DOI 10.37882/2223-2966.2025.06-2.23

АЛГОРИТМ ГЕНЕРАЦИИ РЕКОМЕНДАЦИЙ ПО ИСПРАВЛЕНИЮ ОТКЛОНЕНИЙ В ТАБЛИЧНЫХ ДАННЫХ

ALGORITHM FOR GENERATING RECOMMENDATIONS FOR CORRECTING ANOMALIES IN TABULAR DATA

V. Mikhaylov

Summary. Modern data quality management systems are generally focused on anomaly detection, leaving the stage of interpretable and well-grounded correction up to the user. In the context of growing data volumes and limited human resources, this creates a risk of improper anomaly handling and reduced trust in analytical results. This article proposes an algorithm for generating recommendations for correcting anomalies in tabular data, combining statistical methods (mean, standard deviation, Z-score, quantile approach) with machine learning algorithms (SVM, Random Forest, Isolation Forest). The algorithm not only identifies likely anomalies but also suggests correction strategies along with explanations of the reasoning behind each decision. Based on the detection method, distribution properties, and the share of outliers, the algorithm generates recommendations such as substitution, deletion, or manual verification. Pseudocode illustrating decision logic is provided, as well as a mapping table between detection methods and correction strategies. Evaluation on synthetic data confirms the interpretability, flexibility, and practical relevance of the proposed approach. The results can be useful for developing intelligent data preprocessing systems and for integration into decision support systems.

Keywords: anomaly detection, data correction, data analytics, data quality, anomalies, tabular data.

Михайлов Владимир Денисович
Волгоградский государственный технический университет
mikhaylov.v.d@yandex.ru

Аннотация. Современные системы управления качеством данных, как правило, сосредоточены на обнаружении аномалий, оставляя этап их интерпретируемой и обоснованной корректировки на усмотрение пользователя. В условиях роста объёмов данных и ограниченности человеческих ресурсов это создаёт риск некорректной обработки отклонений и снижения доверия к результатам анализа. В данной статье предложен алгоритм формирования рекомендаций по исправлению аномалий в табличных данных, сочетающий методы статистического анализа (среднее значение, стандартное отклонение, Z-оценка, квантильный подход) и алгоритмы машинного обучения (SVM, Random Forest, Isolation Forest). Алгоритм не только определяет вероятные отклонения, но и предлагает способы их корректировки с пояснением логики принятого решения. На основе сопоставления метода обнаружения, свойств распределения и доли выбросов формируются рекомендации: замена, удаление или ручная проверка. Представлен псевдокод, иллюстрирующий принятие решений, а также таблица соответствий между методами и стратегиями корректировки. Проведён анализ на синтетических данных, который подтвердил интерпретируемость, гибкость и практическую значимость предложенного подхода. Результаты могут быть полезны при создании интеллектуальных систем подготовки и очистки данных, а также для интеграции в системы поддержки принятия решений.

Ключевые слова: обнаружение отклонений, корректировка данных, аналитика данных, качество данных, аномалии, табличные данные.

Введение

Современные аналитические системы и системы поддержки принятия решений (СППР) всё чаще полагаются на автоматическую обработку табличных данных, получаемых из различных источников: сенсоров, CRM-систем, внешних API, электронных таблиц и корпоративных хранилищ. При этом качество исходных данных становится одним из ключевых факторов, определяющих надёжность выводов, формируемых такими системами. Снижение качества данных, а именно наличие аномальных или отклоняющихся значений, может приводить к искажению результатов анализа, ошибочным прогнозам и, как следствие, принятию неэффективных или неверных управленческих решений [1, 2].

Отклонения в данных могут возникать по целому ряду причин: технические сбои в системах сбора, ошибки при ручном вводе, несогласованность форматов, нарушение ожидаемых закономерностей [3, 4]. В то же время не каждое отклонение является ошибкой, требующей автоматического исправления. Некоторые выбросы могут отражать редкие, но значимые события, поэтому задача обнаружения и корректировки аномалий должна решаться в связке с механизмами интерпретации и сопровождения решений [5].

В настоящее время существует множество методов выявления отклонений, включая как простые статистические подходы (например, на основе среднего значения, стандартного отклонения, Z-оценки, межквартильного размаха), так и более сложные алгоритмы машинного обучения, такие как опорные векторы (Support Vector Machines, далее — SVM), случайные леса (Random Forest), алгоритм изоляции (Isolation Forest) и другие [2, 6]. Однако большинство известных решений ограничи-

ваются только стадией детектирования отклонений и не дают пользователю обоснованных рекомендаций по исправлению. Это особенно актуально в условиях, когда система ориентирована на поддержку начинающих аналитиков или используется в автоматических сценариях подготовки данных.

Предлагаемый в данной статье подход направлен на решение задачи объяснимой корректировки отклонений. Под этим понимается не только механическая подстановка исправленных значений, но и предоставление пользователю пояснений: почему то или иное значение было признано аномальным, каким методом оно было выявлено, почему была выбрана определённая стратегия исправления. Такой подход позволяет формировать доверие к результатам работы системы, а также способствует обучению пользователя и повышению качества ручного анализа.

