ЕРМАКОВА Н.А.,
ФГАОУ ВО РНИМУ им. Н.И. Пирогова, Москва, Россия, e-mail: ermakova_na@rsmu.ru

ГУСЕВ А.В.,
к.т.н., ФГБУ «Центральный научно-исследовательский институт организации и информатизации здравоохранения» Минздрава России, Москва, Россия, e-mail: agusev@webiomed.ru

РЕБРОВА О.Ю.,
д.м.н., ФГАОУ ВО РНИМУ им. Н.И. Пирогова, Москва, Россия; ГНЦ РФ ФГБУ «НМИЦ эндокринологии» Минздрава России, Москва, Россия, e-mail: rebrova_oyu@rsmu.ru

ОШИБКИ В ДАННЫХ РЕАЛЬНОЙ КЛИНИЧЕСКОЙ ПРАКТИКИ: ОБЗОР ЛИТЕРАТУРЫ

DOI: 10.25881/18110193_2024_1_28

Аннотация. *В последнее время возрастает интерес к использованию больших данных реальной клинической практики для разработки систем искусственного интеллекта в интересах врачебной практики – моделей диагностики заболеваний и состояний и прогноза их течения. При этом качество этих данных обычно невысоко из-за допускаемых ошибок при вводе, неоптимальной архитектуры информационных систем, отсутствия стандартизации и др. В обзоре рассмотрены критерии надежности данных реальной практики, наиболее часто встречающиеся проблемы и способы их устранения: оценка соответствия набора данных дизайну разрабатываемой модели, выявление и удаление дублирующих записей в наборах данных, обработка пропущенных значений, обнаружение и обработка выпадающих значений, выявление и обработка несогласованности в данных. Делается вывод о том, что требуется дальнейшее развитие методик создания наборов данных на основе реальной клинической практики в части повышения их качества, так как наличие ошибок в них может приводить к снижению качества создаваемых моделей машинного обучения для диагностики и прогнозирования.*

Ключевые слова: *реальная клиническая практика; качество данных; пропущенные значения; противоречия; выпадающие значения; дублирующие записи; электронные медицинские карты; искусственный интеллект.*

Для цитирования: *Ермакова Н.А., Гусев А.В., Реброва О.Ю. Ошибки в данных реальной клинической практики: обзор литературы. Врач и информационные технологии. 2024; 1: 28-43.*

ERMAKOVA N.A.,
Pirogov Russian National Research Medical University, Moscow, Russia, e-mail: ermakova_na@rsmu.ru

GUSEV A.V.,
PhD, Federal Research Institute for Health Organization and Informatics, Moscow, Russia,
e-mail: agusev@webiomed.ai

REBROVA O.YU.,
DSc, Pirogov Russian National Research Medical University, Moscow, Russia,
e-mail: rebrova_oyu@rsmu.ru

ERRORS IN REAL-WORLD DATA: A REVIEW

DOI: 10.25881/18110193_2024_1_28

Abstract. *There is increasing interest in using big data of real clinical practice to develop artificial intelligence systems for diagnostic and predictive models of diseases and conditions. At the same time, the quality of this data is usually low due to errors during input, suboptimal architecture of information systems, lack of standardization, etc. The review examines criteria for the reliability of real-world data, the most common problems, and ways to eliminate them: assessing the compliance of the data set with the design of the model being developed, identifying, and removing duplicate records in data sets, handling missing values, detecting, and handling outliers, identifying and handling inconsistencies in data. We conclude that further development of methods for creating data sets based on real-world data is required in terms of improving their quality, can lead to lower quality of the created machine learning models for diagnosis and prognosis.*

Keywords: *Real-world data; data quality; missing values; contradictions; outliers; duplicate entries; electronic health record; artificial intelligence.*

For citation: *Ermakova N.A., Gusev A.V., Rebrova O.Yu. Errors in real-world data: a review. Medical doctor and information technology. 2024; 1: 28-43.*

ВЕДЕНИЕ

В последнее время возрастает интерес к системам поддержки принятия врачебных решений (СППВР), разработанным на данных реальной клинической практики (ДРКП, англ. real-world data, RWD). ДРКП определяют как данные о состоянии здоровья пациентов и/или об оказании медицинской помощи, полученные из различных источников вне рамок клинических исследований [1]. Различные источники ДРКП лежат в основе всех процессов и этапов получения доказательств – real-world evidence (RWE), которые трактуются Британской медицинской академией наук как доказательства, полученные на основе клинически значимых данных, собранных вне контекста рандомизированных контролируемых испытаний [2, 3].

Одним из основных источников ДРКП является электронная медицинская карта (ЭМК) [4]. ЭМК пациента хранит информацию обо всех случаях обращения пациента за медицинской помощью в пределах одной медицинской организации (МО). По классификации источников ДРКП ЭМК относится к вторичным источникам данных (первичным, в свою очередь, является проспективный сбор данных в условиях рутинного лечебно-диагностического процесса), то есть это данные, изначально собранные для других целей, без исходного фокуса на исследовательские аспекты [2, 5]. ГОСТ Р ИСО/ТС 18308-2008 «Информатизация здоровья. Требования к архитектуре электронного учета здоровья» предусматривает использование обезличенных записей из ЭМК в качестве источника данных для проведения научных и клинических исследований [6].

По объему ЭМК превосходят многие существующие источники данных, а повторное их использование может позволить снизить затраты, которые необходимы при проведении клинических исследований. Исследования, в которых используются ДРКП, не требуют набора пациентов, что является трудоемким дорогостоящим процессом, но позволяют получить представления о состоянии здоровья, результатах обследования и лечения различных групп населения, репрезентативных по отношению к общей популяции пациентов. Вторичное использование данных ЭМК является многообещающим шагом на пути к снижению затрат на исследования и ускорению темпов развития отрасли здравоохранения [7].

Несмотря на некоторые преимущества исследований, основанных на ДРКП, такие как длительные периоды наблюдения, выявление отсроченных эффектов лечения, высокая внешняя валидность, использование ЭМК подвергается критике по ряду причин: наличие большого объема неструктурированной информации, пропуски в данных и некачественное заполнение ЭМК, повторное использование ранее введенной информации, включая подстановку уже введенной информации из других документов или разделов медицинской информационной системы (МИС) в новый документ/раздел, отсутствие стандартизации при сборе ряда данных, низкая методологическая строгость таких исследований и другие [3, 4, 8].

Сведения, поступающие из ЭМК, образуют т.н. озера данных, которые впоследствии используют для формирования наборов данных для конкретного исследования. После сбора медицинских документов из различных МО для формирования базы данных из неструктурированной медицинской информации используются методы машинного обучения, в т.ч. Natural language processing (NLP).

Каждый этап процесса получения необходимого набора данных может сопровождаться появлением ошибок в данных [4]. Для того чтобы снизить риски появления ошибок, вызываемых указанными выше проблемами, и повысить доверие к ЭМК как к источнику данных для использования технологий искусственного интеллекта и машинного обучения, необходимо обеспечивать не только централизованный сбор данных и их хранение, но и автоматизированный контроль качества данных, поступающих в базу данных и отбираемых для построения моделей.

1. Проблема качества данных реальной клинической практики

Во всем мире отмечают, что из-за различий в приоритетах между клиническими и исследовательскими задачами, ДРКП не регистрируются с той же тщательностью, что и данные клинических исследований. Именно поэтому некоторые исследователи являются противниками вторичного использования данных ЭМК в качестве ДРКП. Такие опасения существуют с тех пор, как

впервые были разработаны ЭМК, и до настоящего времени нет единого мнения о том, что же такое «качество данных» в их контексте [7, 9].

