving video camera*. Doklady Nacional′noj akademii nauk Belarusi [*Doklady of the National Academy of Sciences of Belarus*], 2020, vol. 64, no. 2, pp. 144–149 (In Russ.).

7. Zalesky B. A. *Multilevel algorithm for color clustering of images*. Doklady Nacional′noj akademii nauk Belarusi [*Doklady of the National Academy of Sciences of Belarus*], 2021, vol. 65, no. 3, pp. 209–274 (In Russ.).

Информация об авторе

Залесский Борис Андреевич, доктор физико-математических наук, Объединенный институт проблем информатики Национальной академии наук Беларуси.
E-mail: zalesky@newman.bas-net.by

Information about the author

Boris A. Zalesky, D. Sc. (Phys.-Math.), The United Institute of Informatics Problems of the National Academy of Sciences of Belarus.
E-mail: zalesky@newman.bas-net.by

УДК 004; 004.932
https://doi.org/10.37661/1816-0301-2024-21-2-94-106

Оригинальная статья
Original Article

Верификация динамической подписи человека по ограниченному числу образцов

В. В. Старовойтов

Объединенный институт проблем информатики
Национальной академии наук Беларуси,
ул. Сурганова, 6, Минск, 220012, Беларусь
E-mail: valerys@newman.bas-net.by

Аннотация

Цели. Целью исследования является разработка нового метода оценки подлинности подписи одного человека, выполненной на планшете стилусом, при наличии ограниченного числа образцов подписей этого человека.

Методы. Показано, как строить индивидуальный образ динамических подписей произвольного человека, который описывается точками в многомерном признаковом пространстве и предназначен для последующих проверок подлинности подписей данного человека. Образ строится по $5<N<20$ образцам подлинных подписей человека. Он представляет собой выпуклую фигуру в многомерном признаковом пространстве и описывает индивидуальные признаки выполнения подписи конкретным человеком.

Результаты. Динамика исполнения подписи представлена тремя дискретными параметрическими функциями: координатами стилуса X, Y и его давлением на планшет P, зарегистрированными через фиксированные промежутки времени. В процессе исследований отобран ряд вычисляемых по ним вторичных функций-признаков. Поскольку эти массивы данных имеют разную длину, для их сравнения используется алгоритм динамической трансформации временно́й шкалы. Результатами данного преобразования являются расстояния между динамическими признаками двух подписей, которые служат координатами точки в признаковом пространстве, описывающей сходство этих подписей. Множество таких точек описывает сходство всех пар подлинных подписей человека, предъявленных для верификации, в многомерном признаковом пространстве. Выпуклая оболочка облака этих точек используется как образ подписи конкретного человека. Подлинные подписи любого человека всегда отличаются друг от друга, существенные отличия между ними могут исказить результат верификации.

Заключение. Экспериментальные исследования по формированию индивидуальных образов подписей 498 человек из базы динамических подписей DeepSignDB показали точность верификации порядка 98 % при анализе 24 900 подписей. Из них половина подлинные, половина поддельные.

Ключевые слова: верификация, динамическая подпись, *dtw*-преобразование, параметрические функции, признаковое пространство

Для цитирования. Старовойтов, В. В. Верификация динамической подписи человека по ограниченному числу образцов / В. В. Старовойтов // Информатика. – 2024. − Т. 21, № 2. – С. 94–106.
https://doi.org/10.37661/1816-0301-2024-21-2-94-106

Конфликт интересов. Автор заявляет об отсутствии конфликта интересов.

Поступила в редакцию | Received 08.05.2024
Подписана в печать | Accepted 24.05.2024
Опубликована | Published 28.06.2024

Verification of the person's dynamic signature on a limited number of samples

Valery V. Starovoitov

The United Institute of Informatics Problems
of the National Academy of Sciences of Belarus,
st. Surganova, 6, Minsk, 220012, Belarus
E-mail: valerys@newman.bas-net.by

Abstract

Objectives. The goal of the research is to develop a new person-dependent method for verification of a signature of one person made on a tablet with a stylus in the presence of a limited number of signature samples of this person.

Methods. The paper shows how to construct an individual pattern of the dynamic signatures of any person, which is described by points in a multidimensional feature space and is intended for subsequent verification of the authenticity of the signatures of a given person. It is constructed using $5<N<20$ samples of genuine human signatures. The pattern forms a convex object in a multidimensional feature space. It describes the peculiar properties of a signature performed by a specific person.

Results. The dynamics of signature execution is represented by three discrete parametric functions: coordinates of the stylus *X*, *Y* and its pressure on the tablet *P*, recorded at fixed time intervals. In the process of research, a number of secondary functions-features were selected and calculated from them. Since these data sets have different lengths, the dynamic time warping algorithm is used to compare them. The results of this transformation are distances between the dynamic features of two signatures, which serve as coordinates of a point in the feature space that describes the similarity of these signatures. The set of such points describes similarity of all pairs of genuine human signatures presented for verification in a multidimensional feature space. The convex hull of the cloud of these points is used as a pattern of a particular person's signature. The genuine signatures of any person are always different from each other; significant differences between them can distort the verification result.

Conclusion. Experimental studies performed on genuine and fake signatures of 498 people from the largest available database of dynamic signatures, DeepSignDB, showed a verification accuracy of about 98 % when analyzing 24,900 signatures. Half of them are genuine, half are fake.

Keywords: verification, dynamic signature, *dtw* transformation, parametric functions, feature space

For citation. Starovoitov V. V. *Verification of the person's dynamic signature on a limited number of samples*. Informatika [*Informatics*], 2024, vol. 21, no. 2, pp. 94−106 (In Russ.).
https://doi.org/10.37661/1816-0301-2024-21-2-94-106

Conflict of interest. The author declares of no conflict of interest.

Введение. Рукописная подпись удостоверяет личность человека, подписавшего некий документ, либо заверяет подписанный документ. С точки зрения информационных технологий распознавание подписи относится к биометрическим технологиям, которые различаются при анализе подписей, выполненных разными способами: статических (выполненных на бумаге) и динамических, или онлайн (выполненных на планшете). Динамическая подпись может быть представлена изображением, построенным в виде кривых (чем сводится к статической подписи), но она содержит дополнительные данные о динамике исполнения подписи, которые невидимы и делают подпись более уникальной. Динамические характеристики имеют больше степеней свободы и точнее характеризуют особенности исполнения подписи конкретным человеком, поэтому они лучше защищены от подделок.

Следует отличать динамическую подпись от электронной цифровой подписи (ЭЦП), введенной в нашей стране Законом Республики Беларусь от 28 декабря 2009 г. № 113-З «Об электронном документе и электронной цифровой подписи». Основным элементом ЭЦП является криптографический ключ. Этот тип подписи используют в основном юридические лица. С 3 марта 2018 г. постановлением Правления Национального банка Республики Беларусь (URL: https://pravo.by/novosti/novosti-pravo-by/2018/march/27952/) разрешена к использованию цифро-

вая рукописная подпись, она и является динамической, или онлайн, подписью. Согласно указанному постановлению цифровая рукописная подпись – это собственноручная подпись клиента, учиненная с помощью соответствующих программных средств (в том числе планшетов) для подтверждения целостности и подлинности подписываемого документа в электронном виде. Официальных методик исследования экспертами подлинности таких подписей в Республике Беларусь до настоящего времени нет, а в России отсутствует даже официально принятое понятие цифровой рукописной подписи.

