Теоретические вопросы информатики, прикладной математики, компьютерных наук и когнитивно-информационных технологий

https://doi.org/10.25559/SITITO.019.202304.883-892
УДК 681.3.016

Современные тенденции в построении моделей данных многомерных информационных систем с использованием методов классификации

Ж. М. Муратова

Оригинальная статья

НАО «Западно-Казахстанский университет имени Махамбета Утемисова», г. Уральск, Республика Казахстан

Адрес: 090000, Республика Казахстан, г. Уральск, пр. Н. Назарбаева, д. 162

jansik.86@mail.ru

Аннотация

Данное исследование посвящено современным тенденциям в построении моделей многомерных информационных систем с использованием методов классификации. Было определено, что в построении таких моделей предварительная обработка данных является важным этапом с помощью, которого можно решить задачу редукции данных, также достичь оптимальной классификации данных. Для проведения исследования была выбрана база данных об успеваемости студентов технического университета программы бакалавриата. Полученная база данных использовалась для сравнения моделей классификации CART, LDA, SVM, KNN и RF с применением программного обеспечения Rstudio. В данной работе отмечено, что одной из современных тенденций в построении моделей данных многомерных информационных систем является использование ансамблей классификаторов, которые объединяют несколько моделей или алгоритмов классификации для достижения более точных и надежных результатов. Путем сравнения полученных результатов классификации можно определить, насколько прогнозы, полученные с помощью двух разных алгоритмов, коррелируют друг с другом. В целом, современные методы классификации стремятся улучшить точность и эффективность классификации, адаптироваться к различным типам данных и структурам, а также расширить возможности использования данных моделей в реальных приложениях. В исследовании предполагается, что с учетом развития технологий и роста объемов данных дальнейшее развитие методов классификации будет продолжаться.

Ключевые слова: многомерные информационные системы, методы классификации, многомерная модель, сравнительный анализ моделей классификации, машинное обучение

Конфликт интересов: автор заявляет об отсутствии конфликта интересов.

Для цитирования: Муратова Ж. М. Современные тенденции в построении моделей данных многомерных информационных систем с использованием методов классификации // Современные информационные технологии и ИТ-образование. 2023. Т. 19, № 4. С. 883-892. https://doi.org/10.25559/SITITO.019.202304.883-892

Theoretical Questions of Computer Science, Computational Mathematics, Computer Science and Cognitive Information Technologies

Current Trends in the Construction of Data Models of Multidimensional Information Systems Using Classification Methods

Zh. M. Muratova

Mahambet Utemisov West Kazakhstan University, Uralsk, Republic of Kazakhstan

Address: 162 N. Nazarbayev Ave., Uralsk 090000, Republic of Kazakhstan

jansik.86@mail.ru

Original article

Abstract

This study is devoted to current trends in the construction of models of multidimensional information systems using classification methods. It was determined that in the construction of such models, data preprocessing is an important step with the help of which it is possible to solve the problem of data reduction, as well as to achieve optimal data classification. A database on the academic performance of students of the Technical University of the Bachelor's degree program was selected for the study. The resulting database was used to compare CART, LDA, SVM, KNN and RF classification models using Rstudio software. In this paper, it is noted that one of the current trends in the construction of data models of multidimensional information systems is the use of ensembles of classifiers that combine several models or classification algorithms to achieve more accurate and reliable results. By comparing the classification results obtained, it is possible to determine to what extent the predictions obtained using two different algorithms correlate with each other. In general, modern classification methods strive to improve the accuracy and efficiency of classification, adapt to different types of data and structures, and expand the possibilities of using these models in real applications. The study assumes that, taking into account the development of technologies and the growth of data volumes, further development of classification methods will continue.

Keywords: multidimensional information systems, classification methods, multidimensional model, comparative analysis of classification models, machine learning

Conflict of interests: The author declares no conflict of interests.

For citation: Muratova Zh.M. Current Trends in the Construction of Data Models of Multidimensional Information Systems Using Classification Methods. *Modern Information Technologies and IT-Education.* 2023;19(4):883-892. https://doi.org/10.25559/SITITO.019.202304.883-892

1 Введение

В настоящее время с развитием современных информационных технологий и необходимости получения информативных данных, объём которых стремительно растёт, построение эффективных моделей данных многомерных информационных систем становится все более актуальным. Многомерные информационная системы (МИС) используется во многих сферах, так как надежное хранилище данных способствует эффективному управлению и обеспечивает поддержку аналитики для любой организации. В современную эпоху общества, основанного на данных, организации сталкиваются с острой необходимостью эффективного управления большими объемами информации и ее анализа, в то время как многомерные информационные системы (MIS) появились как мощный подход, помогающий в процессах принятия решений, предоставляя всесторонний взгляд на сложные наборы данных. Тем не менее, разработка эффективных моделей данных для многомерных информационных систем сопряжена с рядом проблем из-за присущей им сложности и разнообразия данных, в решениях которых методы классификации являются одним из важных инструментов. Следовательно, считается актуальным выявление современных тенденций, регулирующих создание моделей данных в многомерных информационных системах, с особым акцентом на использование методов классификации. Не менее важным считается изучение того, как эти методы могут повысить точность и эффективность моделирования данных, облегчая организацию и анализ огромных объемов информации.