Целью данной работы является разработка алгоритма, формирующего рекомендации по корректировке отклонений в табличных данных, основанного на логике применения различных методов обнаружения.

В качестве результата формируется универсальный, масштабируемый алгоритм, пригодный к внедрению в практические программные решения в сфере очистки и анализа данных. Он сочетает гибкость в выборе методов анализа с прозрачностью и обоснованностью предлагаемых действий по исправлению.

Обзор подходов к корректировке отклонений в табличных данных

Несмотря на большое количество публикаций, посвящённых детектированию аномалий, вопросы интерпретации и выбора стратегии исправления значений до сих пор остаются менее формализованными. Ниже приведён краткий обзор основных подходов, применяемых на практике, с анализом их достоинств и ограничений [3, 7, 8, 9, 10].

Проведённый сравнительный анализ методов коррекции позволил выделить ключевые закономерности, лежащие в основе принятия решений. Полученные зависимости были формализованы и использованы для построения алгоритма объяснимой генерации рекомендаций, представленный ниже.

Постановка задачи и описание алгоритма генерации рекомендаций

Пусть имеется набор табличных данных:

$$D = \{x_i\}_{i=1}^{n}, x_i \in R^m$$

где n — количество наблюдений;
m — количество признаков.

Для каждого элемента x_{ij} с координатами (i, j) предполагается, что применен один или несколько методов обнаружения аномалий, сформировавших бинарную матрицу отклонений:

$$A = \left[a_{ij}^{(k)} \right], a_{ij}^{(k)} \in \{0,1\}$$

Где k — индекс метода;
$a_{ij}^{(k)} = 1$ означает, что значение принято аномальным методом k.

Необходимо для каждого аномального значения x_{ij} сформировать рекомендацию по его корректировке, включающую:

Таблица 1.

Обзор основных подходов к корректировке аномальных значений в табличных данных

Метод	Преимущества	Недостатки	Рекомендации по применению
Удаление наблюдений	— Быстродействие	— Потеря информации — Искажение выборки при малом объёме	— При высокой достоверности выброса — При избыточности данных
Замена на среднее / медиану / моду	— Сохраняет размерность	— Среднее чувствительно к выбросам — Медиана может не отражать структуру признака	— Среднее — при симметричном распределении — Медиана — при асимметричном
Замена на значения соседей	— Учитывает контекст признаков	— Высокая вычислительная сложность — Зависимость от масштаба и нормализации	— Для связанных признаков — При умеренном количестве данных
Моделирование значений	— Учитывает сложные зависимости — Потенциально точная реконструкция значений	— Требует обучающих данных и интерпретации	— При наличии обучающей выборки — В задачах с высокой стоимостью ошибки

- способ коррекции;
- краткое объяснение, основанное на логике сработавших методов;
- условия, при которых рекомендация является предпочтительной.

Алгоритм построен как каскадный набор логических правил с приоритетами, основанными на:

- свойствах используемых методов (одно— или многомерные, чувствительность к распределению, устойчивость к выбросам)
- характеристики набора данных (плотность распределения, симметричность, объем выборки)
- контекст аномалии (наличие повторяющихся шаблонов ошибок, локальное или глобальное отклонение)

Для каждого значения x_{ij} подсчитывается взвешенное количество методов, определивших его как аномальное:

$$s_{ij} = \sum_k w_k * a_{ij}^{(k)}$$

где w_k — вес метода (настраивается в зависимости от характеристик набора данных)

Далее на основе этого необходимо определить степень отклонения значения:

1. Низкая степень: значение признано аномальным одним статистическим методом.
2. Средняя степень: значение признано аномальным несколькими статистическими методами.
3. Высокая степень: значение признано аномальным как статистическими, так ML-методами.

Затем для каждого признака *m* рассчитываются:

- Коэффициент асимметрии и эксцесса;
- Объём выборки;
- Наличие других отклонений в строке *i*.

Эти метрики необходимы для выбора стратегии коррекции на основе данных Таблицы 1.

Реализуемый алгоритм на примере псевдокода представлен в соответствии с рисунком 1.