Система признаков, используемых на данный момент экспертами-почерковедами, была разработана еще в советское время и на сегодняшний день не особо претерпела существенных изменений [1]. В статье [2] отмечено, что методические основы отечественной почерковедческой экспертизы заложены в 60–70-х гг. прошлого столетия и что в настоящее время общая методика проведения почерковедческого исследования особых изменений не претерпела. «Используя одну и ту же методику проведения идентификационной почерковедческой экспертизы, разные эксперты могут по-разному оценить выявленные совпадения и различия. В результате по одному и тому же исследуемому объекту могут быть сделаны совершенно противоположные (иногда категоричные) выводы. Во многом это связано с тем, что используемые идентификационные признаки в большинстве своем носят качественный, оценочный характер и формируются на основе субъективной оценки эксперта» [2]. Эксперты-юристы все еще используют визуальный анализ графического представления цифровых подписей, а он является субъективным.

Приведенные факты свидетельствуют об актуальности разработки интеллектуальных систем проверки подлинности динамических подписей для повышения объективности анализа, выполняемого экспертом. Во многих странах ведутся активные исследования, направленные на поиск решения задачи верификации (проверки подлинности) динамической подписи [3]. На сегодняшний день исследуется возможность применения разных методов для разработки систем верификации динамических подписей. Наиболее популярными подходами являются динамическое искажение временной шкалы (Dynamic Time Warping, DTW), скрытая марковская модель (Hidden Markov Models, HMM), искусственные нейронные сети и метод опорных векторов (Support Vector Machine, SVM) [3].

1. Представление динамических подписей. Подпись, зарегистрированная на планшете с помощью специального стилуса, представляет собой несколько дискретных параметрических функций. Обязательными из них являются: координаты X и Y положения стилуса, время фиксации этих координат T и давление стилуса на поверхность планшета P в каждой точке. Точки нахождения кончика стилуса фиксируются через равные промежутки времени. У разных производителей они составляют 5, 10 или 15 мс. Некоторые типы планшетов дополнительно регистрируют две угловые координаты положения стилуса. Все параметры чаще всего представлены целыми числами в определенной шкале в текстовом формате. Некоторые производители не сохраняют координаты стилуса, перемещаемого без давления на планшет, а только отмечают нулями места разрывов дискретных функций, т. е. начала и окончания сегментов подписи. На рис. 1 приведены примеры визуального разнообразия подлинных динамических подписей одного человека, при визуализации разные сегменты подписи показаны разным цветом. Справа внизу представлена поддельная подпись этого же человека.

При анализе подлинности цифровых подписей в экспертной практике чаще всего используется их представление в виде графиков дискретных параметрических функций X, Y, P (рис. 2). Некоторые планшеты регистрируют углы наклона стилуса, но в настоящей работе эти данные не применяются. Функции являются параметрическими, поскольку значения каждой из них зависят только от момента регистрации t.

2. Постановка задачи. Задачу верификации динамической подписи, ориентированную на практическое применение, можно сформулировать следующим образом. Даны N ($5<N<20$) подлинных динамических подписей некоторого человека и одна подпись, подлежащая проверке. Требуется разработать метод автоматического вычисления объективной оценки сходства верифицируемой подписи для определения ее подлинности. Эта оценка должна помочь эксперту принять более объективное решение при исследовании подлинности динамической подписи.

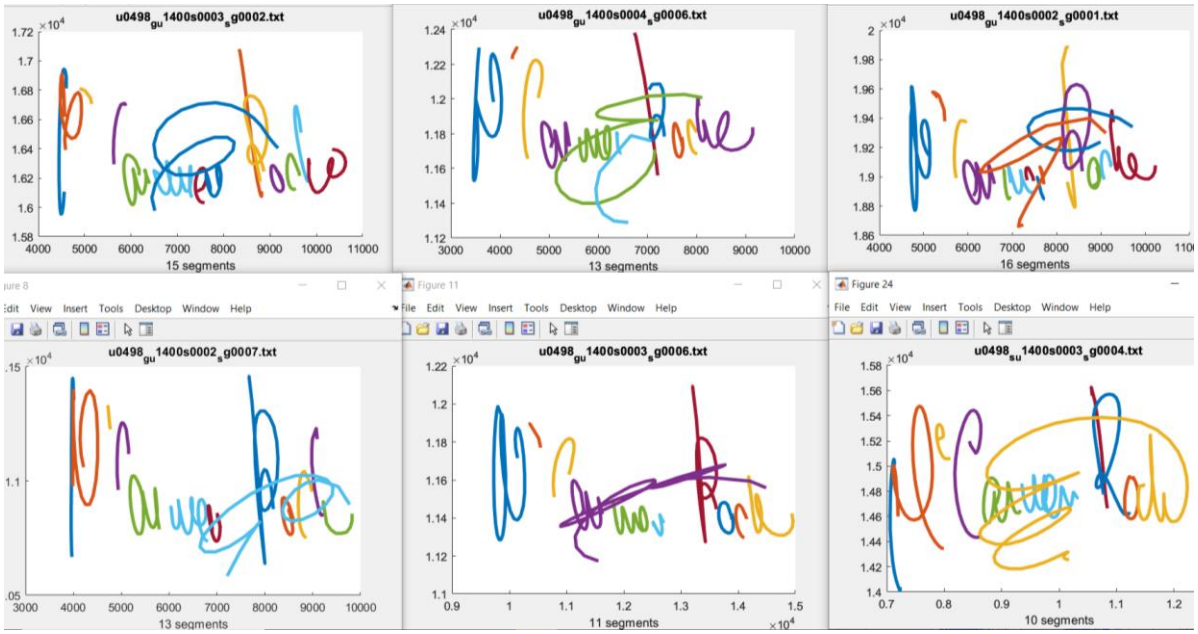


Рис. 1. Визуализация динамических подписей

Fig. 1. Visualization of dynamic signatures

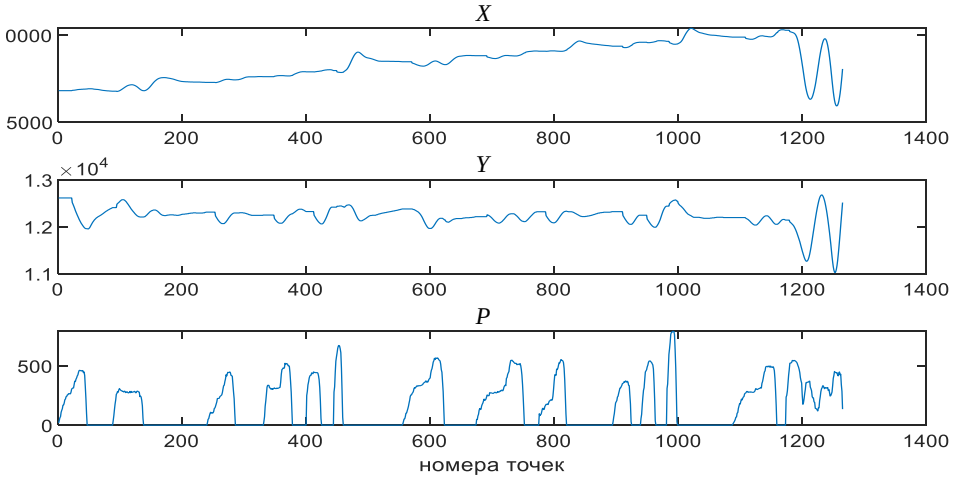


Рис. 2. Основные данные о динамической подписи, представленные в виде дискретных параметрических функций *X, Y, P*

Fig. 2. Basic data on dynamic signature are presented in the form of discrete parametric functions X, Y, P

Для решения поставленной задачи необходимо построить образ динамической подписи человека, который численно описывает индивидуальные особенности исполнения подписи этим человеком и учитывает вариабельность цифрового представления подписи. Он должен позволять выполнять проверку подлинности других подписей этого человека и выявлять поддельные подписи.