Как правило, моделирование данных в MIS опиралось на такие методы, как пространственное моделирование и моделирование взаимосвязей сущностей, но рост сложности данных параллельно с технологическим прогрессом привёл к тому, что методы классификации приобретают все большее значение благодаря их эффективности при обработке объемных и разнородных наборов данных [1, 2].

Известно, что в методах классификации используются алгоритмы машинного обучения для классификации данных по предопределенным классам. В таких алгоритмах за основу взяты определенные атрибуты или характеристики, которые дают возможность построить такие модели данных, которые не только классифицируют, но и организуют информацию осмысленным образом с целью получения ценной информации для принятия решений. Преимущество методов классификации в построении моделей данных для многомерных информационных систем заключается в способности обрабатывать многомерные данные. Эта способность не только позволяет проводить более сложный анализ, но и понимать взаимосвязи между различными измерениями данных. Методы классификации находят применение в различных областях MIS, включая сегментацию клиентов, выявление мошенничества, оценку рисков и анализ настроений. Внедряя эти методы, организации могут извлекать ценную информацию и принимать решения, основанные на данных.

В этой статье будут рассмотрены современные тенденции, регулирующие построение моделей данных в многомерных информационных системах, с особым акцентом на использование методов классификации. В рамках этой работы будут рассмотрены различные методы классификации, используемые при построении моделей данных для MIS, включая методы k ближайших соседей, опорных векторов, случайного леса. Оценка производительности методов классификации, выявление ограничений и их применений в построении модели данных и анализе многомерных информационных систем считается важным этапом. Следует отметить, по мере того как технологии продолжают развиваться и генерировать огромные объемы данных, появятся новые методы классификации, предлагающие более сложные инструменты построения моделей данных и извлечения ценной информации из многомерных информационных систем. Таким образом, построение моделей данных для многомерных информационных систем представляет собой сложную и важную задачу. Тем не менее, внедрение современных методов классификации предлагает многообещающие решения для эффективного распределения и анализа огромных объемов данных. Используя эффективность этих методов, можно открыть новые возможности для принятия обоснованных решений, что приведет к повышению производительности и конкурентным преимуществам цифровых технологий.

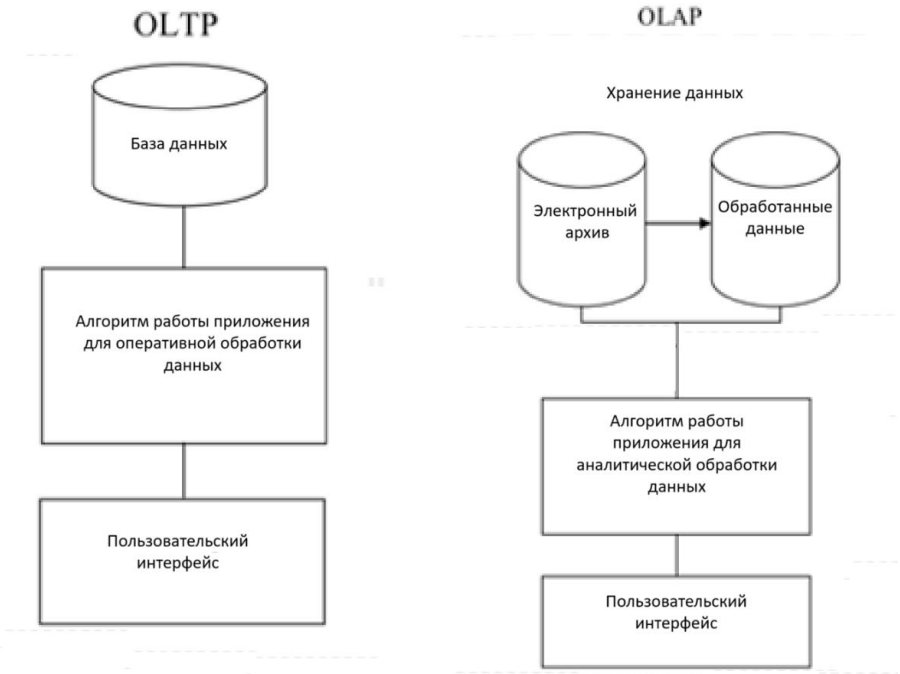
В современном мире, основанном на данных, способность эффективно управлять большими объемами информации и анализировать их стала критически важной для организаций. Более того, методы классификации могут быть применены к широкому спектру приложений в рамках MIS, включая сегментацию клиентов, выявление мошенничества, оценку рисков и анализ настроений. Используя эти методы, организации могут получать ценную информацию и принимать обоснованные решения для выполнения планов.

2 Определение многомерных информационных систем и их роль в современном мире

Многомерная модель данных – это многомерное логическое представление информационных структур, которая основано на манипулировании данными. Объемные данные составляют многомерную базу данных (MDB), где положение каждой ячейки определяется набором переменных (измерения). Каждая ячейка обозначает определенное событие, а значения измерений указывают момент и местоположение, где происходит событие. Удобство и эффективность аналитической обработки больших объемов данных являются основными достоинствами многомерной модели данных, в то время как,

недостатком считается ее громоздкость для простейших задач обычной оперативной обработки информации.

