УДК 004.852

ВЫЯВЛЕНИЕ АНОМАЛИЙ В СЛОЖНЫХ ДАННЫХ С ПОМОЩЬЮ КЛАСТЕРИЗАЦИИ

Шевцова Т.А.
бакалавр, Уральский федеральный университет имени первого Президента России Б.Н. Ельцина (Екатеринбург, Россия)

ANOMALY DETECTION IN COMPLEX DATA USING CLUSTERING

Shevtsova T.
bachelor's degree, Ural Federal University named after the First President of Russia B.N. Yeltsin (Ekaterinburg, Russia)

Аннотация

В статье рассматриваются современные методы кластеризации, используемые для выявления аномалий в сложных данных. Основное внимание уделяется таким алгоритмам, как K-средних, DBSCAN и иерархическая кластеризация, которые эффективно применяются для обнаружения аномальных данных в различных областях, включая кибербезопасность, финансовую аналитику и здравоохранение. Приведены примеры использования этих методов для анализа данных, а также описаны ключевые особенности и различия между ними. Особое внимание уделено применению алгоритма K-средних для выделения кластеров на основе схожих характеристик данных, а также алгоритму DBSCAN, который позволяет выявлять выбросы и аномалии без предварительного указания количества кластеров. Иерархическая кластеризация рассматривается как подход для более глубокого анализа данных, когда необходимо выявить сложные связи между объектами. Статья также включает примеры кода на языке Java, которые демонстрируют реализацию различных методов кластеризации. В результате анализа показано, что выбор метода зависит от особенностей данных и целей исследования, а также от таких факторов, как размер данных, их плотность и наличие выбросов. Рекомендации по выбору метода кластеризации помогут улучшить точность и эффективность обнаружения аномалий в реальных задачах.

Ключевые слова: кластеризация, аномалии, DBSCAN, K-средних, выявление выбросов, алгоритмы анализа данных.

Abstract

This article examines modern clustering methods used for anomaly detection in complex data. The focus is on algorithms such as K-means, DBSCAN, and hierarchical clustering, which are effectively applied to identify anomalous data in various fields, including cybersecurity, financial analytics, and healthcare. Examples of applying these methods to data analysis are provided, along with a description of their key features and differences. Particular attention is given to the K-means algorithm, which clusters data based on similar characteristics, and the DBSCAN algorithm, which detects outliers and anomalies without the need to predefine the number of clusters. Hierarchical clustering is explored as an approach for more in-depth data analysis, especially when uncovering complex relationships between objects. The article also includes Java code examples demonstrating the implementation of various clustering methods. The analysis shows that the choice of method depends on the characteristics of the data and the goals of the study, as well as factors such as data size, density, and the presence of outliers. Recommendations for selecting a clustering method aim to enhance the accuracy and efficiency of anomaly detection in real-world applications.

Keywords: clustering, anomalies, DBSCAN, K-means, outlier detection, data analysis algorithms.

Введение

В условиях бурного роста объемов данных и их разнообразия, эффективные методы их обработки и анализа становятся критически важными для многих областей науки и бизнеса. Одной из таких задач является выявление аномалий в сложных данных, что может касаться выявления мошенничества, ошибок в данных, технических сбоев или иных необычных событий. Аномалии могут быть как явными, так и скрытыми, что делает их выявление трудной задачей. Традиционные методы, такие как статистический анализ или простая кластеризация, зачастую не могут обеспечить необходимую точность и эффективность при обработке больших объемов данных, что создает потребность в новых методах, таких как алгоритмы кластеризации.

Алгоритмы кластеризации, применяемые для обнаружения аномалий, позволяют не только выделять необычные данные, но и эффективно группировать объекты с похожими характеристиками. В последние годы кластеризация становится одним из наиболее востребованных методов для анализа сложных многомерных наборов данных. Она широко применяется в таких областях, как финансовая аналитика, здравоохранение, кибербезопасность и интернет вещей, где необходимость быстрого и точного выявления аномальных событий имеет большое значение. Важность этого подхода заключается в том, что он позволяет не только обнаружить аномалии, но и классифицировать их, предоставляя более полное понимание происходящих процессов.