Проблемы с качеством ДРКП являются следствием многих факторов: высокая рабочая нагрузка на медицинских работников, непрямой сбор данных, дезинформация от пациентов, небрежность документирования [10, 11]. Помимо социальных факторов в литературе отмечаются также технические (связанные с конструкцией информационной системы и с автоматизированным извлечением данных из структурированной и неструктурированной медицинской информации) и организационные (связанные со сбоями в рабочем процессе, недостаточной компьютерной грамотностью персонала, ротацией персонала), что также непреднамеренно способствует сбору некачественных данных о состоянии здоровья пациентов, особенно с точки зрения их полноты, непротиворечивости и правдоподобия [12, 13]. Несмотря на то, что с развитием МИС записи по большей части стали фиксироваться с использованием форм и шаблонов, тем не менее разработчики МИС обычно оставляют свободные поля для ввода дополнительной (непредусмотренной заранее и, соответственно, не формализованной) клинической информации, которые часто (в силу удобства ввода) используются для внесения основной информации вместо заполнения многих необязательных формализованных полей.

Было проведено несколько исследований, в которых подчеркивается ведущая роль оценки качества данных и повышения качества информации, особенно в медицинской области. Для оценки пригодности ЭМК в применении к исследованиям различные авторы предлагают свои критерии качества данных. Основными из них являются полнота, которая определяет степень регистрации всех необходимых данных, точность, которая характеризует корректность собранных реальных данных, согласованность и своевременность данных [14, 15].

При рассмотрении общих вопросов работы с ДРКП Li Cai, Yangyong Zhu [16] предложили стандарт качества больших данных. Их концепция заключается в том, что стандарт качества данных состоит из 5 интегральных показателей: доступность, удобство использования, надежность, актуальность и качество представления. Основные критерии качества данных по мнению этих авторов представлены в таблице 1.

Зарубежные исследователи по-разному оценивают качество ДРКП в зависимости от клинической направленности или используемых МИС МО. Так, например, полнота колеблется от 1,1% до 100%, а точность данных от 44% до 100% [7].

В исследованиях вопросов качества ДРКП в литературе низкое качество связывают с появлением таких ошибок, как пропущенные значения, орфографические ошибки, использование синонимов и омонимов, некорректное значение и др. Перечень ошибок, описанных в различных источниках, представлен в таблице 2 [8, 16, 17].

В работе М.В. Ионова и соавт. [18] по внедрению СППВР для повышения качества медицинских данных пациентов с артериальной гипертензией авторы исследовали ошибки, возникающие при ведении ЭМК, связанные с неполнотой или несогласованностью информации в данных об осложнениях, обусловленных артериальной гипертензией за период 2010–2015 гг. Всего было идентифицировано девять видов ошибок, которые были связаны с отсутствием сведений о каких-либо назначениях, наличием опечаток, появлением заключений без лабораторного подтверждения, отсутствием диагноза при наличии лабораторного подтверждения и др.

Таблица 1 — Критерии надежности данных [16]

Критерий	Комментарий
Точность	Для определения точности значения оно сравнивается с известным эталоном
Согласованность	Определяется полнотой и корректностью логической связи между скоррелированными данными
Целостность	Данные имеют полную структуру, а значения данных стандартизованы в соответствии с моделью данных и типом данных
Полнота	Значения всех компонентов одного элемента данных являются действительными (нет отсутствующих значений в определенном компоненте)

Таблица 2 — Перечень ошибок, выявленных при обзоре литературы

Вид ошибки	Пример
Отсутствующие значения	Незаполненный атрибут (например, пол пациента)
Неверные значения	Атрибут содержит значение, которое не является правильным
Нарушение уникальности	Два (или более) кортежа имеют одинаковое значение в атрибуте, содержащем уникальные значения
Противоречивые значения	Несогласованность значений атрибута в одном или разных источниках (например, в разных документах врачи могли указать противоречивые характеры заболевания – острый и хронический)
Неоднородность единиц измерения	Использование различных единиц измерения в связанных атрибутах как внутри одного источника данных, так и в различных источниках данных (например, в разных источниках данных представление температуры в шкалах Цельсий/Фаренгейт)
Неоднородность представления	Использование разных наборов значений для кодирования одного и того же свойства реального мира внутри одного источника данных или в разных источниках данных (например, в одном источнике пол может быть представлен значениями 1 и 2, а в другом источнике – значениями M и F)
Наличие омонимов	Использование синтаксически одинаковых значений с разными лексическими значениями среди связанных атрибутов из нескольких источников данных (например, вентиляция имеет по крайней мере два разных лексических значения, одно относится к биологическому явлению дыхания, а другое относится к потоку воздуха в окружающей среде)

Накопление большого объема данных затрудняет ручную обработку, в том числе и в области очистки данных и формирования наборов для исследований. Наличие ошибок приводит к использованию некорректной информации или использованию меньшего количества признаков при построении моделей из-за неполноты данных или их недостаточной детализации, и вследствие отсутствия необходимой подготовки данных исследуемая выборка может сократиться в несколько раз. Противоречия в информации из разных источников вызывает недоверие к этим источникам, сомнения и трудности при их выборе [16, 17].

Ряд авторов указывают на необходимость проектирования автоматизированных платформ, интегрирующих в себе возможности реализации всех процессов управления данными. В функции системы управления данными должно входить: (1) автоматическое обнаружение отсутствующих значений, (2) обнаружение и удаление выбросов, (3) обнаружение сходства, т.е. идентификации повторяющихся полей и сильно коррелированных распределений, (4) идентификация атрибутов и группировка данных [19, 20]. Автоматизированное обнаружение ошибок поможет снизить время аналитиков данных на подготовку необходимых для исследования наборов данных и повысит качество доказательств, получаемых на основе этих данных.

2. Подходы к работе с ошибками в данных

Методы, применяемые для обнаружения и обработки ошибок, зависят от типов данных и типа ошибки. Наиболее часто встречающимися проблемами считаются несоответствие набора данных дизайну модели, дублирующие записи данных, пропущенные значения, нереалистичные значения параметров, несогласованность данных [7, 8, 15, 19].

2.1. Соответствие набора данных дизайну модели

При разработке моделей машинного обучения важно исключить возможность предвзятой интерпретации результатов и добиться максимальной их объективности. С точки зрения подготовки набора данных ключевой момент состоит в проспективном планировании сбора данных, идентификации соответствующей базы

данных, минимизации систематических ошибок. Видами систематических ошибок (ограничениями обсервационных исследований) являются:

(1) систематическая ошибка отбора (связанная с наличием различий в исходных характеристиках групп сравнения),

(2) систематическая ошибка информации (связанная с погрешностью определения исхода или воздействия, приводящая к неправильному разбиению на группы),

(3) систематическая ошибка времени (связанная с установлением неверной последовательности событий),

(4) систематическая ошибка наблюдателя (когда интересующее событие с меньшей или большей вероятностью будет зафиксировано в одной группе, чем в другой) [21].

Чтобы снизить риск систематических ошибок, необходимо выявить и надлежащим образом учесть возможные искажающие факторы с помощью стратегий сопоставления или корректировки. Основным шагом в таком случае является подготовка подробного плана извлечения данных из базы данных, точное определение исследуемой популяции и подгрупп, которые представляют интерес [20, 22].

Разработку моделей машинного обучения на ДРКП относят к наблюдательным исследованиям [20]. При проведении высококачественных наблюдательных исследований цели и план анализа должны быть заранее определены. И хотя наблюдательные исследования часто используют ретроспективно собранные данные, они также должны планироваться. Это помогает обеспечить включение всех потенциально значимых переменных, которые необходимы для характеристики пациентов [20–22].

Наблюдательные исследования, в свою очередь, могут быть поперечными (одномоментными) и продольными (динамическими – про- и ретроспективными). Для наблюдательных исследований важным фактором является время, в которое оцениваются результаты, и период сбора данных. Указанный период сбора данных должен быть ограничен определенным интервалом времени в зависимости от решаемой задачи. В первом случае, который обычно применим в отношении разработки моделей для диагностики, происходит однократное измерение показателей и определение связи между ними. В случае продольных исследований проводится сравнение показателей, полученных за определенный период времени до наступления целевого события [23, 24].