В статистической обработке данных информационные системы играют ключевую роль, так как могут использоваться для обработки данных и для решения более сложных задач. Например, статистический анализ количественных данных, собранных за определенное время, с целью принимать решения, основанные на этих данных. В первом случае данные могут быть обработаны с использованием технологии онлайн-обработки транзакций *(OLTP technology)*. Транзакция – это отдельная операция, связанная с базой данных, которая рассматривается как единое целое действие. Во втором случае используется система *Operational Analytical Processing (OLAP)*, главное преимущество которой заключается в качестве анализа и скорости обработки больших объемов данных. Такой высокий уровень OLAP можно объяснить мощными многопроцессорными технологиями, сложными методами анализа и специализированными хранилищами данных. На рисунке 1 показаны структуры технологий OLTP и OLAP.



Р и с. 1. Описание структур технологий для аналитической обработки информационных систем
F i g. 1. Description of technology structures for analytical processing of information systems

Источник: здесь и далее в статье все таблицы и рисунки составлены автором.
Source: Hereinafter in this article all tables and figures were made by the author.

Структурную схему системы OLAP можно представить как куб, который состоит из многомерных данных или множества таблиц, где данные не хранятся в виде отдельных значений. Например, если рассмотреть куб оценок учащихся за семестр, то каждая ячейка будет содержать один вычисленный результат оценки, а не количество всех оценок, полученных в разных ячейках.

3 Основные методы и их применение для построения моделей данных с использованием методов классификации

3.1 Предобработка данных. Выделение информативных признаков многомерной модели

При решениях задач классификации многомерной модели данных не менее важную часть составляют методы предварительной обработки данных, так как они могут значительно повлиять на предсказательную точность модели. Известно, что задача предварительной обработки данных заключается в очистке, преобразовании и сокращении данных. Во многих исследованиях наблюдалось, что преобразование данных для уменьшения воздействия асимметрии данных или выбросов, приводит к улучшению производительности модели классификации.

Предобработка данных также включает метод отбора переменных, цель которого заключается в выборе наиболее важных признаков для использования в модели. Значимость этого метода понимается в двух аспектах: релевантность и избыточность. По первому аспекту значимости метода отбора переменных выбранные признаки должны достаточно сильно влиять на зависимую переменную и отражать связи и закономерности данных. А по второму аспекту признаки не должны быть коррелированы и нести одну и ту же информацию. Это особенно актуально для моделей линейной регрессии, где незначимые и избыточные переменные не только увеличивают размерность задачи, но и снижают результативность. Избыточные и незначимые признаки могут быть

удалены без важной потери информации и ухудшения качества модели. Для отбора значимых переменных в моделях классификации считается актуальным использование алгоритма обратного исключения, когда отбор начинается с модели, содержащей все переменные. По алгоритму обратного исключения на каждом шаге определяется наименее значимая переменная и выявляется правило остановки [3].

Сложность подзадачи выбора признаков распознавания обусловливается ее переборным характером. Достаточно большая размерность данных, исследуемых в задачах распознавания, ограничивают возможности применения точных математических методов, реализующих общий подход к решению данной подзадачи – полный перебор, что привело к появлению многочисленных математических методов выделения признаков, направленных на сокращение полного перебора [4]. Следует отметить, что методами выделения информативных признаков можно эффективно решить задачу уменьшения размерности многомерной модели данных, повышая предсказательную точность модели классификации, также уменьшив время обучения модели.

3.2 Определение модели классификации для многомерной модели информационных систем

Система определения модели классификации для многомерной модели информационных систем состоит из алгоритмов и методов, которые позволяют улучшить качество аналитики, обеспечить безопасность данных и улучшить понимание информации, что делает ее важной для эффективного функционирования информационной системы. Классификация многомерных данных помогает упорядочить информацию и представить ее в такой форме, которая позволит лучше понимать данные и принимать обоснованные решения на основе этой информации.

Данные многомерной модели, которые были получены после предварительной обработки, используются для дальнейшего решения задачи классификации. Для задачи классификации необходимо выбрать модель, которая будет использоваться для классификации данных в информационной системе. Определение модели классификации можно поделить на несколько важных шагов: 1) Анализ требований: определение требований к информационной системе и их анализ, включающий в себя понимание целей и задач системы, а также описание требуемых функциональностей и возможных видов классификации данных. 2) Исследование моделей: изучение и анализ различных моделей классификации, которые могут быть применены к многомерной модели информационной системы. Так как важно изучить особенности каждой модели, ее преимущества и недостатки, а также ее применимость к данной информационной системе. 3) Выбор модели: на основе анализа требований и результатов исследования моделей классификации, выбирается наиболее подходящая модель для данной информационной системы. Выбор модели может зависеть от множества факторов, таких как тип данных, объем данных, задачи классификации и доступные ресурсы. 4) Проектирование модели: после определения модели классификации, происходит ее проектирование. Это включает в себя определение атрибутов и характеристик классификации, создание структуры модели и определение алгоритмов и методов классификации данных. 5) Разработка и тестирование: в итоге, происходит разработка модели классификации и ее тестирование на реальных данных информационной системы. Важно проверить работоспособность модели, ее точность и эффективность в классификации данных. Стадия определения модели классификации для многомерной модели информационных систем важным этапом и требует тщательного анализа и выбора подходящей модели для достижения требуемых результатов классификации.