Следует учесть, что все подписи одного человека различаются числом точек, особенно когда они выполняются с разной скоростью и на планшетах разных типов. Кроме того, подписи одного человека неизбежно имеют локальные вариации формы (рис. 3).

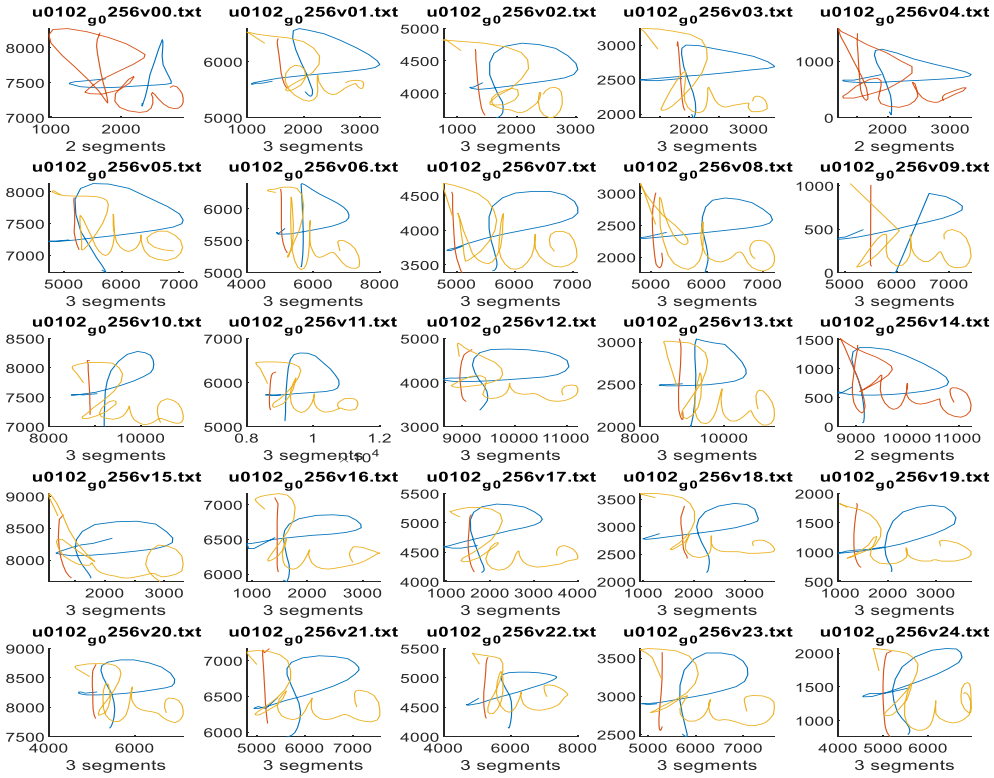


Рис. 3. Подлинные подписи человека с идентификатором U0102
(разным цветом показаны различные сегменты подписей)

Fig. 3. Genuine signatures of a person with ID U0102
(different colors show different segments of the signatures)

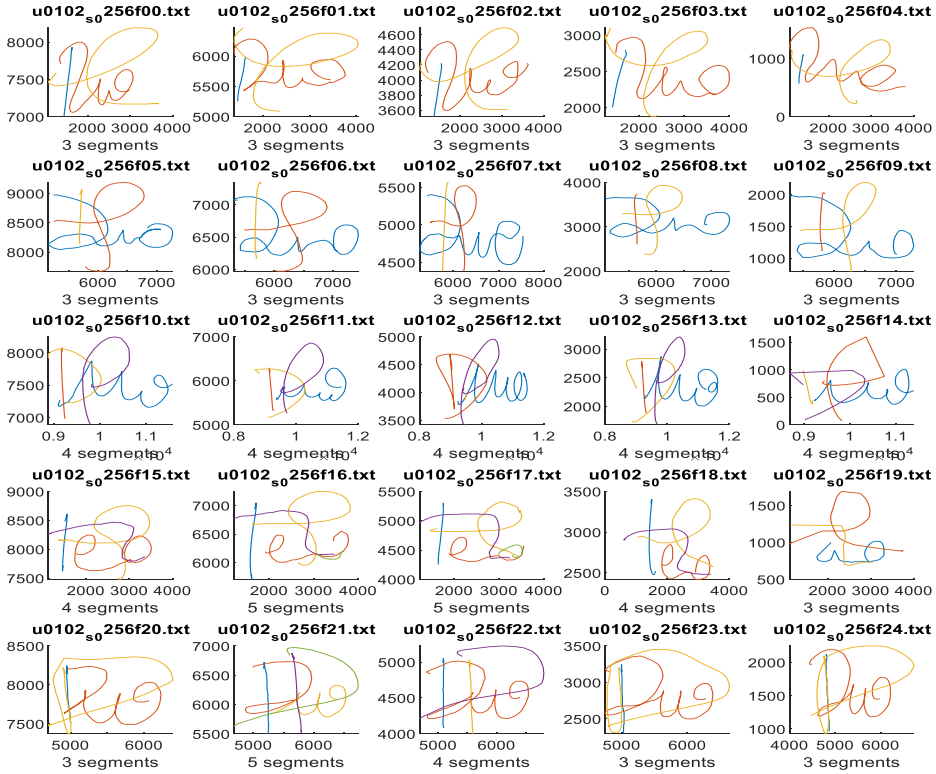


Рис. 4. Фальшивые подписи человека с идентификатором U0102

Fig. 4. Fake signatures of a person with ID U0102

3. Особенности выполнения динамической подписи. Все подписи, сделанные одним человеком даже на одном планшете, всегда имеют разную длину (число точек), геометрические размеры и ориентацию на плоскости *XY*, несколько отличающиеся форму и динамику исполнения (рис. 3 и 4).

Авторы статьи [4] выполнили исследования по анализу зависимости результатов *dtw*-преобразований от пространственной нормализации данных (приведения функций к фиксированному числу значений), примененных к данным из четырех разных баз. Они доказали, что двусторонний тест с уровнем значимости 0,05 между парами параметрических функций переменной и равной длины показывает отсутствие статистически значимой разницы между точностью вычисления *dtw*-расстояний.