Цель данной статьи заключается в анализе применения методов кластеризации для выявления аномалий в сложных данных. Рассматриваются основные подходы и алгоритмы, такие как K-средних, DBSCAN, иерархическая кластеризация, их преимущества и ограничения в контексте решения задач обнаружения аномалий. Также уделяется внимание практическим примерам применения этих методов в реальных задачах, включая анализ финансовых транзакций и медицинских данных.

Основная часть

Кластеризация является одним из ключевых методов для выявления аномалий в сложных данных [1]. Одним из наиболее популярных методов кластеризации является алгоритм K-средних. Этот метод группирует данные в K кластеров, минимизируя внутрикластерное расстояние. Для обнаружения аномалий можно выделить те данные, которые находятся далеко от центров кластеров или не могут быть отнесены к ни одному кластеру [2].

Пример кода для алгоритма K-средних на языке Java:

```
import org.apache.commons.math3.ml.clustering.KMeansPlusPlusClusterer;
import org.apache.commons.math3.ml.clustering.Cluster;
import org.apache.commons.math3.ml.clustering.DoublePoint;

import java.util.ArrayList;
import java.util.List;

public class KMeansExample {

    public static void main(String[] args) {
        // Данные, которые будут кластеризованы (например, точки с координатами)
        List<DoublePoint> points = new ArrayList<>();
        points.add(new DoublePoint(new double[]{1.0, 2.0}));
        points.add(new DoublePoint(new double[]{2.0, 3.0}));
        points.add(new DoublePoint(new double[]{10.0, 10.0}));
        points.add(new DoublePoint(new double[]{12.0, 12.0}));
        points.add(new DoublePoint(new double[]{50.0, 50.0}));
```

```
        // Количество кластеров
        int k = 2;

        // Кластеризация с использованием KMeans++
        KMeansPlusPlusClusterer<DoublePoint> clusterer = new KMeansPlusPlusClusterer<>(k);
        List<Cluster<DoublePoint>> clusters = clusterer.cluster(points);

        // Вывод кластеров
        for (Cluster<DoublePoint> cluster : clusters) {
            System.out.println("Cluster center: " + cluster.getCenter());
            for (DoublePoint point : cluster.getPoints()) {
                System.out.println("Point: " + point);
            }
        }
    }
}
```

В данном примере используется библиотека Apache Commons Math для реализации алгоритма K-средних с использованием стратегии KMeans++. Важно отметить, что данный код выполняет кластеризацию нескольких точек в двухмерном пространстве, а затем выводит результаты кластеризации. Кластеры можно использовать для выявления аномальных точек, которые значительно удалены от центров кластеров.

Аналогичный код может быть использован для обработки более сложных данных, таких как финансовые транзакции или медицинские данные, где каждый «точка» может представлять собой многомерный набор характеристик [3]. Например, при анализе транзакций точкой может быть транзакция, а координатами – сумма, частота и другие параметры.

Следующим важным методом для выявления аномалий является алгоритм DBSCAN (Density-Based Spatial Clustering of Applications with Noise), который отличается от K-средних тем, что не требует заранее заданного количества кластеров и позволяет выявлять аномалии, которые могут быть расценены как выбросы (noise) [4].

Пример кода для DBSCAN на языке Java:

```
import org.apache.commons.math3.ml.clustering.DBSCANClusterer;
import org.apache.commons.math3.ml.clustering.DoublePoint;

import java.util.ArrayList;
import java.util.List;

public class DBSCANExample {

    public static void main(String[] args) {
        // Данные для кластеризации
        List<DoublePoint> points = new ArrayList<>();
        points.add(new DoublePoint(new double[]{1.0, 2.0}));
        points.add(new DoublePoint(new double[]{2.0, 3.0}));
        points.add(new DoublePoint(new double[]{1.1, 2.1}));
        points.add(new DoublePoint(new double[]{10.0, 10.0}));
        points.add(new DoublePoint(new double[]{50.0, 50.0}));

        // Параметры DBSCAN: минимальное количество точек в кластере и радиус
        int minPts = 2;
        double eps = 3.0;

        // Кластеризация с использованием DBSCAN
        DBSCANClusterer<DoublePoint> clusterer = new DBSCANClusterer<>(eps, minPts);
        List<List<DoublePoint>> clusters = clusterer.cluster(points);
```

```
        // Вывод кластеров
        int clusterId = 1;
        for (List<DoublePoint> cluster : clusters) {
            System.out.println("Cluster " + clusterId + ":");
            for (DoublePoint point : cluster) {
                System.out.println("Point: " + point);
            }
            clusterId++;
        }
    }
}
```