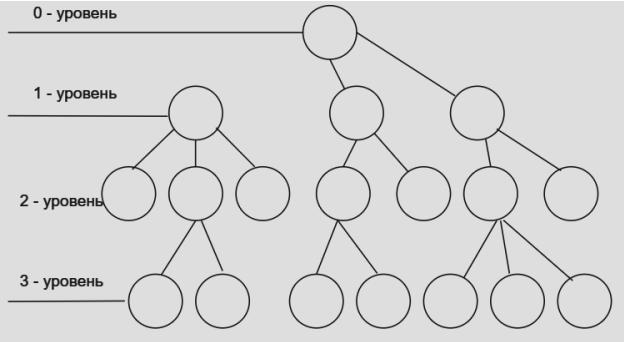
3.3 Определение иерархической классификации информационных систем для моделей многомерных данных

Для информационных систем с многомерными моделями применяются классификационные подходы, которые формируют правила для разбиения ячеек куба на группы. Среди них можно выделить основные классификационные подходы, такие как иерархические и фасетные. По иерархическому методу классификации любой объект на каждом уровне относится к одному классу, характеризуемым конкретным значением своего классификационного признака [7]. В операциях группировки в любом новом классе задаются свои конкретные классификационные признаки. На уровне иерархии с группировкой семантическое содержание класса влияет на набор классификационных признаков. Как правило, соответствующие количеству классификационных признаков уровни классификации, определяет её глубину.

Проведено немало исследований в области интеллектуального анализа данных, статистического распознавания образов и машинного обучения, которые сосредоточились на задачах плоской классификации. Важные проблемы в решении задач классификации в реальной жизни рассматриваются как задачи иерархической классификации, когда классы прогнозирования группируются в иерархию классов. Например, дерево или DAG, который считается ациклическим графом. Задачи иерархической классификации принято определять четко, так как могут возникать проблемы упущения из виду или спутывания с другими задачами, которые имеют одинаковые названия.

Иерархическая система классификация представлена в иллюстрации [2].

0-й уровень составляет набор элементов и делится по выбранному классификационному признаку на группировки, который образуют 1-й уровень. Далее каждый класс 1-го уровня по классификационному признаку делится на подклассы, которые образуют 2-й уровень. Следовательно, каждый класс 2-го уровня делится на группы, которые образуют 3-й уровень и т.д.



Р и с. 2. Иерархическая система классификации
F i g. 2. Hierarchical classification system

В построении моделей данных в информационных системах иерархические системы классификации отличаются с достоинствами в простоте построения и использовании независимых друг от друга классификационных признаков в ветвях иерархии. Но также существуют недостатки, которые описываются жесткой структурой. Данный недостаток может привести к сложности внесения изменений в классификационных группировках. Следовательно, в иерархической системе классификации особое внимание следует уделить выбору обоснованных классификационных признаков.

3.4 Современные тенденции в выборе модели классификации для многомерных информационных систем

Классификация – это форма анализа данных, которая извлекает модель из данных для классификации будущих данных. Он параллельно изучался в области статистики и машинного обучения и в настоящее время является основным методом интеллектуального анализа данных с широким спектром применения. Поскольку многие прикладные задачи могут быть сформулированы как задача классификации, а объем доступных данных стал огромным, разработка масштабируемых, эффективных, зависящих от предметной области алгоритмов классификации, сохраняющих конфиденциальность, имеет важное значение.

Важное значение приобретают масштабируемые алгоритмы классификации. Хотя были предложены некоторые масштабируемые алгоритмы для изучения дерева решений, все еще существует необходимость в разработке масштабируемых и эффективных алгоритмов для других типов методов классификации, таких как изучение правил принятия решений. Ранее изучение методов классификации было сосредоточено на изучении различных механизмов обучения для повышения точности классификации на невидимых примерах. Однако недавнее исследование несбалансированных наборов данных показало, что точность классификации не является подходящим показателем для оценки эффективности классификации, когда набор данных крайне несбалансирован, в котором почти все примеры принадлежат к одному или нескольким более крупным классам и гораздо меньшее количество примеров относится к меньшему, обычно более интересному классу. Поскольку многие наборы данных реального мира несбалансированы, наметилась тенденция к корректировке существующих алгоритмов классификации, чтобы лучше идентифицировать примеры из класса редких. Еще одним вопросом, который становится все более важным при интеллектуальном анализе данных, является защита конфиденциальности. По мере того, как инструменты интеллектуального анализа данных применяются к большим базам данных личных записей, проблемы конфиденциальности возрастают. Интеллектуальный анализ данных с сохранением конфиденциальности в настоящее время является одной из самых популярных тем исследований в области интеллектуального анализа данных и останется таковой в ближайшем будущем.

4 Сравнительный анализ моделей классификации на многомерные данные

С целью сравнительного анализа была рассмотрена задача классификации на многомерной модели данных, со следующими методами классификации: случайный лес, kNN, LDA, также методы опорных векторов и деревьев решений. Для этого была использована среда программного обеспечения Rstudio, а в качестве многомерной базы данных были взяты данные из журнала успеваемости студентов, с разными атрибутами, определяющие итоговую оценку прохождения аттестации студентами.

Были применены разные модели классификации, с целью определить сдал ли студент предмет или нет. В рисунке 2 представлен фрагмент многомерной модели базы данных, который был применен для сравнительного анализа.