В работе [5] сравнивались четыре различных типа нормализации данных динамических подписей: отсутствие нормализации, нормализация по времени, масштабирование амплитуды давления *P* до диапазона [0; 1] и нормализация по времени с последующей нормализацией давления *P*. Нормализация была применена до выполнения *dtw*. Авторы выделили 15 лучших вариантов нормализованных признаков с наименьшими ошибками верификации EER. Почти все из 15 вариантов связаны со скоростью.

Таким образом, динамические подписи хотя и похожи после визуализации на привычные статические подписи, сделанные на бумаге, но в статическом представлении теряется вся информация о динамике исполнения подписи. Между тем динамические признаки существенно увеличивают число степеней свободы, что при малом числе образцов подлинных подписей усложняет задачу верификации, а собирать и хранить их в большом количестве проблематично.

В настоящей статье основными признаками динамической подписи считаются те, которые вычисляются локально, по соседним отсчетам функций *X*, *Y* и *P*. В работе были использованы избранные признаки, описанные в статьях [6, 7] (табл. 1).

Таблица 1
Динамические признаки подписи

Table 1
Dynamic signature features

Функция *Function*	Формула *Formula*
Горизонтальная скорость, *dx*	$dx_i = x_{i+1} - x_i$
Вертикальная скорость, *dy*	$dy_i = y_{i+1} - y_i$
Абсолютная скорость, *dxy*	$dxy_i = \sqrt{dx_i^{\,2} - dy_i^{\,2}}$
Скорость изменения давления, *dp*	$dp_i = p_{i+1} - p_i$
Горизонтальное ускорение, *ddx*	$ddx_i = dx_{i+1} - dx_i$
Вертикальное ускорение, *ddy*	$ddy_i = dy_{i+1} - dy_i$
Абсолютное ускорение, *ddxy*	$ddxy_i = dxy_{i+1} - dxy_i$
Ускорение изменения давления, *ddp*	$ddp_i = dp_{i+1} - dp_i$
Центростремительное ускорение, *ac*	$ac_i = ((x_{i+3} - x_{i+2})(y_{i+3} + y_{i+4} - 2y_{i+2}) - -(y_{i+3} - y_{i+2})(x_{i+3} + x_{i+4} - 2x_{i+2}))/8$
Рывок по вертикали (третья производная), *dddy*	$dddy_i = ddy_{i+1} - ddy_i$
Абсолютный рывок (третья производная), *dddxy*	$dddxy_i = ddxy_{i+1} - ddxy_i$
Векторное произведение, *s*	$s_i = x_i y_{i+2} + y_i x_{i+2}$

Окончание табл. 1
End of table 1

Функция *Function*	Формула *Formula*
v	$v_i = \sin(ang_i) dxy_i dxy_{i+1}$
*d*2	$d2_i = x_i y_{i+1} - x_{i+1} y_i$
Угол между соседними точками, *ang*	$ang_i = \arctan(x_{i+1}, y_{i+1}) - \arctan(x_i, y_i)$
Косинус угла между двумя точками, cos*XY*	$\cos XY_i = \frac{dx_i}{\sqrt{dx_i^2 + dy_i^2}}$
Скорость изменения угла, *dang*	$dang_i = ang_{i+1} - ang_i$
Площадь треугольника из трех последовательных точек подписи, *area*	$area_i = ((x_i - x_{i+2})(y_{i+1} - y_{i+2}) - (x_{i+1} - x_{i+2})(y_i - y_{i+2}))/2$

Для данных каждой подписи вычисляются указанные признаки. Как описать образ подписи, представленной массивами признаков, длина которых различна? Одним из вариантов является нормализация длины массивов исходных данных *X, Y, P* до единого размера. Этот подход имеет два недостатка: массивы поддельных подписей часто длиннее, чем массивы оригинальных, и нормализация частично теряет данные о динамике подписи; размерность массивов равна сотням, что делает признаковое пространство с такой размерностью бесполезным для 7–20 образцов.

Решением проблемы является сравнение близости однотипных признаков между парами подписей. Подобный подход был успешно апробирован при верификации статических подписей в работе [8], но там использовались только два признака и вычислялась ранговая корреляция между гистограммами значений этих признаков. Гистограммы позволяли уменьшить размерность признакового описания до количества используемых интервалов. После экспериментов по сравнению коэффициентов корреляции гистограмм вышеуказанных признаков было решено использовать другой метод их сравнения, так как гистограммное представление теряет локальную динамику исполнения подписи.

Для сравнения массивов признаков был выбран алгоритм *dtw*, разработанный Беллманом [9]. Он позволяет вычислять расстояния между двумя дискретными кривыми, которые имеют разное число точек. По результатам международного конкурса алгоритмы на базе *dtw* были признаны лучшими для сравнения признаков динамических подписей [10].

В настоящей работе $dtw(dy^1, dy^2)$, например, обозначает расстояние, вычисленное между массивами вертикальных скоростей двух подписей (верхние индексы указывают на номера подписей). Все *dtw*-расстояния неотрицательны и вычисленные для признаков из табл. 1 могут использоваться как координаты, описывающие близость пар подписей вдоль осей в *K*-мерном признаковом пространстве.

Признаковое пространство, построенное на базе *dtw*-расстояний между парами массивов признаков, которые вычисляются для разных подписей, позволяет увеличивать количество образцов, описывающих *N* подлинных подписей человека, до N_2 образцов, но уже описывающих сходство пар подписей. По ним строится образ подлинных подписей одного человека в признаковом пространстве размерностью *K*. В данной статье использовалось $K = 18$. В табл. 2 показано, как возрастает число образов N_2 в признаковом пространстве при разных значениях доступных подлинных подписей *N*. Для сравнения приведено количество образцов подписей, используемых в общепринятых методах верификации.

Таблица 2
Число образцов при разном количестве доступных подлинных подписей

Table 2
Number of samples with different numbers of available genuine signatures

Общепринятая верификация *Common verification*		Попарная верификация *Pairwise verification*	
Верифицируемая подпись *Verifiable signature*	Подлинные подписи, *N* штук *Original signatures, N things*	Образцы пар с верифицируемой подписью, *N* штук *Sample pairs with a verifiable signature, N things*	Образцы пар подлинных подписей, $N_2=N(N-1)/2$ штук *Sample pairs of genuine signatures,* $N_2=N(N-1)/2$ *things*
1	7	7	21
1	10	10	45
1	15	15	105
1	19	19	171
1	24	24	276

Возникает задача описания образа подписей человека и определения критерия, позволяющего отличить его подлинные подписи от поддельных. Для этого следует построить область в признаковом пространстве, которая охватывает подписи конкретного человека. Самый простой вариант – построить гиперсферу, охватывающую множество *dtw*-расстояний между парами признаков подписей человека. Это простейшая выпуклая фигура, охватывающая все образцы в признаковом пространстве.

