ОБНАРУЖЕНИЕ АНОМАЛИЙ В НАБОРАХ ДАННЫХ: МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ

Ярош Е. А. Пилецкий И. И.
Кафедра программного обеспечения информационных технологий, Белорусский государственный университет информатики и радиоэлектроники
Минск, Республика Беларусь
E-mail: yarosh.catherine@gmail.com

В данной статье проводится обзор типов аномалий, а так же набор современных методов машинного обучения, используемых для классификации и идентификации аномалий в наборах данных, таких как автокодер, изолирующий лес и эллипсоидальная аппроксимация данных.

Введение

Обнаружение аномалий в наборах данных - это важная задача для обеспечения безопасности и получения экономической выгоды. Аномалиями называют отклонения или выбросы от нормального поведения при наблюдении различных явлений или экспериментальных данных. Важно понимать, что это за аномалии и их причины. Это могут быть сбои системы или злонамеренные действия. Поэтому в современных условиях, в системах принятия решений это крайне важная задача. Для обнаружения аномалий в современных системах применяются различные методы и алгоритмы машинного обучения.

Поиск аномалий и выявление подозрительных операций с помощью машинного обучения широко применяется в клиентской аналитике, банковском аудите и многих других видах аналитики.

I. Виды аномалий

Аномалии можно условно разделить на три типа: точечные аномалии, контекстуальные аномалии и коллективные аномалии.

Точечная аномалия [1] часто представляет собой неровность или отклонение, которое происходит случайно и может не иметь особой интерпретации. Например, на рисунке 1 транзакции кредитной карты с большими расходами, зарегистрированной в ресторане Monaco, первая транзакция кажется точечной аномалией, поскольку она значительно отклоняется от остальных сделок.

Аномальные наборы отдельных точек данных известны как *коллективные аномалии* [3], при этом каждая из отдельных точек в отдельности выглядит как экземпляры нормальных данных, в то же время в группе они демонстрируют необычные характеристики. Например, рассмотрим иллюстрацию мошеннической транзакции по кредитной карте в данных журнала, показанных на рисунке 1, если произошла бы одна транзакция «MISC», это, вероятно, не могло бы показаться аномальным. Следующая группа из транзакций на сумму 75 долларов, безусловно, кажутся кандидатами на коллективную или групповую аномалию.

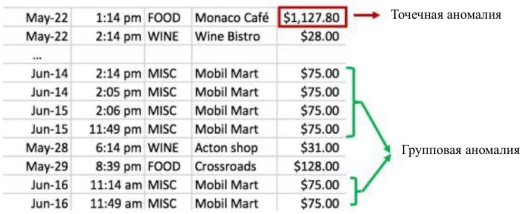

Рис. 1 – Обнаружение мошенничества с кредитными картами: иллюстрации точечной и коллективной аномалии

Контекстуальная аномалия также известная как условная аномалия, представляет собой экземпляр данных, который можно рассматривать как аномальный в некотором конкретном контексте. Это означает, что наблюдение за одной и той же точкой в разных контекстах не всегда будет свидетельствовать об аномальном поведении. Контекстуальная аномалия определяется сочетанием контекстуальных и поведенческих признаков [3]. Для контекстуальных признаков чаще всего используются время и пространство, тогда как поведенческие признаки зависят от анализируемой области — суммы потраченных денег, средней температуры или какой-либо другой количественной меры, которая используется в качестве признака.

Рисунок 2 иллюстрирует пример контекстуальной аномалии с учетом данных о температуре, обозначенных резким падением незадолго до июня. Это значение не указывает на нормальное значение, найденное за это время.

Рис. 2 – Температурные данные Hayes и Capretz

II. Методы машинного обучения для обнаружения аномалий

На данный момент самыми популярными алгоритмами машинного обучения являются следующие:

1. Автокодер
2. Изолирующий лес
3. Эллипсоидальная аппроксимация данных

Автокодер [2] представляет собой нейросеть прямого распространения, обученную с помощью обратного распространения ошибки. Автокодеры особым образом используют свойство нейронной сети для реализации некоторых эффективных методов обучения сетей для изучения нормального поведения. Когда поступает точка данных с выбросом, автокодер не может ее хорошо кодировать. Он научился представлять шаблоны, не существующие в этих данных. При попытке восстановить исходные данные из их компактного представления реконструкция не будет похожа на исходные данные. Это помогает обнаруживать аномалии при их возникновении. Цель такого процесса — попытаться восстановить исходный ввод из закодированных данных, что имеет решающее значение при создании модуля обнаружения аномалий.