Одной из современных тенденций в построении моделей данных многомерных информационных систем является использование ансамблей классификаторов. Ансамбли классификаторов объединяют несколько моделей или алгоритмов классификации для достижения более точных и надежных результатов. С помощью сравнения результатов классификации моделей можно ответить на вопрос о том, коррелируют ли прогнозы, полученные с помощью двух разных алгоритмов. Если они слабо коррелированы, то являются хорошими кандидатами для объединения в ансамблевом прогнозировании.

В рисунке 4 получены результаты точности моделей классификации Cart, LDA, SVM, kNN, RF, где можно заметить, что высокие точности классификации наблюдаются в моделях LDA и RF, с точностями 0.965 и 0.986, соответственно.

По рисунку 5 можно увидеть распределение оценочных значений точности для различных моделей классификации и на то, как они соотносятся. Так как ячейки упорядочены от наибольшей к наименьшей средней точности.

Рисунок 6 показывает распределение точности модели в виде графиков плотности.

ID *	Homework1	Homework2	Classwork3	Homework3	SIS1	SIS2	SIS3	SIS4	Conspect lectures	Midterm	Attestation	Quiz	SIS5	SIS6	SIS7	SIS8	SIS9	SIS10	Bonus conspect	Bonus surprise	Bonus game	Endterm	Attestations	TotalA	Total	Exam	Sum	Absences	Mark
19B03059	1,5	0	2	2	2	2	3	2	2	9	25,5	0	3	2	4	3	2	6	2	2	2	2,7	28,7	54,2	54,2	38	92,2		2 PASS
19B03050	0	0	2	2	2	2	3	3	0	10	24	3,7	0	2	3	3	2	3	0	2	2	2,5	23,2	47,2	47,2	40	87,2		5 PASS
19B03005	0	0	2	2	2	2	3	3	0	7	21	3,5	0	2	3	3	2	3	4	2	0	3	25,5	46,5	46,5	40	86,5		5 PASS
19B03009	0	2	2	0	1,9	2	3	3	0	9	22,9	4	3	2	4	3	2	6	0	2	2	2,5	30,5	53,4	53,4	36	89,4		4 PASS
19B03061	1,8	0	2	2	2	2	3	1,5	2	9	25,3	5	1,5	2	0	1,5	2	3	4	2	2	1,5	24,5	49,8	49,8	35	84,8		2 PASS
19B03018	2	2	2	2	2	2	3	3	2	11	31	5,5	3	2	4	3	2	6	2	2	3	4	36,5	67,5	60	40	100		0 PASS
19B03019	0	0	0	0	0	0	0	0	0	7	7	6	3	2	4	3	2	6	4	2	3	2,5	37,5	44,5	44,5	40	84,5		9 PASS
19B03018	2	0	2	0	0	0	1,5	1,5	2	8	17	5,5	1,5	2	0	2	2	0	2	2	4	3,5	24,5	41,5	41,5	32	73,5		6 PASS
19B03060	2	2	2	2	2	2	3	2	2	10	29	4,5	3	2	4	3	2	6	2	2	2	4	34,5	63,5	60	40	100		0 PASS
18BD1373	0	0	0	0	0,5	0,5	0	0	2	6	9	4,3	0	2	0	3	2	2	2	2	2	3,5	22,8	31,8	31,8	0	31,8		8 FAIL
19B03022	2	2	2	2	2	2	3	0	2	10	27	3	3	2	4	0	2	0	2	2	2	3,5	23,5	50,5	50,5	36	86,5		3 PASS
20B03014	0	0	0	0	0	0	0	0	0	6	6	5,5	0	2	0	3	2	2	0	2	0	0	16,5	22,5	22,5	0	22,5		14 FAIL
19B03052	2	0	2	0	2	0	0	0	2	11	19	6	3	2	2	0	0	0	0	2	5	4	24	43	43	36	79		9 PASS
19B03011	2	2	2	2	2	2	3	3	2	9	29	6	2	2	4	1	1	4	2	2	5	3,5	32,5	61,5	60	40	100		0 PASS
18BD1109	0	0	0	0	0	0	1,5	1,5	0	6	9	0	1	2	0	1	1	3	2	2	5	4	21	30	30	30	60		9 PASS
19B03023	0	0	2	0	0	0	0	0	0	6	8	3,5	0	0	0	0	0	0	4	0	0	0	7,5	15,5	15,5	0	15,5		17 FAIL
19B03020	2	2	2	2	0	1	3	3	2	7	24	5,8	2,5	1	3	1,5	0	2	0	2	5	0	22,8	46,8	46,8	40	86,8		4 PASS
19B03008	2	2	2	2	0	2	3	2	2	9	26	5,3	0	2	0	0	2	0	2	2	2	3	18,3	44,3	44,3	33	77,3		5 PASS
19B03018	2	2	2	2	2	2	3	3	2	9	29	4,5	2	2	4	1	1	4	2	2	2	3,7	28,2	57,2	57,2	35	92,2		0 PASS
19B03021	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		22 FAIL
19B03006	0	2	2	2	1,8	2	3	1,5	2	9	25,3	4,2	1,5	2	4	3	2	4	2	2	2	3,5	30,2	55,5	55,5	38	93,5		1 PASS
19B03061	1,8	2	2	2	2	2	3	1,5	2	9	27,3	4,8	1,5	2	0	1,5	2	3	2	2	2	4	24,8	52,1	52,1	38	90,1		1 PASS
19B03006	1,8	2	2	2	2	2	3	0	2	9	25,8	4	3	2	4	0	2	0	2	2	3	4	26	51,8	51,8	40	91,8		3 PASS
19B03060	0	0	0	0	0	0	1,5	1,5	0	6	9	3,5	0	2	0	3	2	0	7,5	2	0	2,5	22,5	31,5	31,5	0	31,5		11 FAIL
19B03033	1,8	2	2	2	2	1,5	3	0	2	10	26,3	3,3	2	2	0	1,5	2	3	2	2	3	3	23,8	50,1	50,1	38	88,1		2 PASS
19B03017	2	2	2	2	2	1,8	3	2	2	11	29,8	4,3	3	2	4	3	2	6	2	2	5	2,5	35,8	65,6	60	40	100		0 PASS
19B03037	1,8	0	2	2	2	2	3	2	2	7	23,8	4,5	3	2	4	3	2	6	2	2	2	4	34,5	58,3	58,3	38	96,3		1 PASS
19B03042	1,6	2	0	0	0		1,5	1,5	0	7	13,6	0	1	2	0	1	1	1	4,4	2	2	2	16,4	30	30	30	60		6 PASS
19B03019	1	2	2	2	0	1	3	3	2	8	24	5	2,5	1	3	1,5	0	2	2	2	2	3,5	24,5	48,5	48,5	36	84,5		2 PASS
19B03021	2	2	2	2	2	2	3	3	2	8	28	2,5	3	2	4	3	2	6	0	2	2	3	29,5	57,5	57,5	40	97,5		1 PASS