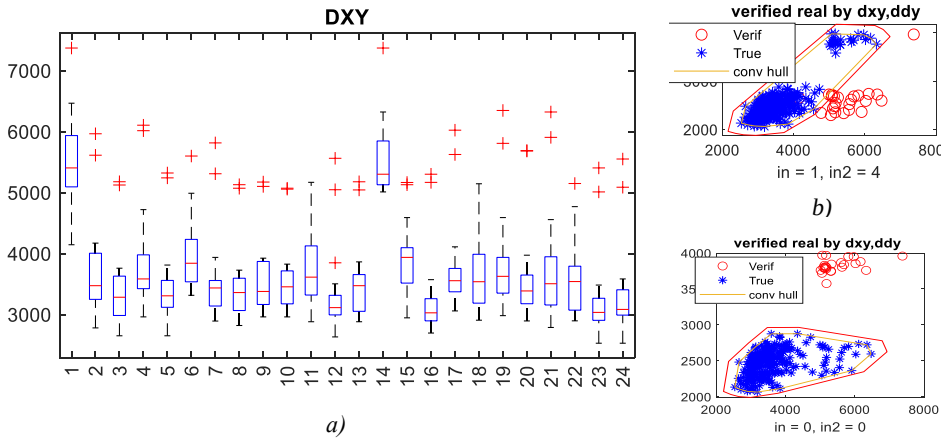
Был выполнен ряд экспериментов на динамических подписях из самой большой доступной базы данных DeepSignDB (URL: https://github.com/BiDAlab/DeepSignDB). Следует отметить, что данные подписей перед вычислением *dtw*-расстояний не подвергались нормализации, поскольку ранее в работе [11] было экспериментально доказано, что она снижает точность верификации.

Эксперименты показали, что совокупности значений некоторых признаков представляют собой очень узкие области, а *dtw*-расстояния между признаками разных типов имеют существенно различные диапазоны значений. Автором было принято решение использовать проекции признакового пространства на плоскости, определенные двумя признаковыми осями, и строить выпуклые множества, описывающие точки, которые представляют собой *dtw*-расстояния, соответствующие этим осям (рис. 5, *b*, *c*). Возможное число таких проекций равно $K(K-1)/2$. В данной работе число проекций ограничено числом $K/2$, причем для проекций использованы оси, образованные парами признаков, которые указаны в табл. 1.

Критерий подлинности определен следующим образом. Если новый образец, образованный парой (исходная подлинная подпись, верифицируемая подпись), попадает в большинство выпуклых множеств, образованных проекциями образа пар подлинных подписей в признаковом пространстве, принимается решение, что исследуемая подпись может считаться похожей на ту подлинную, с которой она сравнивалась. Так повторяется $K/2$ раз для всех проекций и верифицируемых пар подписей. Если бóльшая часть из них попадает M раз ($M>K/4$) в выпуклые множества проекций образа подлинной подписи, принимается решение о подлинности верифицируемой подписи. Иначе подпись считается поддельной.

Выбор наиболее значимых для верификации признаков можно выполнять на глобальном уровне и на локальном. В данной работе первоначально модель образа подписей человека строилась в 18-мерном признаковом пространстве. Далее на глобальном уровне оценивалось число ошибок верификации на большом количестве людей и их подписей из доступных баз данных. На рис. 6 показаны графики суммарных попаданий 2500 подлинных (синяя линия) и 2500 поддельных (красная линия) подписей 100 человек в выпуклые многоугольники, построенные на проекциях образов подписей каждого человека. Очевидно, что шестая и восьмая проекции дают гораздо больше ошибок при верификации поддельных подписей. По этой причине четыре признака, определяющие эти проекции, были исключены во второй модели обра-

за подлинных подписей произвольного человека, описанного в 14-мерном признаком пространстве.



a) *b)* *c)*

Рис. 5. Диаграммы размаха *dtw*-расстояний абсолютных скоростей *dxy* между каждой подлинной подписью и остальными 24 подписями человека с идентификатором U0184 (*a*); проекции образа подписей этого человека на плоскость (*dxy, ddy*), построенные с использованием 14-й подписи (*b*) и без нее (*c*)

Fig. 5. Box-plots of dtw-distances of absolute velocities dxy between each genuine signature and the remaining 24 signatures of a person with the IDr U0184 (a); projections of this person signature pattern onto the plane (dxy, ddy), constructed using the 14th signature (b) and without it (c)

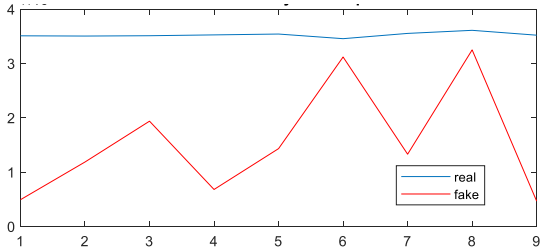


Рис. 6. Графики попадания образов подписей 100 человек (по 25 подлинных и фальшивых каждого) в девять проекций при N=15 (единицы измерений вертикальной шкалы десятки тысяч)

Fig. 6. Graphs of signature patterns of 100 people (25 genuine and fake each) falling into nine projections with N=15 (vertical scale units tens of thousands)

4. Уточнение образа подлинных подписей человека. Пусть имеется M подлинных подписей некоего человека. Тогда на локальном уровне можно оценить качество образа подписей этого человека, построенного по принципу сравнения одной подлинной подписи и модели образа, построенного по остальным (M–1) подлинным. Если признаки подписи существенно искажают образ подписи, ее не следует использовать для его построения. На рис. 5, *a* показано, что для M = 25 признак скорости написания 14-й подписи *dxy* человека с идентификатором U0184 существенно отличается от значений этого признака при сравнении других подписей между собой. На рис. 5, *b* показано, что такая подпись не попадает в проекцию образа подписей этого человека по данному признаку.

Для подписей каждого человека строился индивидуальный образ на базе признаков, извлеченных из N первых подлинных подписей. Полученные результаты позволяют сделать вывод о том, что для достижения высокой точности объективной верификации динамических подписей необходимо не менее 10 подлинных подписей одного человека. Анализируя признаки его подписи, следует удалять те подписи, которые существенно отличаются от остальных, и не использовать динамические признаки, которые чаще других не позволяют определять фальшивые подписи.

Из рис. 5, *а* видно, что первая и 14-я подлинные подписи человека U0184 существенно отличаются от остальных по признаку абсолютной скорости исполнения подписи. Представлены два варианта проекций образов подписи, описанные признаками *dxy* и *ddy* в виде 276 синих точек, которые очерчены выпуклым многоугольником. Красными кружками показаны *dtw*-расстояния от 14-й подписи до остальных 24, по которым построен образ подлинных подписей человека. Если при построении образа подписей данного человека использовать первую и 14-ю подписи, образ существенно увеличится в признаковом пространстве, что может привести к неверной верификации поддельных подписей. *Dtw*-расстояния 14-й подписи по признаку *ddy* превышают расстояния между парами остальных подлинных подписей в 1,5–2 раза.

Графики на рис. 7 также указывают на выбросы *dtw*-расстояний по признакам *ddxy*, *ddy*, *dy* от 14-й подлинной подписи человека U0184 до его остальных подписей. При построении классификатора подписей человека необходимо использовать подобный численный анализ данных о признаках при построении образа его подписей в признаковом пространстве. На примере подписей этого человека показано, что образ подлинных подписей каждого человека следует строить индивидуально, выбирая наиболее общие признаки сходства подписей. Подлинные подписи могут существенно отличаться по разным причинам: подпись не выработана, подпись выполнена в неудобных условиях, человек болен, подпись сознательно искажается с целью признания ее фальшивой впоследствии и др.