2.2. Выявление и удаление дублирующих записей в наборах данных

При сборе данных из разных источников в одно хранилище объем данных растет, но при этом возрастает вероятность дублирования информации. Дублирование снижает качество результата интеллектуального анализа данных и делает его менее достоверным [25, 26].

Причиной большого количества дубликатов в ДРКП является тот факт, что для каждого случая обращения за медицинской помощью вносится своя запись, например, пациенту с диабетом при каждом посещении будет вноситься запись о его диабете. При этом, если в рамках одного медицинского учреждения при формировании набора данных можно сопоставить уникальные идентификаторы пациентов, то при работе с хранилищем данных, собранных из разных учреждений, идентификаторы одного и того же пациента могут отличаться или отсутствовать вовсе [26, 27].

В таком случае аналитики прибегают к обнаружению дубликатов на основе правил, например, предлагая для различных наборов данных сопоставление всех наблюдений по набору признаков, которые в своей совокупности являются наиболее уникальными для объекта [26, 28]. В одном из исследований авторы считали процент повторяющихся записей для разных учреждений по совпадению ФИО и даты рождения пациентов. При этом совпадение имени и фамилии колебалось от 16,49% до 40,66% записей, а при включении даты рождения показатели снизились до диапазона от 0,16% до 15,47% [28].

В вопросе обнаружения повторяющихся записей, как и практически в любом вопросе подготовки и анализа данных, есть две проблемы, которые следует учитывать: точность и скорость. При этом исследователи делают важный акцент не только на обнаружении дубликатов, но и на их устранении в том понимании, что для грамотной предварительной обработки данных необходим осмысленный выбор повторяющегося наблюдения, который будет сохранен при исключении оставшихся. Критериями в пользу

выбора наблюдения выступают полнота данных в одном наблюдении и качество этих данных [26].

2.3. Обработка пропущенных значений

Наличие пропусков в наборе данных приводит к ограничению доступных методов интеллектуального анализа. Несмотря на существование небольшого количества алгоритмов машинного обучения, таких как деревья решений, нейронные сети, которые могут принимать на вход данные с пропусками, создавать полные наборы данных нужно, чтобы расширять количество возможных для применения алгоритмов машинного обучения и сравнивать их точность.

Проблема отсутствующих значений при предварительной подготовке данных обычно решается с помощью двух подходов: удаление объектов при наличии пропуска хотя бы в одном признаке; заполнение пропусков с помощью значений, полученных на основе исходного набора данных [29–31]. Сами методы обработки отсутствующих данных должны удовлетворять трем правилам: они не должны изменять распределений признаков; при применении методов связь между признаками должна быть сохранена; методы не должны быть слишком сложными в вычислительном отношении [29, 30].

Еще в 1976 году Дональд Рубин формализовал классификацию отсутствующих данных, выделив три категории [29]:

1. Отсутствующие полностью случайно (missing completely at random; MCAR);
2. Отсутствующие частично случайно (missing at random; MAR);
3. Отсутствующие неслучайно (not missing at random; NMAR).

С тех пор считается, что в процессе обработки первых двух категорий вероятность пропуска в данных определяется на основе информации, содержащейся в наборе данных, поэтому удаление пропусков не приводит к значительному искажению. Третья категория предполагает, что пропуски появляются в зависимости от других факторов, не представленных в наборе данных, поэтому их удаление может вызвать значительное искажение [29, 30].

В применении к анализу ДРКП, подход, связанный с удалением наблюдений с пропусками (методы «complete case analysis» и «pairwise deletion»), хоть и является наиболее распространенным, однако приводит к потере статистической мощности и возникновению систематической ошибки, связанной с тем, что пациенты, у которых отсутствует какая-либо информация, могут значительно и систематически отличаться от пациентов, имеющих более полную информацию о состоянии их здоровья (например, пациенты с более тяжелыми заболеваниями чаще обследуются у врачей и, соответственно, имеют меньше отсутствующих данных) [32–35]. Другой подход обработки пропусков (заполнение на основе имеющихся данных) включает в себя алгоритмы одиночной импутации (single imputation) и множественной импутации (multiple imputation). Такие методы позволяют исследователям избежать потери информации, но и они также могут не давать необходимой точности [31, 36–38].

Одиночная замена пропущенных значений представляет собой замену, основанную на оценке истинного значения наблюдаемой переменной. Основным недостатком методов одиночной импутации является то, что значение, которым заменяется пропуск, рассматривается как известное и не всегда отражает изменчивость выборки в рамках модели данных. В зависимости от задачи применяются следующие методы: замена средним, замена модой, замена медианой, замена последним наблюдением (last observation carried forward, LOCF), метод горячей колоды [29, 30, 40].

Методы заполнения средним, модой или медианой заключаются в заполнении любых пропущенных значений соответствующей мерой центральной тенденции для непропущенных значений по признаку. Предполагается, что меры центральной тенденции переменной являются наилучшей оценкой для наблюдения с пропуском, но, несмотря на простоту, при наличии большого количества пропусков эта стратегия может сильно исказить распределение переменной [39–41]. Есть две основных вариации данного метода: заполнение общей по выборке мерой центральной тенденции или заполнение значениями, рассчитанными для каждого из прогнозируемых классов. Первый является более консервативным и предпочтительным [42].

Другим методом одиночной импутации является метод LOCF в динамических исследованиях с многократным наблюдением, в основе которого лежит подстановка последнего наблюдаемого значения вместо пропуска. Чаще всего метод применяется при анализе данных рандомизированных контролируемых испытаний, например, при разработке лекарственных препаратов, или при анализе временных рядов. Основной недостаток метода заключается в предположении о неизменности наблюдения, что зачастую является нереалистичным [41, 43, 44].

Еще один подход, используемый для замены пропусков, – это метод горячей колоды (Hot-Deck). Метод был разработан в бюро переписи населения США и применялся для восстановления пропусков в данных, собранных путем анкетирования [41]. Основная идея метода – заполнение отсутствующих значений признака в наблюдении значением другого признака полного наблюдения, при условии минимального расстояния между наблюдениями, мера которого выбирается исходя из задачи [41, 45].

Методы множественной импутации (MI) основаны на обработке данных путем последовательных итераций и направлены на учет неопределенности в отношении отсутствующих данных. В основе этой группы методов лежит создание нескольких различных наборов данных с заполненными пропусками и последующее объединение результатов [36, 46, 47]. Одним из примеров, который часто применяется при анализе наборов ДРКП, является метод множественной импутации с использованием цепных уравнений (multiple imputation by chained equations, MICE) [46, 48, 49]. В процедуре MICE последовательно запускаются регрессионные модели, где каждая переменная с пропусками моделируется в зависимости от других переменных. Таким образом, переменные могут быть смоделированы в соответствии с их распределением (бинарные переменные смоделированы с использованием логистической регрессии, а непрерывные переменные смоделированы с использованием линейной регрессии) [50, 51].

При выборе стратегии заполнения пропусков важно определиться с методом, учитывая количество пропусков в данных. В работе [42] авторы рекомендуют опираться на следующее правило: если доля наблюдений, в которых отсутствует какая-либо переменная, меньше 0,03, то допустимо использование анализа полных наблюдений или для непрерывных переменных – заполнение пропусков медианой, а для категориальных переменных – модой, если же доля наблюдений с пропусками больше 0,03, то рекомендуется применять методы множественной импутации.