Р и с. 3. Фрагмент использованной многомерной модели базы данных
F i g. 3. Fragment of the multidimensional database model used

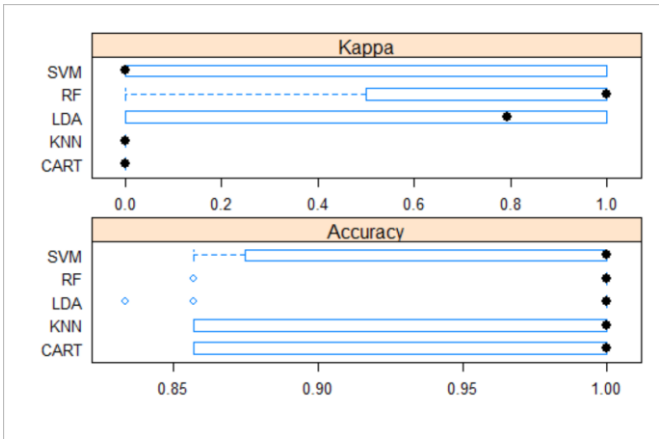
```
Accuracy
          Min.   1st Qu. Median      Mean 3rd Qu. Max. NA's
CART 0.8571429 0.8571429      1 0.9440476       1    1    0
LDA  0.8333333 1.0000000      1 0.9650794       1    1    0
SVM  0.8571429 0.9062500      1 0.9630952       1    1    0
KNN  0.8571429 0.8571429      1 0.9440476       1    1    0
RF   0.8571429 1.0000000      1 0.9857143       1    1    0

Kappa
     Min. 1st Qu.    Median      Mean 3rd Qu. Max. NA's
CART    0    0.00 0.0000000 0.0000000       0    0   18
LDA     0    0.00 0.7941176 0.5840336       1    1   16
SVM     0    0.00 0.0000000 0.3333333       1    1   18
KNN     0    0.00 0.0000000 0.0000000       0    0   18
RF      0    0.75 1.0000000 0.7500000       1    1   18
```

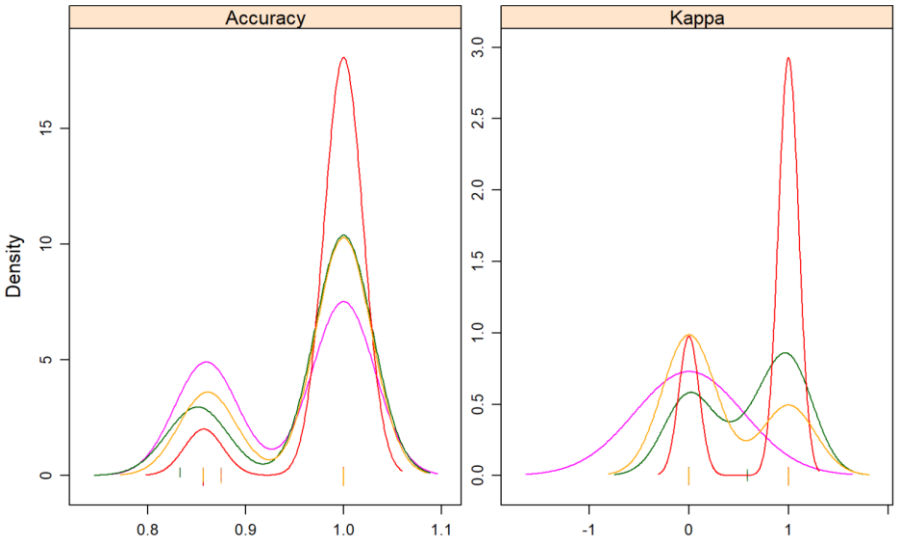
Р и с. 4. Результаты сравнения моделей классификации
F i g. 4. Results of the classification models comparison

Графики в рисунке 7, показывают как среднюю оценочную точность, так и 95% доверительный интервал. Следует отметить, что поведение алгоритмов, а также их предсказательные точности распределены по возрастанию, с модели классификации Cart до модели RF, точность которой наивысшая.