В архитектуре автокодера есть 3 основные части:

- *Кодировщик* уменьшает размерность набора данных высокой размерности до низкоразмерного;
- *Код* содержит сокращенное представление ввода, который подается в декодер;
- *Декодер* расширяет данные низкой размерности до данных высокой размерности.

Изолирующий лес — это алгоритм обучения без учителя, который выявляет аномалии, изолируя выбросы в данных [3].

Изолирующие леса строятся на основе деревьев решений. В них случайно отобранные данные обрабатываются в древовидной структуре на основе случайно выбранных признаков. Образцы, которые перемещаются глубже в дерево, с меньшей вероятностью будут аномалиями, поскольку для их изоляции требуется больше разрезов. Точно так же образцы, которые заканчиваются более короткими ветвями, указывают на аномалии, поскольку дереву было легче отделить их от других наблюдений.

Изолирующий лес рекурсивно создает разделы в наборе данных, случайным образом выбирая объект, а затем случайным образом выбирая значение разделения для объекта. Предположительно, для аномалий требуется изолировать меньшее количество случайных разделов по сравнению с «нормальными» точками в наборе данных, поэтому аномалии будут точками с меньшей длиной пути в дереве, причем длина пути представляет собой количество ребер, пройденных от корневого узла.

В **эллипсоидальной аппроксимации данных** [4], как следует из названия, облако точек моделируется как внутренность эллипсоида. Метод хорошо работает только на одномодальных данных, а особенно хорошо – на нормально распределённых. Степень новизны здесь фактически определяется по расстоянию Махалонобиса.

Одним из распространенных способов обнаружения аномалий появляется в предположении, что данные распределены каким-то известным способом, например, по Гауссу. В таком случае задача заключается в определении вида этого распределения и выделении тех объектов, которые не удовлетворяют найденному распределению.

Степень аномальности объекта определяется по расстоянию Махалонобиса, в математической статистике являющимся мерой расстояния между векторами случайных величин и обобщающим понятие евклидова расстояния. В задаче определения принадлежности точки одному из классов необходимо найти матрицы ковариации всех классов, что, обычно, делается на основе известных выборок из каждого класса. При вычислении расстояния Махаланобиса до каждого класса выбирается тот класс, для которого это расстояние оказалось минимальным, что эквивалентно методу максимального правдоподобия (Maximum likelihood estimation, MLE). Точка, имеющая наибольшее расстояние Махаланобиса до остального множества точек, считается аномалией. Такая точка имеет наибольшее влияние на кривизну и на коэффициенты уравнения регрессии. Также расстояние Махаланобиса может быть использовано в задаче определения многомерных выбросов.

III. Заключение

В данной статье были рассмотрены три типа аномалий в наборах данных. Были описаны наиболее эффективные алгоритмы машинного обучения для обнаружения аномалий на сегодняшний день: автокодера, изолирующего леса и эллипсоидальной аппроксимации данных.

IV. Список литературы

1. V. Chandola, A. Banerjee, V. Kumar, “Anomaly detection: A survey”, ACM Computing Surveys, vol. 41(3), 2009, pp. 1–58
2. Anomaly detection in cardio dataset using deep lerning technique: autoencoder [Electronic resource]: Suchimisita Sahu, 2021.– Mode of access: https://medium.com/analytics-vidhya/anomaly-detection-in-cardio-dataset-using-deep-learning-technique-autoencoder-fd24ca9e5c69. – Date of access: 10.10.2022.
3. Ho, Tin Kam. Random Decision Forests // Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC. — 1995. — Pp. 278–282.
4. G. Calafiore, “Approximation of n-dimensional data using spherical and ellipsoidal primitives”, IEEE Transactions on Systems Man and Cybernetics - Part A Systems and Humans, vol. 10, April 2002.