В этом примере используется также библиотека Apache Commons Math, но в отличие от метода K-средних, DBSCAN не требует указания количества кластеров. Вместо этого он использует два параметра: минимальное количество точек в кластере (minPts) и радиус поиска (eps), чтобы определить, какие точки считаются частью одного кластера, а какие являются выбросами (шумом) [5].

При анализе сложных данных, таких как интернет-трафик или медицинские данные, DBSCAN может быть особенно полезен, так как он способен выявить аномалии в данных с высокой плотностью и выделить их как отдельные объекты или выбросы.

В таблице 1 представлены основные методы кластеризации, которые применяются для выявления аномалий. Каждый метод имеет свои преимущества и ограничения, которые необходимо учитывать при выборе алгоритма для конкретной задачи [6]. Например, метод K-средних эффективен для простых задач, но его точность может страдать при наличии выбросов. В то время как DBSCAN, благодаря своей способности выявлять выбросы, является более подходящим для работы с шумными данными.

Таблица 1

Применение методов кластеризации для выявления аномалий

Метод кластеризации	**Преимущества**	**Ограничения**	**Пример применения**
K-средних	Простота реализации, быстрое выполнение	Требуется заранее заданное количество кластеров, чувствителен к выбросам	Финансовые транзакции
DBSCAN	Не требует заранее заданного числа кластеров, хорошо работает с выбросами	Требует настройки параметров eps и minPts, чувствителен к масштабу данных	Кибербезопасность, интернет-трафик
Иерархическая кластеризация	Гибкость в анализе данных, возможность создавать иерархические структуры	Высокая вычислительная сложность, не подходит для больших данных	Медицинская диагностика

Метод K-средних, несмотря на свою простоту и высокую скорость работы, может сталкиваться с проблемами, если данные содержат выбросы или имеют сильно различающиеся плотности. В таких случаях алгоритм может неправильно классифицировать

данные, что приводит к снижению точности. Тем не менее, его широко применяют в задачах, где данные хорошо структурированы и не содержат значительных выбросов, например, в финансовых транзакциях, где необходимо разделить данные на группы с максимальной внутренней однородностью [7].

DBSCAN, в свою очередь, является более гибким и устойчивым к выбросам методом, так как не требует заранее заданного количества кластеров. Этот алгоритм работает на основе плотности, что позволяет эффективно выявлять аномальные точки, которые не соответствуют плотным группам данных. Однако его применение требует тщательной настройки параметров, таких как радиус поиска (eps) и минимальное количество точек для формирования кластера (minPts), что может быть вызовом при работе с большими и многомерными наборами данных. Тем не менее, DBSCAN показывает свою эффективность в таких областях, как кибербезопасность и анализ интернет-трафика, где важно выделить аномальные события или подозрительные действия [8].

Иерархическая кластеризация предоставляет еще один интересный подход, особенно когда необходимо получить иерархическую структуру кластеров для дальнейшего анализа. Этот метод позволяет создать древовидную структуру, в которой кластеры могут быть подразделены на более мелкие группы, что дает более детализированную картину данных. Однако его вычислительная сложность значительно выше, особенно при работе с большими объемами данных, что ограничивает его применение в реальных задачах с высокой нагрузкой. Несмотря на это, иерархическая кластеризация может быть полезна в задачах медицинской диагностики, где требуется не только выявить аномалии, но и понять, как они связаны между собой на разных уровнях данных [9].

В итоге, выбор метода кластеризации для выявления аномалий зависит от особенностей анализируемых данных, требуемой точности и скорости работы алгоритма. Для каждого конкретного применения важно учитывать как преимущества, так и ограничения каждого из методов. Например, в задачах с большими объемами данных и высокой плотностью полезнее будет использовать DBSCAN, в то время как для упрощенных случаев с хорошо структурированными данными K-средних будет наилучшим выбором.