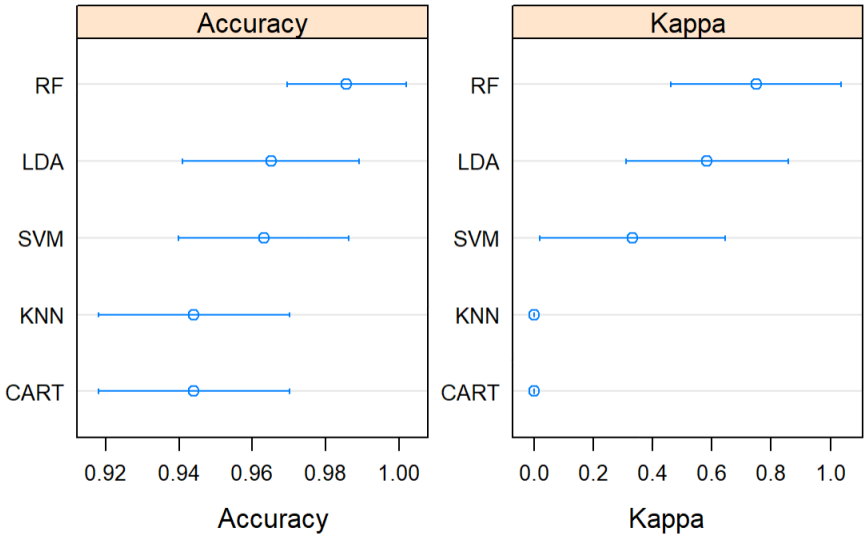
По рисунку 8, можно ответить на вопрос о том, коррелируют ли прогнозы, полученные с помощью двух разных алгоритмов. Если они слабо коррелированы, то являются хорошими кандидатами для объединения в ансамблевом прогнозировании. Например, по графикам можно отметить, что LDA и SVM сильно коррелируют, как и SVM и RF, в то время как SVM и CART выглядят слабо коррелированными.



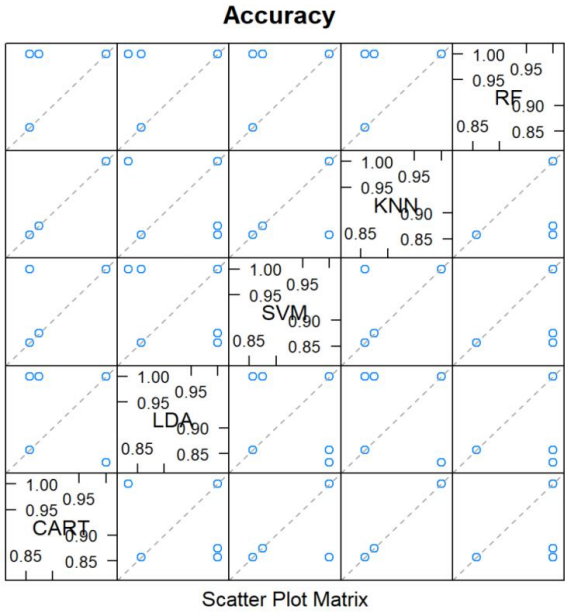
Р и с. 5. Результаты сравнения точности и Каппа-статистики моделей классификации
F i g. 5. Results of the comparison of accuracy and Kappa statistic for classification models



Р и с. 6. Результаты сравнения моделей классификации в R Density plots
F i g. 6. Results of the classification models comparison in R Density plots



Р и с. 7. Результаты сравнения моделей классификации в R Dot plots
F i g. 7. Results of the classification models comparison in R Dot plots



Р и с. 8. Матрица точечной диаграммы
F i g. 8. Scatter plot matrix

5 Обсуждение и заключение

В работе был сделан обзор на современные тенденции в построении моделей многомерных информационных систем с методами классификации. По определению модели классификации для многомерной модели информационных систем была отмечена важность методов предварительной обработки данных, которые помогут нам решить задачу редукции данных, тем самым оптимально классифицировать данные. После решения задачи редукции данных с методами выделения информативных признаков необходимо выбрать модель классификации. Также в работе отмечена актуальность использования иерархических методов классификации. Следовательно, в работе в качестве базы данных был взят журнал успеваемости студентов технического университета программы бакалавриата. В рамках работы проведена сравнительная оценка моделей классификации CART, LDA, RF, KNN, SVM в среде программного обеспечения Rstudio. Так как многомерные

информационные системы появились как мощный подход к облегчению процессов принятия решений, необходимо выбрать оптимальную модель классификации, при этом предоставляя всестороннее представление о сложных наборах данных. Одной из современных тенденций в построении моделей данных многомерных информационных систем является использование ансамблей классификаторов. В целом, современные тенденции в построении моделей данных многомерных информационных систем с использованием методов классификации направлены на улучшение точности и эффективности классификации, адаптации к различным типам данных и структурам, а также расширение возможностей использования моделей в реальных приложениях. С учетом развития технологий и роста объемов данных развитие методов классификации будет продолжаться и в дальнейшем.