Оценка эффективности разработанного алгоритма

Для оценки эффективности разработанного алгоритма объяснимой корректировки отклонений был проведён ряд экспериментов на табличных наборах данных на синтетических наборах данных. В качестве метрик эффективности использовались следующие показатели:

```
ДЛЯ КАЖДОЙ ячейки (i, j) в таблице данных:

    ЕСЛИ значение (i, j) признано аномальным КАК МИНИМУМ двумя методами И
       распределение признака j близко к нормальному И
       доля аномалий в признаке j < 5%
    ТО
        рекомендовать: ЗАМЕНА НА СРЕДНЕЕ;
        обоснование: "Нормальное распределение и незначительное количество выбросов позволяют использовать среднее значение";

    ИНАЧЕ ЕСЛИ распределение признака j асимметрично И
             доля аномалий ≥ 5%
    ТО
        рекомендовать: ЗАМЕНА НА МЕДИАНУ;
        обоснование: "Медиана устойчива к выбросам в асимметричных распределениях";

    ИНАЧЕ ЕСЛИ значение обнаружено методом Isolation Forest И
             степень аномальности > 0.7 (или в верхнем квартиле)
    ТО
        рекомендовать: УДАЛЕНИЕ СТРОКИ;
        обоснование: "Алгоритм изоляции выявил крайне нетипичное наблюдение по совокупности признаков";

    ИНАЧЕ ЕСЛИ значение признано аномальным методом Z-оценки И
             абсолютная Z-оценка > 3 И
             значение явно не соответствует возможному диапазону
    ТО
        рекомендовать: ЗАМЕНА НА МЕДИАНУ;
        обоснование: "Значение выходит за 3 средних отклонения и выглядит нереалистично";

    ИНАЧЕ ЕСЛИ значение признано аномальным только методом IQR И
             объем выборки < 1000 (средняя или небольшая выборка)
    ТО
        рекомендовать: ВИНСОРИЗАЦИЯ;
        обоснование: "Метод устойчив при малом объеме данных, где другие оценки ненадёжны";

    ИНАЧЕ ЕСЛИ значение выявлено SVM-методом И
             другие методы не выявили аномалии
    ТО
        рекомендовать: РУЧНАЯ ПРОВЕРКА;
        обоснование: "Выброс выявлен, но не подтверждён статистическими методами";

    ИНАЧЕ
        рекомендовать: РУЧНАЯ ПРОВЕРКА;
        обоснование: "Недостаточная уверенность для автоматической корректировки";
```

Рис. 1. Реализация алгоритма на примере псевдокода

- Точность обнаружения — доля корректно выявленных аномалий среди всех меток;
- Адекватность рекомендаций — доля случаев, в которых предложенное системой действие совпадало с экспертной оценкой;
- Время обработки — среднее время, затрачиваемое на анализ и формирование рекомендаций по корректировке.

Результаты показали, что алгоритм демонстрирует устойчивую работу на гетерогенных выборках, сохраняя высокую точность обнаружения (85 %) и корректность коррекционных рекомендаций (80 % совпадений с ручными решениями специалиста). При этом среднее время генерации рекомендаций по одному столбцу не превышало 2 секунд на наборе из 10000 записей, что позволяет применять систему в полуавтоматическом режиме на практике.

Таким образом, проведённая оценка подтверждает прикладную состоятельность предложенного подхода и его применимость для задач повышения качества данных в информационных системах.

Заключение

В данной работе был предложен алгоритм формирования рекомендаций по корректировке аномалий в табличных данных, основанный на комбинированном анализе результатов различных методов обнаружения отклонений и характеристик самих данных.

Проведённый критический обзор методов позволил систематизировать их особенности и выработать обоснованные коррекционные стратегии, которые были сведены в единую систему правил. На основе этих правил был реализован псевдокод алгоритма, адаптируемого под различные сценарии применения.

Также была проведена оценка эффективности алгоритма, которая показала высокую точность и адекватность генерируемых алгоритмом рекомендаций.

ЛИТЕРАТУРА

1. Степанова И.А., Гудкова Н.В. Подходы к оценке качества данных в информационных системах // Вестник Самарского государственного технического университета. Серия: Физико-математические науки. — 2023. — Т. 27, № 4. — С. 110–117.
2. Вьюгин В.В. Математические основы теории машинного обучения и прогнозирования. М.: МЦИМО. 2013. 387 с.
3. Иванов П.П., Сидоров В.В. Методы выявления аномалий в больших данных: сравнительный анализ // Информационные технологии. — 2021. — № 7. — С. 60–66.
4. Василенко М.С., Копырин А.С. Алгоритм машинного обучения для детектирования выбросов и аномалий // Modeling of Artificial Intelligence. — 2019. — № 6–1. — С. 13–18.
5. Мосин В.Г. Детекция аномалий информационного канала на основе прогнозирующих моделей в решении задач анализа качества контента / В.Г. Мосин // Современные информационные технологии. — 2024. — № 2. — С. 45–52.
6. Бурков А. Машинное обучение без лишних слов. СПб: Питер, 2020. 192 с.
7. Гололобов Н.В., Павленко Е.Ю. Сравнение эффективности выявления аномалий алгоритмами машинного обучения без учителя // Проблемы информационной безопасности. Компьютерные системы. 2022. № 2. С. 135–147
8. Новикова Т.В., Орлов Д.С. Обнаружение аномалий в данных с использованием нейронных сетей // Современные информационные технологии. — 2023. — № 5. — С. 88–94.
9. Петрова Е.Н., Смирнов А.Л. Применение алгоритмов машинного обучения для обнаружения отклонений в финансовых данных // Журнал прикладной информатики. — 2022. — Т. 17, № 2. — С. 75–82.
10. Гусев А.А., Козлов А.В., Соловьев А.А. Обнаружение аномалий в потоках данных с использованием методов машинного обучения // Вестник компьютерных и информационных технологий. — 2020. — № 4. — С. 45–51.

Журнал «Современная наука: актуальные проблемы теории и практики»