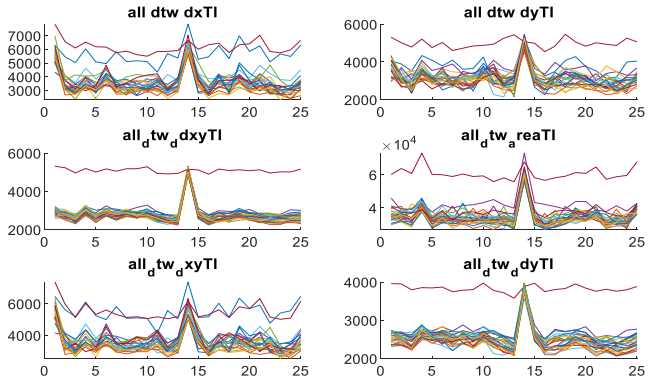


Рис. 7. Выбросы *dtw*-расстояний, вычисленных по признакам *ddxy*, *ddy*, *dy* 14-й подписи человека U0184

Fig. 7. Dtw-distances outliers calculated by the features ddxy, ddy, dy of the fourteenth signature of person U0184

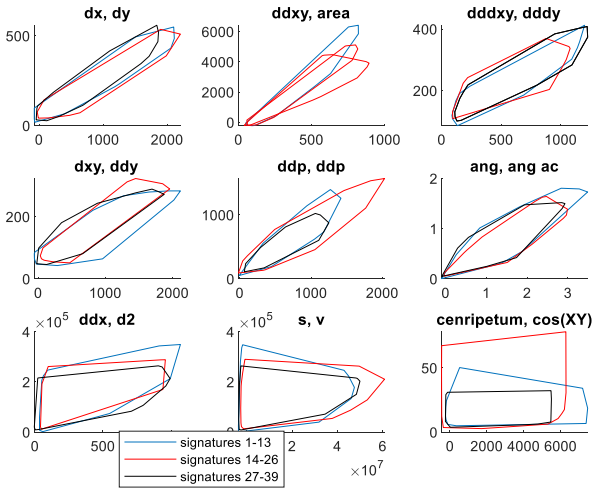
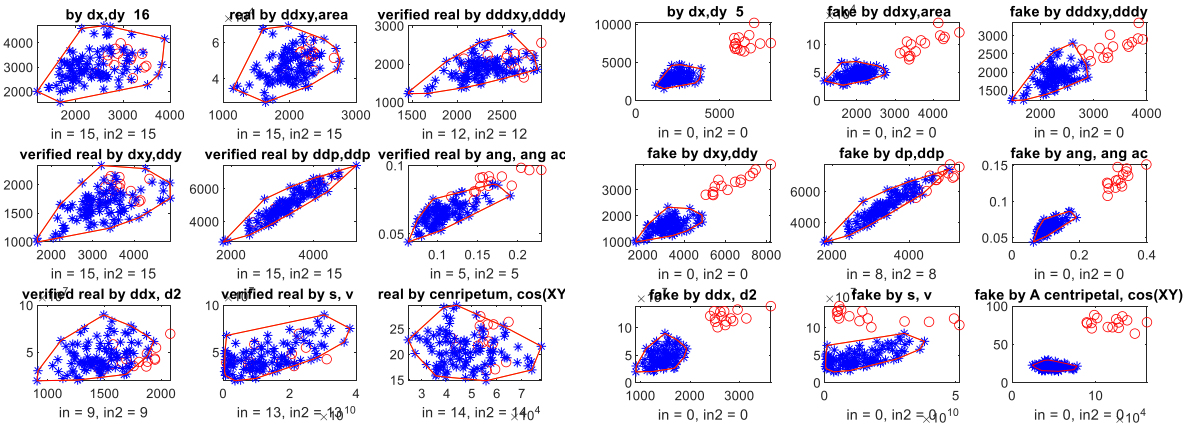


Рис. 8. Проекции трех образов подписей человека U1009, построенных по трем непересекающимся группам подписей для N=13

Fig. 8. Projections of three signature patterns of person U1009, constructed from three disjoint groups of signatures for N=13

На рис. 8 показано, что образ подлинных подписей человека существенно зависит от изменчивости конкретных подписей, использованных для его построения. Все оси на рис. 8 имеют разный масштаб.

Верифицируемая подлинная подпись на рис. 9, *a* не попадает в проекцию, образованную признаками «угол» и «изменение угла» (*ang*, *ang_ac*). Поддельная подпись человека, наоборот, попадает в проекцию, образованную признаками «скорость» и «ускорение изменения давления» (*dp*, *ddp*). Учитывая попадание в образ большей части признаков, вычисленных по парам подписей, можно сделать верное заключение о подлинности анализируемой подписи. Отметим, что признаки на базе углов не используются во второй модели образа подписей.



a) *b)*

Рис. 9. Проекции образа подлинных подписей человека (синие точки), верификация подписи (красные кружки) подлинной (*a*) и поддельной (*b*)

Fig. 9. Projections of a pattern of a person's genuine signatures (blue dots), verification of a signature (red circles) genuine (a) and fake (b)

При использовании девяти проекций образа подлинных подписей человека общая точность не превышает 90,9 % при $N = 10$, а при увеличении N она немного снижается (рис. 10). При семи проекциях точность распознавания поддельных подписей составляет примерно 99 % независимо от N, а точность распознавания подлинных подписей растет с увеличением N.

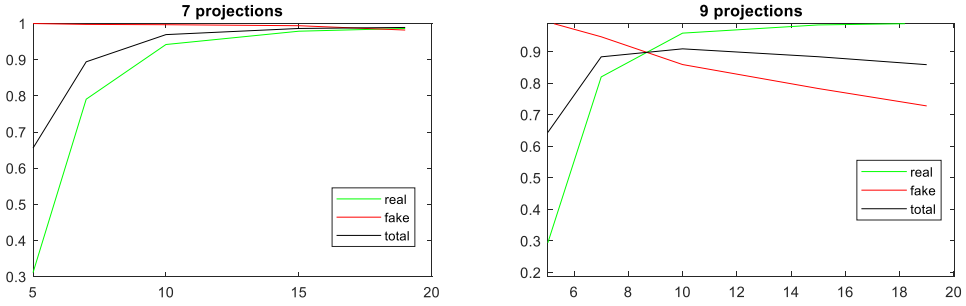


Рис. 10. Точность верификации подписей 100 пользователей при использовании семи и девяти проекций образов подлинных подписей

Fig. 10. Accuracy of signatures verification of 100 clients using seven and nine projections of patterns of genuine signatures

В табл. 3 собраны результаты верификации подписей 230 человек, по 25 подлинных и 25 поддельных каждого, а их общее число равно 11 500. Символ R означает подлинные подписи, символ F – поддельные, индекс 1 означает первую модель признакового пространства (18 признаков), индекс 2 – вторую (14 признаков). Образы подписей каждого человека строились по его первым N подлинным подписям, представленным в базе. Затем верифицировались все подписи каждого человека (50 штук).