References

1. Hasan F. A Review Study of Information Systems. *International Journal of Computer Applications*. 2018;179:15-19. https://doi.org/10.5120/ijca2018916307
2. Palominos F.E., Córdova F., Durán C., Nuñez B. A simpler and semantic multidimensional database query language to facilitate access to information in decision-making. *International journal of computers communications and control*. 2020;15(4):3900. https://doi.org/10.15837/ijccc.2020.4.3900
3. Fan C., Chen M., Wang X., Wang J., Huang B. A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data. *Frontiers in Energy Research.* 2021;9:652801. https://doi.org/10.3389/fenrg.2021.652801
4. Patro S.G.K., Sahu, Kumar Sahu K. Normalization: A Preprocessing Stage. *International Advanced Research Journal in Science, Engineering and Technology*. 2015;2(3):20-22. https://doi.org/10.17148/IARJSET.2015.2305
5. Godd E.F. A Relation Model of Data for Large Shared Data Banks. *Communications of the ACM*. 1970;13(6):377-387. https://doi.org/10.1145/362384.362685
6. Han J., Kamber M., Pei J. Data Mining: Concepts and Techniques. 3rd ed. Morgan Kaufmann; 2012. https://doi.org/10.1016/C2009-0-61819-5
7. Aggarwal C.C. Data Mining: The Textbook. Cham: Springer; 2015. 734 p. https://doi.org/10.1007/978-3-319-14142-8
8. Breiman L. Random Forests. *Machine Learning*. 2001;45:5-32. https://doi.org/10.1023/A:1010933404324
9. Cortes C., Vapnik V. Support-vector networks. *Machine Learning*. 1995;20:273-297. https://doi.org/10.1007/BF00994018
10. Cover T., Hart P. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*. 1967;13(1):21-27. https://doi.org/10.1109/TIT.1967.1053964
11. Fisher R.A. The use of multiple measurements in taxonomic problems. *Annals of Eugenics.* 1936;7:179-188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x
12. Quinlan J.R. Induction of decision trees. *Machine Learning.* 1986;1:81-106. https://doi.org/10.1023/A:1022643204877
13. Bishop C.M. Pattern Recognition and Machine Learning. *Information Science and Statistics.* New York, NY: Springer; 2006. 778 p.
14. Liu H., Motoda H. Feature Selection for Knowledge Discovery and Data Mining. *The Springer International Series in Engineering and Computer Science*. Vol. 454. New York, NY: Springer; 1998. 214 p. https://doi.org/10.1007/978-1-4615-5725-8
15. Guyon I., Elisseeff A. An introduction to variable and feature selection. *Journal of Machine Learning Research*. 2003;3:1157-1182. https://doi.org/10.1162/153244303322753616
16. Chandola V., Banerjee A., Kumar V. Anomaly detection: A survey. *ACM Computing Surveys*. 2009;41(3):15. https://doi.org/10.1145/1541880.1541882
17. Domingos P. A few useful things to know about machine learning. *Communications of the ACM*. 2012;55(10):78-87. https://doi.org/10.1145/2347736.2347755
18. Dietterich T.G. Ensemble Methods in Machine Learning. In: Multiple Classifier Systems. MCS 2000. *Lecture Notes in Computer Science*. Vol. 1857. Berlin, Heidelberg: Springer; 2000. p. 1-15. https://doi.org/10.1007/3-540-45014-9_1
19. Zhou Z.-H. Ensemble Methods: Foundations and Algorithms. New York: Chapman and Hall/CRC; 2012. 236 p. https://doi.org/10.1201/b12207
20. Sokolova M., Lapalme G. A systematic analysis of performance measures for classification tasks. *Information Processing & Management.* 2009;45(4):427-437. https://doi.org/10.1016/j.ipm.2009.03.002
21. Japkowicz N., Stephen S. The class imbalance problem: A systematic study. *Intelligent Data Analysis*. 2002;6(5):429-449.
22. Chen C., Liaw A., Breiman L. Using Random Forest to Learn Imbalanced Data (666). University of California, Berkeley. 2004;110(1-12):24. Available at: https://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf (accessed 29.07.2023).

23. Aggarwal C.C. Outlier Analysis. Cham: Springer; 2017. 466 p. https://doi.org/10.1007/978-3-319-47578-3
24. Kimball R., Ross M. The Data Warehouse Toolkit. John Wiley & Sons, Inc.; 2013. 608 p.
25. Inmon W.H. Building the Data Warehouse. John Wiley & Sons, Inc., 2005. 432 p.

Поступила 29.07.2023; одобрена после рецензирования 06.09.2023; принята к публикации 21.10.2023.

Submitted 29.07.2023; approved after reviewing 06.09.2023; accepted for publication 21.10.2023.

Об авторе:

Муратова Жансая Муратовна, старший преподаватель, НАО «Западно-Казахстанский университет имени Махамбета Утемисова» (090000, Республика Казахстан, г. Уральск, пр. Н. Назарбаева, д. 162), **ORCID: https://orcid.org/0000-0003-2841-0178,** jansik.86@mail.ru

Автор прочитал и одобрил окончательный вариант рукописи.

About the author:

Zhansaya M. Muratova, University Senior Lecturer, Mahambet Utemisov West Kazakhstan University (162 N. Nazarbayev Ave., Uralsk 090000, Republic of Kazakhstan), **ORCID: https://orcid.org/0000-0003-2841-0178,** jansik.86@mail.ru

The author has read and approved the final manuscript.