Отсутствие данных является проблемой практически в любых исследованиях медицинской сферы, однако этот вопрос изучается лишь в небольшом количестве публикаций. Burton A. и Altman D.G. в своей работе показывают, что лишь в 10 из 100 исследований сравнивают характеристики или исходы между случаями с полными или неполными данными. Авторы также демонстрируют, что наиболее популярным методом в медицине является анализ полных случаев, что связано с простотой его применения и доступностью в любых программах для статистического анализа данных, при этом большинство статей не содержит упоминаний о методах обработки отсутствующих данных [52]. Это подтверждают и другие более поздние исследования, в которых хоть и фиксируется рост количества статей, где при анализе данных применяются методы заполнения (чаще всего – заполнение средним, очень редко – методы множественной импутации), анализ полных случаев все равно используется чаще [53]. При этом методы заполнения пропусков применяются вне зависимости от количества пропусков или от механизмов их возникновения, что с позиции современной статистики считается ошибочным.

2.4. Обнаружение и обработка выпадающих значений

В настоящее время принято различать две категории выпадающих количественных данных [54]:

1. Выбросы, которые с точки зрения биологии имеют место в исключительных условиях;
2. Экстремальные (нереалистичные) значения, которые являются точно ошибочными и не могут существовать с точки зрения биологии.

В случае категориальных признаков в наборах ДРКП нереалистичными обычно являются значения, которые не предусмотрены в соответствующем справочнике, поэтому для обнаружения несоответствий достаточно организовать

проверку на наличие значения в справочнике. Вместе с этим возможно применение более сложных методов, способных воспроизводить данные и вычислять оценку аномальности категориального значения, например, метод k-ближайших соседей, моделей на основе машинного обучения и др. [55–57].

а. Метод обнаружения выбросов и нереалистичных значений на основе пороговых значений

Из существующих методов наиболее широко используемый подход в оценке качества для непрерывных данных заключается в обнаружении выбросов и нереалистичных чисел с использованием заранее определенных пороговых значений. Знания, относящиеся к конкретной предметной области, могут быть использованы для обнаружения ошибок в данных, которые невозможно идентифицировать с помощью статистического анализа [45].

Для построения правил производится экспертная оценка интервалов и составляются таблицы знаний. В таблицах знаний хранятся минимальные и максимальные значения признаков, допустимых с точки зрения решаемой задачи. Это позволяет при использовании метода разделить задачу обнаружения нереалистичных значений и задачу обнаружения выбросов. В комбинации с одномерными или многомерными статистическими методами метод пороговых значений применяется в компьютерных системах безопасности, для обнаружения повреждений в крупных строительных объектах, а также в медицине при анализе лабораторных параметров и электрофизиологических сигналов [45, 58–60].

б. Одномерные статистические методы обнаружения выпадающих значений

Помимо экспертной оценки пороговых значений для определения неправдоподобных значений и выбросов используются методы одномерной статистики. К группе методов одномерной статистики относят Z-оценку, Метод Тьюки, определение пороговых значений с помощью процентилей, фильтр Хампеля, а также статистические тесты Граббса, Диксона (Q-критерий), Рознера и др. [61–64].

Одномерные статистические методы позволяют обнаружить ошибки в отдельных признаках. Стандартная или z-оценка применяется в случае, если исследуемый признак имеет нормальное распределение. Смысл стандартной оценки состоит в том, что она показывает, на сколько стандартных отклонений определенное значение выборки больше или меньше среднего. Выбросами считаются значения, выходящие за установленный порог (обычно 3 стандартных отклонения) [64, 65]. Поскольку большинство данных реального мира, и в частности медицинских данных, не характеризуются нормальным распределением, то рекомендуют применять другие одномерные методы, например, фильтр Хампеля, который описывает выпадающие значения, как значения, выходящие за интервал плюс-минус 3 медианы абсолютных отклонений (median absolute deviation, MAD), или метод, предложенный американским математиком Джоном Тьюки, основанный на расчете межквартильного размаха (межквартильного диапазона) [54, 65–67].

Критерии Граббса и Диксона применимы для проверки крайних элементов упорядоченной по возрастанию выборки, распределение которой соответствует нормальному. Статистика теста Граббса определяется как наибольшая абсолютная величина отклонения от среднего значения выборки в единицах стандартного отклонения выборки. Q-критерий предложен Уилфридом Диксоном и оценивает отнесение значения к выпадающему путем расчета абсолютной разницы между предполагаемым выпадающим значением (крайним значением) и ближайшим к нему значением. Если рассчитанное значение статистики Граббса или Диксона превышает табличное, то значение считается выпадающим [64, 68]. Несмотря на ограничение в применимости метод Диксона используется в медицинских исследованиях в настоящее время и некоторыми исследователями считается предпочтительным по отношению к другим стандартным одномерным методам для решения ряда задач в области лабораторных и генетических исследований [69–71].

в. Многомерные статистические методы для обнаружения выпадающих значений

В последнее время развиваются и применяются для выявления выпадающих значений в наборах ДРКП многомерные статистические

методы и методы машинного обучения [72–74]. В машинном обучении существует два основных подхода: обучение с учителем (контролируемое обучение) и обучение без учителя (неконтролируемое обучение). И хотя неконтролируемое обучение характеризуется меньшей точностью, в применении к обнаружению выпадающих значений обычно придерживаются второго подхода, в связи с отсутствием размеченных для обучения с учителем данных.

Из существующих многомерных алгоритмов для выявления ошибок используют:

1. Методы на основе близости и плотности (метод k-ближайших соседей, локальный коэффициент выбросов);
2. Линейные методы (метод главных компонент, метод опорных векторов одного класса);
3. Вероятностные модели (обнаружение выбросов на основе копулы, обнаружение выбросов на основе угла);
4. Изолирующий лес;
5. Нейронные сети.

В 1998 г. Knorr E.M. и Ng R.T. предложили метод определения выбросов на основе расстояния [75], который был позднее расширен за счет использования расстояния к k-ближайшим соседям, чтобы ранжировать выбросы [76]. В моделях, основанных на расстоянии, основная гипотеза состоит в том, что неправдоподобные наблюдения достаточно редко встречаются, что при выполнении кластерного анализа на большом наборе данных делает их разреженными, а значит и детектируемыми. Так, например, в исследовании [73] применение алгоритма к данным, извлеченным из ЭМК, показывает высокую чувствительность и специфичность и меньшее число ложноположительных результатов.

Еще один метод кластеризации для обнаружения выбросов в многомерных данных, в основе которого лежит мера плотности, был предложен в 2000 г. Breunig M.M с соавторами. Локальный коэффициент выбросов (local outlier factor, LOF) – это подход, при котором вычисляется отклонение локальной плотности данной точки данных по отношению к ее соседям. Авторы утверждают, что LOF является первой концепцией, которая количественно определяет, насколько удален объект [77]. Чем выше значение, тем более вероятно, что точка является выпадающей. В медицинской литературе данный подход применяли для автоматического обнаружения артефактов ЭЭГ сигналов [78], аномалий на ЭКГ сигналах [79], а также при анализе генетических полиморфизмов [80].

Некоторые авторы [80–82] в своих задачах используют робастный метод главных компонент, который предложили в 2005 г. Hubert et al. [83, 84]. Метод предполагает, что при вложении исходных данных в подпространство более низкой размерности, выбросами будут являться точки, которые естественным образом не соответствуют модели, то есть ведут себя иначе, чем другие [85, 86].

Опираясь на размышления о том, что существующие методы детекции выбросов оптимизированы для профилирования нормальных экземпляров данных и, как следствие, вызывают много ложных срабатываний, в 2008 г. Zhi-Hua Zhou и коллеги описали новый алгоритм – изолирующий лес [87]. Изолирующий лес является частью семейства алгоритмов деревьев решений. Алгоритм заключается в построении ансамблей деревьев для заданного набора данных на основе случайно выбранных функций для разделения данных. Модель изолирует ошибочные экземпляры исходя из двух их количественных свойств: ошибочных экземпляров мало, ошибочные экземпляры сильно отличаются от нормальных. На первом этапе построения каждого дерева формируется однородная выборка из данных. Затем случайным образом выбирается точка разделения, и выборка делится. Эти шаги повторяются до тех пор, пока не достигается заданное ограничение на высоту деревьев (количество разбиений), либо пока не останется одна точка (или несколько точек с одинаковым значением). После построения ансамбля деревьев для каждой точки данных вычисляется оценка ошибки на основании этого ансамбля.