Применение методов кластеризации для выявления аномалий в различных областях

В последние годы применение методов кластеризации для выявления аномалий активно расширяется в различных областях, таких как кибербезопасность, финансовые технологии, здравоохранение и промышленность. Аномальные данные, такие как неожиданные отклонения или нехарактерные события, могут существенно повлиять на принятие решений и поведение системы, поэтому их своевременное выявление играет критически важную роль в предотвращении негативных последствий [10]. В различных областях используются разные методы кластеризации в зависимости от сложности данных и специфики задач.

В области **кибербезопасности** методы кластеризации применяются для выявления аномальных паттернов в сетевых данных, что позволяет обнаружить потенциальные угрозы и вторжения в реальном времени. Например, анализ трафика сети с использованием алгоритмов кластеризации может помочь в обнаружении нетипичных или подозрительных действий, таких как попытки несанкционированного доступа или распространение вредоносного ПО. Однако для таких целей необходимо учитывать большую изменчивость данных, что требует использования гибких методов, таких как DBSCAN, который может выявлять аномалии в плотных и разреженных областях данных.

В **финансовой сфере** методы кластеризации используются для выявления мошенничества, аномальных транзакций и других несанкционированных действий. Применение этих методов позволяет автоматизировать процесс мониторинга транзакций, определять высокие риски и принимать меры по их предотвращению. Например, алгоритм K-средних используется для группировки транзакций по схожести, что помогает найти подозрительные транзакции, которые не вписываются в нормальный контекст. Однако для

более сложных и многомерных финансовых данных, таких как прогнозирование рисков или динамики рынка, часто используется иерархическая кластеризация.

В области **здравоохранения** методы кластеризации применяются для выявления аномальных данных, связанных с диагнозами, медицинскими записями пациентов или даже для анализа биомедицинских изображений. Аномалии в медицинских данных могут указывать на ошибки в диагностике или неожиданные отклонения в состоянии пациента, которые требуют дополнительного внимания. В этой области часто используются как методы K-средних для группировки пациентов по общим признакам заболевания, так и DBSCAN для обнаружения аномальных случаев, которые не попадают в существующие кластеры.

Таблица 2 демонстрирует различные области применения методов кластеризации для выявления аномалий.

Таблица 2

Применение методов кластеризации в различных областях [11-13]

Область применения	**Метод кластеризации**	**Преимущества**	**Ограничения**
Кибербезопасность	DBSCAN	Идентификация аномальных паттернов в сетевых данных	Требует настройки параметров eps и minPts
Финансовые технологии	K-средних	Простой и быстрый для анализа транзакций	Может не распознавать аномалии в сложных данных
Здравоохранение	Иерархическая кластеризация	Позволяет выявить скрытые паттерны в медицинских данных	Высокая вычислительная сложность
Промышленность	DBSCAN	Выявление аномалий в производственных процессах	Может не справляться с большими объемами данных

В каждой области используется метод, который наиболее эффективно справляется с характерными особенностями данных. Например, в области кибербезопасности предпочтение отдается DBSCAN, поскольку этот алгоритм способен эффективно выявлять аномалии в условиях динамических и часто изменяющихся данных. В то же время в финансовых технологиях, где данные часто имеют структуру и хорошо распределены по определенным категориям, методы K-средних оказываются более эффективными [14].

С другой стороны, иерархическая кластеризация, несмотря на свою вычислительную сложность, находит свое применение в здравоохранении, где важно не только выявить аномалии, но и понять их иерархическую структуру. Этот метод помогает выявить редкие, но критически важные отклонения, например, в биомедицинских данных. Таким образом, выбор метода кластеризации зависит от специфики задач и данных, с которыми работает организация.

Методы кластеризации активно применяются для решения широкого спектра задач в различных областях. Их возможность выявлять аномалии в данных способствует повышению точности анализа, улучшению принятия решений и, в конечном счете, повышению эффективности работы в таких высоко конкурентных и критически важных отраслях, как финансы, здравоохранение и кибербезопасность.