Таблица 3
Точность верификация 11 500 подписей 230 человек из базы DeepSignDB
Table 3
Verification accuracy of 11,500 signatures of 230 people from the DeepSignDB database

Число подлинных подписей для построения образа, *N* *Number of genuine signatures to build the pattern, N*	5	7	10	13	15	19
Точность *R*1 *Accuracy R1*	0,2977	0,8181	0,9517	0,9746	0,9802	0,9873
Точность *F*1 *Accuracy F1*	0,9918	0,9344	0,8398	0,7965	0,7567	0,7068
Общая точность 1 *Overall accuracy 1*	0,6448	0,8763	**0,8957**	0,8856	0,8684	0,8470
Точность *R*2 *Accuracy R2*	0,3150	0,7911	0,9344	0,9656	0,9737	0,9830
Точность *F*2 *Accuracy F2*	0,9998	0,9976	0,9960	0,9939	0,9923	0,9910
Общая точность 2 *Overall accuracy 2*	0,6574	0,8943	0,9652	0,9797	**0,9830**	**0,9870**

Анализируя табл. 3, можно сделать два вывода:

– первая модель дает несколько больше ошибок верификации;

– числа подлинных подписей $N<10$ для построения образа подписи человека недостаточно для надежной верификации. Оптимальным является $N=15$. Увеличение N не дает существенного прироста точности, но увеличивает время вычислений. Отметим, что точность верификации поддельных подписей во второй модели более 0,99 при любом N, но точность определения подлинных подписей существенно зависит от N из-за вариативности исполнения подписи человеком. Лучшие значения точности верификации выделены жирным шрифтом.

Заключение. В работе были выполнены эксперименты по верификации подлинных и поддельных подписей 498 человек (по 25 подлинных и поддельных подписей) из базы DeepSignDB. Для каждого человека строился индивидуальный образ его подписей. При использовании $N = 15$ и без оптимизации образа конкретного человека была получена точность распознавания порядка 98 % при анализе 24 900 подписей этих людей.

На примере анализа подписей сотен людей показано, что простейший вариант верификации динамической подписи может быть реализован на базе порогового классификатора. Если одномерный или многомерный признак попадает в определенный диапазон значений, подпись относится к классу подлинных, иначе – фальшивых. Обобщением такого классификатора является выпуклый K-мерный многогранник, где K – размерность признакового пространства. Такой многогранник дает более точные результаты, чем совокупность его проекций на несколько плоскостей, однако выпуклые многоугольники на плоскости строить проще и быстрее.

На примерах десятков тысяч подписей сотен человек продемонстрирована высокая точность нового метода верификации динамических подписей с использованием классификатора, который строится для каждого человека индивидуально по ограниченному множеству его подлинных подписей.

Список использованных источников

1.Ярощук, И. А. Проблемные вопросы экспертизы подписи как малообъемного почеркового объекта / И. А. Ярощук, К. В. Гриневич // Актуальные проблемы российского права. – 2021. – Т. 129, № 8. – С. 141–151.

2. Мещеряков, В. А. Оценка возможностей почерковедческой экспертизы сквозь призму современных информационных технологий / В. А. Мещеряков, В. В. Бутов // Вестн. Воронежского ин-та МВД России. – 2017. – № 2. – С. 40–46.

3. Kaur, H. Signature identification and verification techniques: state-of-the-art work / H. Kaur, M. Kumar // J. of Ambient Intelligence and Humanized Computing. – 2023. – Vol. 14, no. 2. – P. 1027–1045.

4. Ratanamahatana, C. A. Everything you know about dynamic time warping is wrong / C. A. Ratanamahatana, E. Keogh // Third Workshop on Mining Temporal and Sequential Data, Seattle, USA, 22 Aug. 2004. – Seattle, 2004. – P. 50–60.

5. Fenton, D. Evaluation of features and normalization techniques for signature verification using dynamic timewarping / D. Fenton, M. Bouchard, T. H. Yeap // 2006 IEEE Intern. Conf. on Acoustics Speed and Signal Processing Proc., Toulouse, France, 14–19 May 2006. – Toulouse, 2006. – Vol. 3. https://doi.org/10.1109/icassp.2006.1660860

6. Discriminative feature selection for on-line signature verification / X. Xia [et al.] // Pattern Recognition. – 2018. – Vol. 74. – P. 422–433.

7. Mobile signature verification: Feature robustness and performance comparison / M. Martinez-Diaz [et al.] // IET Biometrics. – 2014. – Vol. 3, no. 4. – P. 267–277.

8. Starovoitov, V. Writer-dependent approach to off-line signature verification / V. Starovoitov, U. Akhundjanov // Pattern Recognition and Information Processing (PRIP'2023) : Proc. of the 16th Intern. Conf., Minsk, Belarus, 17–19 Oct. 2023. – Minsk, 2023. – P. 241–244.

9. Bellman, R. On the theory of dynamic programming / R. Bellman // Proc. of the National Academy of Sciences. – 1952. – Vol. 38(8). – P. 716–719.

10. BioSecure signature evaluation campaign (BSEC'2009): Evaluating online signature algorithms depending on the quality of signatures / N. Houmani [et al.] // Pattern Recognition. – 2012. – Vol. 45, no. 3. – P. 993–1003.

11. Старовойтов, В. В. Следует ли нормализовать данные динамических подписей перед верификацией методом DTW? / В. В. Старовойтов // BIG DATA и анализ высокого уровня : сб. науч. ст. X Междунар. науч.-практ. конф., Минск, 13 марта 2024 г. : в 2 ч. – Минск, 2024. – Ч. 2. – С. 391–400.

References

1. Jaroshhuk I. A., Grinevich K. V. *Problematic issues of examination of a signature as a small-volume handwriting object*. Aktual'nye problemy rossijskogo prava [*Current Problems of Russian Law*], 2021, vol. 129, no. 8, pp. 141–151 (In Russ.).

2. Meshherjakov V. A., Butov V. V. *Assessing the capabilities of handwriting examination through the prism of modern information technologies.* Vestnik Voronezhskogo instituta Ministerstva vnutrennih del Rossii [*Bulletin of the Voronezh Institute of the Ministry of Internal Affairs of Russia*], 2017, no. 2, pp. 40–46 (In Russ.).

3. Kaur H., Kumar M. Signature identification and verification techniques: state-of-the-art work. *Journal of Ambient Intelligence and Humanized Computing*, 2023, vol. 14, no. 2, pp. 1027–1045.

4. Ratanamahatana C. A., Keogh E. Everything you know about dynamic time warping is wrong. *Third Workshop on Mining Temporal and Sequential Data, Seattle, USA, 22 August 2004*. Seattle, 2004, pp. 50–60.

5. Fenton D., Bouchard M., Yeap T. H. Evaluation of features and normalization techniques for signature verification using dynamic timewarping. *2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings, Toulouse, France, 14–19 May 2006*. Toulouse, 2006, vol. 3. https://doi.org/10.1109/icassp.2006.1660860

6. Xia X., Song X., Luan F., Zheng J., Chen Z., Ma X. Discriminative feature selection for on-line signature verification. *Pattern Recognition*, 2018, vol. 74, pp. 422–433.

7. Martinez-Diaz M., Fierrez J., Krish R. P., Galbally J. Mobile signature verification: Feature robustness and performance comparison. *IET Biometrics*, 2014, vol. 3, no. 4, pp. 267–277.

8. Starovoitov V., Akhundjanov U. Writer-dependent approach to off-line signature verification. *Pattern Recognition and Information Processing (PRIP'2023): Proceedings of the 16th International Conference, 17–19 October 2023*. Minsk, Belarus, 2023, pp. 241–244.