Характерной особенностью представленных выше алгоритмов является их ограниченная интерпретируемость, а увеличение размерности данных ведет к увеличению вычислительной сложности. Описанные проблемы попытались решить Z. Li, Y. Zhao, N. Botta, C. Ionescu and X. Hu, предложив в 2020 г. новый подход – алгоритм обнаружения выбросов на основе копулы (copula based outlier detector, COPOD) [88], и таким образом впервые применив теорию

копул, которая ранее использовалась в задачах финансовой аналитики и эконометрики для детекции ошибок. В основе расчета лежит непараметрический подход подбора эмпирических кумулятивных функций распределения (empirical cumulative distribution function, ECDF) – эмпирических копул. В медицинской литературе применение метода описано в качестве алгоритма сравнения для задачи выявления ухудшения состояния здоровья пациентов при дистанционном мониторинге [89].

Глубокое обучение также применимо к решению задачи на поиск выпадающих значений [53, 90, 91]. При рассмотрении методов глубокого обучения в контексте задачи обнаружения выпадающих значений предлагается следующая классификация подходов [91]:

1. Deep learning for feature extraction (глубокое обучение для извлечения признаков) – сокращает признаковое пространство с последующим решением задачи классификации на новых данных с помощью классических методов обнаружения ошибок;
2. Learning feature representation of normality (обучение особенностям представления нормальности) – не разделяет два процесса, а сочетает обучение с оценкой ошибочности экземпляра данных;
3. End-to-end anomaly score learning (сквозное обучение оценке аномальности) – напрямую изучает показатели аномальности.

На данный момент обнаружение выбросов с помощью нейронных сетей в медицине нашло применение в задаче классификации медицинских изображений, где определяются аномальные паттерны изображения, характеризующих патологию [60, 92].

д. Сравнение методов обнаружения выпадающих значений

Каждый из рассмотренных методов имеет свои особенности и ограничения. Для статистических методов, основанных на близости и плотности, определяющим фактором является размерность и количество данных, которые напрямую влияют на скорость вычислений. Общим ограничением для этих методов является также возможность использования в основном количественных или категориальных порядковых данных. Многие одномерные статистические методы имеют ограничения, связанные с характером распределения анализируемых данных [54, 64, 73].

В линейных моделях критическим является предположение о линейных корреляциях, что может быть неверным для конкретных наборов данных и может сказаться на эффективности моделирования [85]. Нейронные сети применимы при наличии большого набора обучающих данных, однако их основным недостатком в большинстве случаев является невозможность объяснить полученные результаты (неинтерпретируемость), в отличие от использования пороговых значений. С другой стороны, формирование набора пороговых значений для всех признаков ЭМК требует больших временных затрат.

Применение методов машинного обучения к ДРКП позволяет учесть демографические и антропометрические характеристики (возраст, пол, расу, этническую принадлежность, рост и вес), но точность конкретной модели машинного обучения сильно зависит от исследуемых данных. Ряд методов показывает высокую точность на одних выборках и плохо работает в применении к другим [85].

Самым важным ограничением в применимости статистических методов является невозможность объяснять полученные выводы, а также разделять задачу обнаружения выбросов и задачу обнаружения нереалистичных значений, что вызывает трудности при дальнейшей обработке.

е. Обработка выпадающих значений

В литературе описано около 20 методов работы с выбросами и нереалистичными значениями. Широко распространенной является позиция, что выбросы представляют собой наблюдения, которые должны быть удалены путем исключения всего наблюдения либо исправлены. В таком случае обработка выбросов обычно сводится к проблеме заполнения пропусков. Но указанные подходы не всегда являются корректными, поэтому в настоящее время исключать выбросы из исследования не рекомендуется [93, 94].

В случае неправдоподобных значений наиболее предпочтительный в настоящее время подход к анализу данных описан в работе [93].

Предлагается анализировать данные дважды: вместе с выпадающими значениями и без них. Если полученные результаты устойчивы к выпадающим значениям (различаются незначительно), то взять первый результат. Если различаются – привести и прокомментировать и тот, и другой. Такой подход называется анализом чувствительности.

2.5. Выявление и обработка несогласованности в данных

В вопросах выявления несогласованности данных внутри одного документа и между разными документами, из которых извлекаются ДРКП, исследователи сходятся на способе, основанном на правилах [31, 95, 96], который заключается в формировании базы знаний на основе различных подходов инженерии знаний (интервьюирование экспертов, приобретение знаний с помощью вычислительных методов) и использовании этой базы знаний при подготовке наборов данных для отбрасывания нарушающих логику значений.

ЗАКЛЮЧЕНИЕ

Таким образом, вопросы выявления и обработки ошибок в ДРКП широко рассматриваются в литературе, а предложенные методы применяются в различных задачах в области медицины. Несмотря на наличие методологических разработок в части создания и использования наборов данных, основанных на ДРКП, для исследований и разработок, включая создания СППВР и систем искусственного интеллекта [97, 98, 99], требуется дальнейшее развитие этих методик в части повышения качества создаваемых наборов, так как наличие ошибок в них может приводить к снижению качества создаваемых моделей машинного обучения для диагностики и прогнозирования.

ЛИТЕРАТУРА/REFERENCES

1. Гольдина Т.А., Колбин А.С., Белоусов Д.Ю., Боровская В.Г. Обзор исследований реальной клинической практики // Качественная Клиническая Практика. – 2021. – №1. – С.56-63. [Goldina TA, Kolbin AS, Belousov DYu, Borovskaya VG. Review of real-world data study. Kachestvennaya Klinicheskaya Praktika. 2021; 1: 56-63. (In Russ.)] doi: 10.37489/2588-0519-2021-1-56-63.
2. Солодовников А.Г., Сорокина Е.Ю., Гольдина Т.А. Данные рутинной практики (real-world data): от планирования к анализу // Медицинские технологии. Оценка и выбор. – 2020. – Т.41 – №3. – С.9-16. [Solodovnikov AG, Sorokina EYu, Goldina TA. Real-world data: from planning to analysis. Medical Technologies. Assessment and Choice. 2020; 41(3): 9-16. (In Russ.)] doi: 10.17116/medtech2020410319.
3. Maissenhaelter BE, Woolmore AL, Schlag PM. Real-world evidence research based on big data: Motivation-challenges-success factors. Onkologe (Berl). 2018; 24(S2): 91-98. doi: 10.1007/s00761-018-0358-3.
4. Гусев А.В., Зингерман Б.В., Тюфилин Д.С., Зинченко В.В. Электронные медицинские карты как источник данных реальной клинической практики // Реальная клиническая практика: данные и доказательства. – 2022. – Т.2 – №2. – С.8-20. [Gusev AV, Zingerman BV, Tyufilin DS, Zinchenko VV. Electronic medical records as a source of real-world clinical data. Real-World Data & Evidence. 2022; 2(2): 8-20. (In Russ.)] doi: 10.37489/2782-3784-myrwd-13.
5. Гольдина Т.А., Суворов Н.И. Исследования рутинной клинической практики: от получения данных к оценке медицинских технологий и принятию решений в здравоохранении // Медицинские технологии. Оценка и выбор. – 2018. – Т.31 – №1. – С.21-29. [Goldina TA, Suvorov NI. Real-World Data Studies: from Data to Health Technology Assessment and Decision-Making in Healthcare. Medical Technologies. Assessment and Choice. 2018; 1(31): 21-29. (In Russ.)] doi: 10.37489/2782-3784-myrwd-13.
6. Григорьев С.Г., Лобзин Ю.В., Скрипченко Н.В. Роль и место логистической регрессии и ROC-анализа в решении медицинских диагностических задач // Журнал инфектологии. – 2016. – Т.8 – №4. – С.36-45. [Grigoryev SG, Lobzin YuV, Skripchenko NV. The role and place of logistic regression and ROC analysis in solving medical diagnostic task. Jurnal infektologii. 2016; 4(8): 36-45. (In Russ.)] doi: 10.22625/2072-6732-2016-8-4-36-45.
7. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc. 2013; 20(1): 144-151. doi: 10.1136/amiajnl-2011-000681.
8. Cruz-Correia RJ, Rodrigues PP, Freitas A, et al. Data quality and integration issues in electronic health records. In book: Information Discovery on Electronic Health Records. Chapter: 4. Publisher: CRC PressEditors: Hristidis, Vagelis. 2009. P.55-95. doi: 10.1201/9781420090413-c4.