Заключение

Методы кластеризации являются мощным инструментом для выявления аномалий в данных, благодаря своей способности группировать объекты с похожими характеристиками и выделять отклоняющиеся объекты. В данной статье рассмотрены наиболее эффективные

подходы к решению задач обнаружения аномалий с использованием алгоритмов кластеризации, таких как K-средних, DBSCAN и иерархическая кластеризация. Каждый из этих методов обладает своими особенностями, которые делают их подходящими для различных типов данных и задач. Кластеризация позволяет не только обнаружить аномальные объекты, но и понять структуру данных, что является особенно важным в таких сферах, как кибербезопасность, финансовые технологии и здравоохранение. Алгоритм DBSCAN, например, оказался эффективным для работы с шумными данными и выявления выбросов, в то время как K-средних хорошо подходит для более структурированных наборов данных, таких как финансовые транзакции. Несмотря на успешное применение этих методов в практике, важно учитывать их ограничения, такие как чувствительность к параметрам (например, eps и minPts в DBSCAN) и высокую вычислительную сложность некоторых методов, таких как иерархическая кластеризация. В будущем возможно дальнейшее развитие алгоритмов, направленных на улучшение их адаптивности и эффективности при работе с большими и многомерными данными.

Список литературы

1. Smiti A. A critical overview of outlier detection methods // Computer Science Review. 2020. Vol. 38. P. 100306.
2. Грушо А. А., Грушо Н. А., Забежайло М. И., Смирнов Д. В., Тимонина Е. Е., Шоргин С. Я. Статистика и кластеры в поисках аномальных вкраплений в условиях больших данных // Информатика и её применения. 2021. Т. 15. №4. С. 79-86.
3. Li J., Izakian H., Pedrycz W., Jamal I. Clustering-based anomaly detection in multivariate time series data // Applied Soft Computing. 2021. Vol. 100. P. 106919.
4. Ariyaluran Habeeb R. A., Nasaruddin F., Gani A., Amanullah M. A., Abaker Targio Hashem I., Ahmed E., Imran M. Clustering-based real-time anomaly detection—A breakthrough in big data technologies // Transactions on Emerging Telecommunications Technologies. 2022. Vol. 33. No.8. P. e3647.
5. Nurdinova K. Integration of Artificial Intelligence into Accounting as a Tool for Optimization and Risk Management // Bulletin of Science and Practice. 2024. Vol. 10. No. 12. P. 405-410.
6. Pu G., Wang L., Shen J., Dong F. A hybrid unsupervised clustering-based anomaly detection method // Tsinghua Science and Technology. 2020. Vol. 26. №2. С. 146-153.
7. Турашев А. С., Сухомлин В. А. Выявление аномалий в поведении ЦП с применением алгоритмов кластеризации библиотеки Scikit-Learn языка программирования Python // Современные информационные технологии и ИТ-образование. 2024. Т. 20. №1.
8. Ломакин Н. А. Применение методов машинного обучения и интеллектуального анализа данных для усовершенствования управления оборудованием // Синтез науки и общества в решении глобальных проблем. 2023. С. 200.
9. Thudumu S., Branch P., Jin J., Singh J. A comprehensive survey of anomaly detection techniques for high dimensional big data // Journal of Big Data. 2020. Vol. 7. P. 1-30.
10. Lindemann B., Maschler B., Sahlab N., Weyrich M. A survey on anomaly detection for technical systems using LSTM networks // Computers in Industry. 2021. Vol. 131. P. 103498.
11. Жилов Р. А. Интеллектуальные методы кластеризации данных // Известия Кабардино-Балкарского научного центра РАН. 2023. №6 (116). С. 152-159.
12. Багутдинов Р. А., Саргсян Н. А., Красноплахтыч М. А. Аналитика, инструменты и интеллектуальный анализ больших разнородных и разномасштабных данных // Экономика. Информатика. 2020. Т. 47. №4. С. 792-802.
13. Павлычев А. В., Стародубов М. И., Галимов А. Д. Использование алгоритма машинного обучения Random Forest для выявления сложных компьютерных инцидентов // Вопросы кибербезопасности. 2022. №5. С. 51.
14. Schmidl S., Wenig P., Papenbrock T. Anomaly detection in time series: a comprehensive evaluation // Proceedings of the VLDB Endowment. 2022. Vol. 15. No.9. P. 1779-1797.