9. Bellman R. On the theory of dynamic programming. *Proceedings of the National Academy of Sciences*, 1952, vol. 38(8), pp. 716–719.

10. Houmani N., Mayoue A., Garcia-Salicetti S., Dorizzi B., Khalil M. I., …, Vivaracho-Pascual C. BioSecure signature evaluation campaign (BSEC'2009): Evaluating online signature algorithms depending on the quality of signatures. *Pattern Recognition*, 2012, vol. 45, no. 3, pp. 993–1003.

11. Starovoitov V. V. *Should we normalize dynamic signatures data before DTW-based verification?* BIG DATA i analiz vysokogo urovnja : sbornik nauchnyh statej X Mezhdunarodnoj nauchno-prakticheskoj konferencii, Minsk, 13 marta 2024 goda : v 2 chastjah [*BIG DATA and Advanced Analytics : Collection of Scientific Articles of the X International Scientific and Practical Conference, Minsk, 13 March 2024 : in 2 Parts*]. Минск, 2024, part 2, pp. 391–400 (In Russ.).

Информация об авторе

Старовойтов Валерий Васильевич, доктор технических наук, профессор, Объединенный институт проблем информатики Национальной академии наук Беларуси.
E-mail: valerys@newman.bas-net.by

Information about the author

Valery V. Starovoitov, D. Sc. (Eng.), Prof., The United Institute of Informatics Problems of the National Academy of Sciences of Belarus.
E-mail: valerys@newman.bas-net.by

Правила для авторов

Редакция журнала «Информатика» просит авторов руководствоваться приведенными ниже правилами.

I. Статьи принимаются в редакцию через электронную систему подачи по адресу http://inf.grid.by в формате файлов текстовых редакторов Microsoft Word. Объем оригинальной статьи – от 8 до 16 стр., включая рисунки, таблицы и достаточное количество наиболее актуальных ссылок; объем обзорной статьи – от 16 до 32 стр., включая все основные ссылки. Текст набирается с переносами, шрифт Times New Roman 11 пт, интервал между строками одинарный, абзацный отступ 0,5 см, поля по 2,5 см со всех сторон.

Материал статьи должен быть четко структурированным: Введение; основные разделы, в которых изложены цели и задачи, методы, результаты; Заключение (выводы).

II. Статьи о результатах работ, проведенных в научных учреждениях, должны иметь разрешение на публикацию (сопроводительное письмо за подписью руководителя или выписку из заседания ученого совета, отдела или кафедры, акт экспертизы).

III. Статьи в обязательном порядке должны включать аннотацию, ключевые слова, список литературы, информацию об авторах на русском и английском языках.

На титульной странице располагаются следующие метаданные:

1. Индекс по универсальной десятичной классификации (УДК); *на русском и английском языках* тип статьи (оригинальная или обзорная), название статьи, инициалы и фамилии всех авторов, полное наименование учреждений, где работают авторы, с указанием почтового адреса, при наличии указывается ученая степень и ORCID, e-mail ответственного лица.

2. Аннотация (Abstract) объемом 150–250 слов в оригинальной статье должна быть структурирована отдельными подразделами: Цели, Методы, Результаты, Заключение, а также максимально характеризовать содержательную часть рукописи. Сюда не следует включать впервые введенные термины, аббревиатуры (за исключением общеизвестных), ссылки на литературу.

3. Ключевые слова (Keywords) – наиболее значимые слова или словосочетания по теме работы, отражающие специфику темы, объекты и результаты исследования; перечень ключевых слов должен содержать 5–10 слов.

4. В разделе Благодарности (Acknowledgements) указываются все источники финансирования исследования, а также благодарности людям, которые участвовали в работе над статьей.

5. Автор обязан уведомить редакцию о реальном или потенциальном конфликте интересов, включив информацию в раздел Конфликт интересов (Conflict of interest).

6. Формулы, рисунки, таблицы в статье нумеруются в соответствии с порядком их упоминания в тексте. Ссылки на рисунки и таблицы в тексте обязательны. Рисунки должны быть выполнены с хорошим разрешением в масштабе, позволяющем четко различать надписи и обозначения. Цветные иллюстрации печатаются только в том случае, когда это необходимо для понимания излагаемого материала. Подрисуночные подписи с расшифровкой всех позиций, представленных на рисунке, и названия таблиц набираются шрифтом гарнитуры основного текста размером 9 пт. Перевод подрисуночной подписи и пояснений к рисунку, а также перевод названия таблицы, заголовки строк или столбцов располагаются курсивом после русскоязычной версии.

7. Набор формул выполняется в формульном редакторе Microsoft Equation или Math Type. Прямым шрифтом набираются: греческие и русские буквы; математические символы (sin, lg, ∞); символы химических элементов (C, Cl, CHCl3); цифры (римские и арабские); индексы (верхние и нижние), являющиеся сокращениями слов. Курсивом набираются латинские буквы, символы физических величин (в том числе и в индексе).

8. Список использованной литературы оформляется в соответствии с требованиями Высшей аттестационной комиссии Республики Беларусь (ГОСТ 7.5–2008). Номер литературной ссылки в тексте дается порядковым номером в квадратных скобках. Ссылаться на неопубликованные работы не допускается.

9. Отдельно оформляется References со следующей структурой: авторы (транслитерация), транслитерированное название монографии, *Перевод названия монографии на английский язык*. Выходные данные с обозначениями на английском языке. От транслитераций названий статей можно отказаться.

10. Ссылки на учебно-методическую литературу, ГОСТы, авторефераты, статистические отчеты в список не включаются, а оформляются в виде сносок (с подробными рекомендациями можно ознакомиться на сайте журнала в разделе Правила для авторов).

11. В разделе Информация об авторах (Information about the authors) приводятся ФИО авторов полностью, ученая степень, звание, должность, название организации, ORCID (при наличии).

IV. Все поступающие в редакцию рукописи проходят предварительную проверку на соответствие Правилам для авторов. Статья может быть возвращена автору на доработку с просьбой устранить недостатки или дополнить информацию. После проверки на соответствие правилам статья направляется рецензенту с указанием сроков рецензирования.

V. При наличии замечаний рецензента автору предоставляется определенное время на доработку рукописи. Статьи, направляемые на доработку, должны быть возвращены в исправленном виде с ответами на все замечания. Окончательное решение о публикации или отклонении рукописи принимается редколлегией журнала. При положительном заключении рецензента статья передается редактору для подготовки к печати. Редакция оставляет за собой право на редакционные изменения, не искажающие основное содержание статьи.

VI. Редакция журнала предоставляет возможность первоочередного опубликования статей, представленных лицами, которые осуществляют послевузовское обучение (аспирантура, докторантура, соискательство) в год завершения обучения.

VII. Авторы несут ответственность за направление в редакцию статей, уже опубликованных ранее или принятых к публикации другими изданиями.

ИНДЕКСЫ

00827
для индивидуальных
подписчиков

008272
для предприятий
и организаций

ИНФОРМАТИКА ■ INFORMATICS ТОМ ■ VOL. 21 2 | 2024

ISSN 1816-0301 (Print)
9 771816 030000