9. van der Lei J. Use and abuse of computer-stored medical records. Methods Inf Med. 1991; 30(2): 79-80.
10. Mikkelsen G, Aasly J. Consequences of impaired data quality on information retrieval in electronic patient records. Int J Med Inform. 2005; 74(5): 387-394. doi: 10.1016/j.ijmedinf.2004.11.001.
11. Roukema J, Los RK, Bleeker SE, et al. Paper versus computer: feasibility of an electronic medical record in general pediatrics. Pediatrics. 2006; 117(1): 15-21. doi: 10.1542/peds.2004-2741.
12. Kaboli PJ, McClimon BJ, Hoth AB, Barnett MJ. Assessing the accuracy of computerized medication histories. Am J Manag Care. 2004; 10(11 Pt 2): 872-877.
13. Wallace CJ, Stansfield D, Gibb Ellis KA, Clemmer TP. Implementation of an electronic logbook for intensive care units. Proc AMIA Symp. 2002: 840-844.
14. Botsis T, Hartvigsen G, Chen F, Weng C. Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. Summit Transl Bioinform. 2010; 2010: 1-5.
15. Wyatt JC, Liu JL. Basic concepts in medical informatics. J Epidemiol Community Health. 2002; 56(11): 808-812. doi: 10.1136/jech.56.11.808.
16. Cai L, Zhu Y. The Challenges of Data Quality and Data Quality. Assessment in the Big Data Era. Data Science Journal. 2015; 14(2): 1-10. doi: 10.5334/dsj-2015-002.
17. von Lucadou M, Ganslandt T, Prokosch HU, Toddenroth D. Feasibility analysis of conducting observational studies with the electronic health record. BMC Med Inform Decis Mak. 2019; 19(1): 202. doi: 10.1186/s12911-019-0939-0.
18. Ионов М.В., Болгова Е.В., Звартау Н.Э. и др. Внедрение системы поддержки принятия решений для повышения качества медицинских данных пациентов с артериальной гипертензией // Научно-технический вестник информационных технологий, механики и оптики. – 2022. – Т.22 – №1. – С.217-222. [Ionov MV, Bolgova EV, Zvartau NE, et al. Implementation of a clinical decision support system to improve the medical data quality for hypertensive patients. Scientific and Technical Journal of Information Technologies, Mechanics and Optics. 2022; 22(1): 217-222 (In Russ.)] doi: 10.17586/2226-1494-2022-22-1-217-222.
19. Pezoulas VC, Kourou KD, Kalatzis F, et al. Medical data quality assessment: On the development of an automated framework for medical data curation. Comput Biol Med. 2019; 107: 270-283. doi: 10.1016/j.compbiomed.2019.03.001.
20. Roche N, Reddel H, Martin R, et al. Quality standards for real-world research. Focus on observational database studies of comparative effectiveness. Ann Am Thorac Soc. 2014; 11(S2): S99-S104. doi: 10.1513/AnnalsATS.201309-300RM.
21. Collins R, MacMahon S. Reliable assessment of the effects of treatment on mortality and major morbidity, I: clinical trials. Lancet. 2001; 357(9253): 373-380. doi: 10.1016/S0140-6736(00)03651-5.
22. Takahashi Y, Nishida Y, Asai S. Utilization of health care databases for pharmacoepidemiology. Eur J Clin Pharmacol. 2012; 68(2): 123-129. doi: 10.1007/s00228-011-1088-2.
23. Han K, Song K, Choi BW. How to Develop, Validate, and Compare Clinical Prediction Models Involving Radiological Parameters: Study Design and Statistical Methods. Korean J Radiol. 2016; 17(3): 339-350. doi: 10.3348/kjr.2016.17.3.339.
24. Lee YH, Bang H, Kim DJ. How to Establish Clinical Prediction Models. Endocrinol Metab (Seoul). 2016; 31(1): 38-44. doi: 10.3803/EnM.2016.31.1.38.
25. Rahm E, Do H. Data Cleaning: Problems and Current Approaches. IEEE Data Eng. Bull. 2000; 23: 3-13.
26. Tamilselvi J, Gifta C. Handling Duplicate Data in Data Warehouse for Data Mining. International Journal of Computer Applications. 2011; 15.
27. Kristianson KJ, Ljunggren H, Gustafsson LL. Data extraction from a semi-structured electronic medical record system for outpatients: a model to facilitate the access and use of data for quality control and research. Health Informatics J. 2009; 15(4): 305-319. doi: 10.1177/1460458209345889.
28. McCoy AB, Wright A, Kahn MG, et al. Matching identifiers in electronic health records: implications for duplicate records and patient safety. BMJ Qual Saf. 2013; 22(3): 219-224. doi: 10.1136/bmjqs-2012-001419.
29. Little RJA, Rubin DB. Statistical analysis with missing data. John Wiley & Sons, Inc. Hoboken, New Jersey. 2002: 41-93. doi: 10.1002/9781119013563.fmatter.
30. Liu P, El-Darzi E, Lei L, et al. An analysis of missing data treatment methods and their application to health care dataset. In Advanced Data Mining and Applications: First International Conference, ADMA. Wuhan, China, July 22-24, 2005. Proceedings 1: 583-590.
31. Wang Z, Talburt JR, Wu N, et al. A Rule-Based Data Quality Assessment System for Electronic Health Record Data. Appl Clin Inform. 2020; 11(4): 622-634. doi: 10.1055/s-0040-1715567.
32. van der Heijden GJ, Donders AR, Stijnen T, Moons KG. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. J Clin Epidemiol. 2006; 59(10): 1102-1109. doi: 10.1016/j.jclinepi.2006.01.015.

33. Liu W, Ding J. A novel complete-case analysis to determine statistical significance between treatments in an intention-to-treat population of randomized clinical trials involving missing data. Stat Methods Med Res. 2018; 27(4): 1067-1075. doi: 10.1177/0962280216651307.

34. Okpara C, Edokwe C, Ioannidis G, et al. The reporting and handling of missing data in longitudinal studies of older adults is suboptimal: a methodological survey of geriatric journals. BMC Med Res Methodol. 2022; 22(1): 122. doi: 10.1186/s12874-022-01605-w.

35. Wells BJ, Chagin KM, Nowacki AS, Kattan MW. Strategies for handling missing data in electronic health record derived data. EGEMS (Wash DC). 2013; 1(3): 1035. doi: 10.13063/2327-9214.1035.

36. Graham JW. Missing data analysis: making it work in the real world. Annu Rev Psychol. 2009; 60: 549-576. doi: 10.1146/annurev.psych.58.110405.085530.

37. Li J, Yan XS, Chaudhary D, et al. Imputation of missing values for electronic health record laboratory data. NPJ Digit Med. 2021; 4(1): 147. doi: 10.1038/s41746-021-00518-0.

38. Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods. 2002; 7(2): 147-177.

39. Рыженкова К.В. Методы восстановления пропуска данных при проведении статистических исследований // Интеллект. Инновации. Инвестиции. – 2012. – №3. – С.127-133 [Ryzhenkova K. Data omission recovery methods in statistical research. Intellekt. Innovatsii. Investitsii. 2012; 3: 127-133. (In Russ.)]

40. Zhang Z. Missing data imputation: focusing on single imputation. Ann Transl Med. 2016; 4(1): 9. doi: 10.3978/j.issn.2305-5839.2015.12.38.

41. Nakai M, Ke W. Review of the methods for handling missing data in longitudinal data analysis. International Journal of Mathematical Analysis. 2011; 5(1): 1-13.

42. Harrell FE. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Cham: Springer international publishing, 2015. 600 p.

43. Fitzmaurice GM, Laird NM, Ware JH. Applied longitudinal analysis. Hoboken. N.J.: Wiley-Interscience, 2004. 506 p.

44. Powney M, Williamson P, Kirkham J, Kolamunnage-Dona R. A review of the handling of missing longitudinal outcome data in clinical trials. Trials. 2014; 15: 237. doi: 10.1186/1745-6215-15-237.

45. Wang H, Belitskaya-Levy I, Wu F, et al. A statistical quality assessment method for longitudinal observations in electronic health record data with an application to the VA million veteran program. BMC Med Inform Decis Mak. 2021; 21(1): 289. doi: 10.1186/s12911-021-01643-2.

46. Hegde H, Shimpi N, Panny A, et al. MICE vs PPCA: Missing data imputation in healthcare. Informatics in medicine unlocked. 2019; 17: 100275. doi: 10.1016/j.imu.2019.100275.

47. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. BMJ. 2009; 338: b2393. doi: 10.1136/bmj.b2393.

48. Brinton DL, Ford DW, Martin RH, et al. Missing data methods for intensive care unit SOFA scores in electronic health records studies: results from a Monte Carlo simulation. J Comp Eff Res. 2022; 11(1): 47-56. doi: 10.2217/cer-2021-0079.

49. Jerez JM, Molina I, García-Laencina PJ, et al. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artif Intell Med. 2010; 50(2): 105-115. doi: 10.1016/j.artmed.2010.05.002.

50. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res. 2011; 20(1): 40-49. doi: 10.1002/mpr.329.

51. Baneshi MR, Talei AR. Does the missing data imputation method affect the composition and performance of prognostic models? Iran Red Crescent Med J. 2012; 14(1): 31-36.

52. Burton A, Altman DG. Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. Br J Cancer. 2004; 91(1): 4-8. doi: 10.1038/sj.bjc.6601907.

53. Zhao F, Zhang C, Dong N, et al. A Uniform Framework for Anomaly Detection in Deep Neural Networks. Neural Process Lett. 2022; 54: 3467-3488. doi: 10.1007/s11063-022-10776-y.

54. Aggarwal CC. An introduction to outlier analysis. In: Outlier Analysis. Springer, Cham. 2017: 1-34. doi: 10.1007/978-3-319-47578-3_1.

55. Ienco D, Pensa RG, Meo R. A Semisupervised Approach to the Detection and Characterization of Outliers in Categorical Data. IEEE Trans Neural Netw Learn Syst. 2017; 28(5): 1017-1029. doi: 10.1109/TNNLS.2016.2526063.

56. Koufakou A, Ortiz E, Georgiopoulos M, et al. A Scalable and Efficient Outlier Detection Strategy for Categorical Data. 2007; 2: 210-217. doi: 10.1109/ICTAI.2007.125.

57. Suri NNRR, Murty MN, Athithan G. Detecting outliers in categorical data through rough clustering. Nat Comput. 2016; 15: 385-394. doi: 10.1007/s11047-015-9489-2.

58. Akande T, Kaur B, Dadkhah S, Ghorbani A. Threshold based Technique to Detect Anomalies using Log Files. ICMLT 2022: 2022 7th International Conference on Machine Learning Technologies. 2022: 191-198. doi: 10.1145/3529399.3529430.

59. Chen H, Zhang H, Liu C, et al. J Neural Eng. 2022; 19(5): 10.1088/1741-2552/ac954d. doi: 10.1088/1741-2552/ac954d.

60. Li X, Bagher-Ebadian H, Gardner S, et al. An uncertainty-aware deep learning architecture with outlier mitigation for prostate gland segmentation in radiotherapy treatment planning. Med Phys. 2023; 50(1): 311-322. doi: 10.1002/mp.15982.

61. Золотова Т.В., Волкова Д.А. Методы интеллектуальной обработки данных для коррекции атипичных значений котировок акций // Статистика и экономика. – 2022. – Т.19 – №2. – С.4-13. [Zolotova TV, Volkova DA. Intelligent Data Processing Methods for the Atypical Values Correction of Stock Quotes. Statistics and Economics. 2022; 2(19): 4-13. (In Russ.)] doi: 10.21686/2500-3925-2022-2-4-13.

62. Chandola V, Banerjee A, Kumar V. Anomaly detection: A Survey. ACM Comput. Surv. 2009; 41: 1-72. doi: 10.1145/1541880.1541882.

63. Grubbs FE. Sample criteria for testing outlying observations. Ann. Math. Statist. 21(1): 27-58. doi: 10.1214/aoms/1177729885.

64. ГОСТ Р ИСО 16269-4-2017 Статистические методы. Статистическое представление данных: Часть 4. Выявление и обработка выбросов (Система стандартов по информации, библиотечному и издательскому делу) [Электронный ресурс] // Электрон. фонд правовой и норматив.-техн. информ. Доступно по: https://docs.cntd.ru/document/1200146680. Ссылка действительна на 18.02.2024. [GOST R ISO 16269-4-2017 Statisticheskie metody. Statisticheskoe predstavlenie dannykh: Chast' 4. Vyyavlenie i obrabotka vybrosov (Sistema standartov po informatsii, bibliotechnomu i izdatel'skomu delu) [Elektronnyi resurs]. Elektron. fond pravovoi i normativ.-tekhn. inform. [cited 18.02.2024] Available from: https://docs.cntd.ru/document/1200146680. (In Russ.)]

65. Tukey JW. Exploratory data analysis. Addison-Wesley publishing company. 1977. 716 p.

66. Hampel FR. The influence curve and its role in robust estimation // Journal of the American statistical association. 1974; 69(346): 383-393. doi: 10.2307/2285666.

67. Rousseeuw P, Hubert M. Robust statistics for outlier detection. Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery. 2011; 1: 73-79. 10.1002/widm.2.

68. Кузнецов В.В., Ларин С.Л., Романенко С.В. Алгоритм обнаружения серии выбросов по критерию Диксона в инверсионной вольтамперометрии // Аналитика и контроль. – 2014. – Т.18 – №3. – С.310-315. [Kuznetsov VV, Romanenko SV, Larin SL. Detection algorithm of a series of releases by Dikson criterion in inversion voltammetry. Analitika i kontrol`. 2014; 3(18): 310-315. (In Russ.)]

69. Bansal V, Dorn C, Grunert M, et al. Outlier-based identification of copy number variations using targeted resequencing in a small cohort of patients with Tetralogy of Fallot. PLoS One. 2014; 9(1): e85375. doi: 10.1371/journal.pone.0085375.

70. Barkley D, Hatsis P, Glick J, et al. Dixon's Q-test and Student's t-test to assess analog internal standard response in nonregulated LC-MS/MS bioanalysis. Bioanalysis. 2020; 12(21): 1535-1543. doi: 10.4155/bio-2020-0207.

71. Marcks KL, Zhao Y, Motro M, Will LA. Cephalometric Variability Among Siblings: A Pilot Study. Turk J Orthod. 2022; 35(4): 239-247. doi: 10.5152/TurkJOrthod.2022.21237.

72. Estiri H, Murphy SN. Semi-supervised encoding for outlier detection in clinical observation data. Comput Methods Programs Biomed. 2019; 181: 104830. doi: 10.1016/j.cmpb.2019.01.002.

73. Estiri H, Klann JG, Murphy SN. A clustering approach for detecting implausible observation values in electronic health records data. BMC Med Inform Decis Mak. 2019; 19(1): 142. doi: 10.1186/s12911-019-0852-6.

74. Phan HTT, Borca F, Cable D, et al. Automated data cleaning of paediatric anthropometric data from longitudinal electronic health records: protocol and application to a large patient cohort. Sci Rep. 2020; 10(1): 10164. doi: 10.1038/s41598-020-66925-7.

75. Knorr EM, Ng R. Algorithms for mining distance-based outliers in large datasets. VLDB '98: Proceedings of the 24rd International conference on very large data bases. 1998; 392-403.

76. Ramaswamy S, Rastogi R, Shim K. Efficient algorithms for mining outliers from large data sets. ACM SIGMOD Record: proceedings of the 2000 ACM SIGMOD international conference on management of data. Dallas Texas: ACM. 2000; 29(2): 427-438. doi: 10.1145/335191.335437.

77. Breunig M, Kröger P, Ng R, Sander J. LOF: identifying density-based local outliers // ACM SIGMOD Record: proceedings of the 2000 ACM SIGMOD international conference on management of data. Dallas Texas: ACM. 2000; 29(2): 93-104. doi: 10.1145/335191.335388.

78. Kumaravel VP, Buiatti M, Parise E, Farella E. Adaptable and Robust EEG Bad Channel Detection Using Local Outlier Factor (LOF). Sensors (Basel). 2022; 22(19): 7314. doi: 10.3390/s22197314.

79. Karasmanoglou A, Antonakakis M, Zervakis M. ECG-Based Semi-Supervised Anomaly Detection for Early Detection and Monitoring of Epileptic Seizures. Int J Environ Res Public Health. 2023; 20(6): 5000. doi: 10.3390/ijerph20065000.

80. Fowler JW, Alpert BK, Joe YI, et al. A Robust Principal Component Analysis for Outlier Identification in Messy Microcalorimeter Data. J Low Temp Phys. 2019; 199(3-4): 10.1007/s10909-019-02248-w. doi: 10.1007/s10909-019-02248-w.

81. Ebrahimi S, Fleuret J, Klein M, et al. Robust Principal Component Thermography for Defect Detection in Composites. Sensors (Basel). 2021; 21(8): 2682. doi: 10.3390/s21082682.

82. Chen X, Zhang B, Wang T, et al. Robust principal component analysis for accurate outlier sample detection in RNA-Seq data. BMC Bioinformatics. 2020; 21(1): 269. doi: 10.1186/s12859-020-03608-0.

83. Hubert M, Rousseeuw P, Branden K. ROBPCA: A new approach to robust principal component analysis. Technometrics. 2005; 47: 64-79. doi: 10.1198/004017004000000563.

84. Hubert M, Rousseeuw P, Verdonck T. Robust PCA for skewed data and its outlier map. Computational Statistics & Data Analysis. 2009; 53: 2264-2274. doi: 10.1016/j.csda.2008.05.027.

85. Aggarwal CC. Linear models for outlier detection. In: Outlier analysis. Springer, Cham: Springer international publishing. 2017: 65-110. doi: 10.1007/978-3-319-47578-3_3.

86. Wold S, Esbensen K, Geladi P. Principal component analysis. Chemometrics and intelligent laboratory systems. 1987; 2(1-3): 37-52. doi: 10.1016/0169-7439(87)80084-9.

87. Liu FT, Ting KM, Zhou Z. Isolation forest. 2008 Eighth IEEE international conference on data mining. 2009: 413-422. doi: 10.1109/ICDM.2008.17.

88. Li Z, Zhao Y, Botta N, et al. COPOD: Copula-Based Outlier Detection. International conference on data mining: 2020 IEEE International conference on data mining (ICDM), 2020: 1118-1123. doi: 10.1109/ICDM50108.2020.00135.

89. Bijlani N, Nilforooshan R, Kouchaki S. An Unsupervised Data-Driven Anomaly Detection Approach for Adverse Health Conditions in People Living With Dementia: Cohort Study. JMIR Aging. 2022; 5(3): e38211. doi: 10.2196/38211.

90. Pang G, Shen C, Hengel A. Deep anomaly detection with deviation networks // KDD '19: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019: 353-362. doi: 10.1145/3292500.3330871.

91. Pang G, Shen C, Cao L, Hengel A. Deep learning for anomaly detection: a review. ACM Computing surveys. 2021; 54(2): 1-38. doi: 10.1145/3439950.

92. Garcia JB, Tanadini-Lang S, Andratschke N, et al. Suspicious Skin Lesion Detection in Wide-Field Body Images using Deep Learning Outlier Detection. Annu Int Conf IEEE Eng Med Biol Soc. 2022; 2022: 2928-2932. doi: 10.1109/EMBC48229.2022.9871655.

93. Реброва О.Ю. Статистический анализ медицинских данных. Применение пакета прикладных программ STATISTICA. – М.: МедиаСфера; 2006. [Rebrova OYu. Statisticheskii analiz meditsinskikh dannykh. Primenenie paketa prikladnykh programm STATISTICA. Moscow: Media Sfera; 2006. (In Russ.)]

94. Aguinis H, Gottfredson RK, Joo H. Best-practice recommendations for defining, identifying, and handling outliers. Organizational Research Methods. 2013; 16(2): 270-301. doi: 10.1177/1094428112470848.

95. Brown PJ, Warmington V. Data quality probes-exploiting and improving the quality of electronic patient record data and patient care. Int J Med Inform. 2002; 68(1-3): 91-98. doi: 10.1016/s1386-5056(02)00068-0.

96. Carlson D, Wallace CJ, East TD, Morris AH. Verification & validation algorithms for data used in critical care decision support systems. Proc Annu Symp Comput Appl Med Care. 1995; 188-192.

97. Бобровская Т.М., Васильев Ю.А., Никитин Н.Ю., Арзамасов К.М. Подходы к формированию наборов данных в лучевой диагностике // Врач и информационные технологии. – 2023. – №4. – С.14-23. [Bobrovskaya TM, Vasil'ev YUA, Nikitin NYU, Arzamasov KM. Podhody k formirovaniyu naborov dannyh v luchevoj diagnostike. Vrach i informacionnye tekhnologii. 2023; 4: 14-23. (In Russ.)]

98. Васильев Ю.А., Бобровская Т.М., Арзамасов К.М. и др. Основополагающие принципы стандартизации и систематизации информации о наборах данных для машинного обучения в медицинской диагностике // Менеджер здравоохранения. – 2023. – №4. – С.28-41. [Vasil'ev YUA, Bobrovskaya TM, Arzamasov KM, et al. Osnovopolagayushchie principy standartizacii i sistematizacii informacii o naborah dannyh dlya mashinnogo obucheniya v medicinskoj diagnostike. Menedzher zdravoohraneniya. 2023; 4: 28-41. (In Russ.)]

99. Национальный стандарт РФ ГОСТ Р 59921.5-2022 «Системы искусственного интеллекта в клинической медицине. Часть 5. Требования к структуре и порядку применения набора данных для обучения и тестирования алгоритмов». [Nacional'nyj standart RF GOST R 59921.5-2022 «Sistemy iskusstvennogo intellekta v klinicheskoj medicine. CHast' 5. Trebovaniya k strukture i poryadku primeneniya nabora dannyh dlya obucheniya i testirovaniya algoritmov». (In Russ.)]