Обзор литературы по теме:

«Количественные методы выявления аномалий во временных рядах»

Чиркова Юлия, БЭАД221

10 декабря 2023 г.

1 Введение

Наличие аномалий в данных – это обширный и важный вопрос, волнующий специалистов, работающих с данными, не один десяток лет. С одной стороны, аномалии могут возникать из-за обыкновенных ошибок при записи или переносе информации или из-за намеренного сокрытия определенных данных. С другой стороны, выбросы могут быть истинными данными, свидетельствующими об исключительном феномене или о нештатной работе некоторого механизма, платформы или сайта.

Так или иначе, датасаентистам, аналитикам и другим специалистам необходимо уметь точно, качественно и с малыми затратами ресурсов выявлять аномалии в данных. В эпоху больших данных сделать это «вручную» - практически невозможная миссия, поэтому на помощь человечеству здесь приходят статистика, компьютерные алгоритмы и машинное обучение. Данная работа содержит обзор ряда статей, связанных с выявлением аномалий во временных рядах.

Прежде чем приступать к изучению данной темы, необходимо ответить на некоторые вопросы, фундаментальные для понимания всего последующего материала: что такое временные ряды? Что представляют из себя аномалии в данных и почему их необходимо выявлять?

2 Временные ряды

Обширное описание временных рядов приведено в работе “An Introductory Study on Time Series Modeling and Forecasting” авторства R. Adhikari, R. K. Agrawal [1].

Временной ряд – это набор измерений (данных), сделанных в течение некоторого периода и хронологически упорядоченных. Притом могут осуществляться измерения как одной, так и нескольких величин, поэтому с точки зрения математики временной ряд представляется как набор векторов $X(t), t = 0, 1, 2, \ldots$, где t – номер измерения.

Временной ряд называется непрерывным, если наблюдения измеряются без остановки в течение некоторого отрезка времени. Примерами непрерывных временных рядов являются электрокардиограмма пациента, показания температуры воздуха и др. Временной ряд называется дискретным, если он содержит измерения, которые были сделаны не безостановочно, а дискретно, в определенные моменты времени, как правило через относительно равные временные промежутки (час, день, неделя, месяц, год). К дискретным временным рядам можно отнести, например, ВВП страны, численность населения города, значение ключевой ставки и др.

Выделяют 4 вида составляющих декомпозиции временного ряда: трендовые, циклические, сезонные и нерегулярные.

- Трендовые составляющие отвечают за наиболее общую тенденцию временного ряда, т.е. за то, как в целом временной ряд «ведет себя» в долгосрочном периоде: он растет, падает или находится на неизменном уровне.

- Сезонные составляющие представляют колебания измеряемых показателей в течение года по сезонам, на них особое влияние оказывают погода и климат в рассматриваемом регионе, традиционные мероприятия местных народов, праздники и т.д.

- Циклические составляющие отвечают за изменения в среднесрочной перспекти-

ве, которые повторяются циклически: за одной фазой неизменно следует другая. Данные составляющие особенно заметны и важны для временных рядов в экономике и финансах из-за влияния на них экономического роста и делового цикла.

- Наконец, нерегулярные составляющие – это различные факторы, возникновение которых не подвержено никаким закономерностям - напротив, они возникают в непредсказуемые моменты (война, дефолт, забастовка и т.д.).

Обычно в анализе и прогнозировании временных рядов величина $X(t)$ рассматривается как случайный вектор; также предполагается, что данный вектор имеет некоторое (известное, модельное) распределение. Структуру временного ряда принято описывать с помощью случайного, или стохастического, процесса. Как правило, считают, что измерения независимы и распределены нормально, однако на практике это верно не всегда: например, если сегодня на улице отрицательная температура сегодня, то, скорее всего, завтра она тоже будет отрицательной, т.е. на самом деле между измерениями может наблюдаться зависимость.

Выделяют стационарные временные ряды – среднее и дисперсия таких временных рядов не меняются со временем, т.е. они не обладают трендом. Более того, в таких временных рядах ковариация двух измерений зависит только от длины временного промежутка между ними. Со стационарными временными рядам гораздо проще и удобнее работать, а свойство стационарности у временного ряда можно выявить с помощью теста Дики-Фуллера. Зачастую временные ряды, отражающие сезонные показатели или тренды не являются стационарными, однако существуют различные способы (дифференцирование и степенные преобразования), позволяющие получить стационарный временной ряд из нестационарного.

Таким образом, в рассматриваемой работе приведено всеобъемлющее объяснение временных рядов, их свойств, способов моделирования, анализа и прогнозирования. Авторы детально и на продвинутом уровне вводят читателя в новую тему, проходя

путь от базовых определений и простых примеров применения временных рядов до построения и оценки сложных моделей. Преимуществами данной работы являются довольно строгий математический подход, большое количество наглядных графиков и обширный практический блок – изученные метрики и концепции применяются на реальных данных.

3 Аномалии в данных

Работа “On the nature and types of anomalies: a review of deviations in data”, написанная R. Foorthuis, [5] хорошо раскрывает понятие "аномалия"и предоставляет фундаментальную теорию по данной теме.

Согласно статье, аномалии в данных – это некоторая группа измерений, которые отличаются от характерного поведения рассматриваемого показателя. Выделяют несколько видов аномалий временных рядов:

- точечные (в конкретных точках значения показателя отличаются от типичных – например, они значительно отличаются по величине),
- групповые (в каждой конкретной точке значение показателя типично, но группа точек «ведет себя» отлично от других измерений – например, значения могут находиться в привычном интервале, но вид образуемой ими зависимости может отличаться)
- контекстные аномалии (значения, которые являются аномальными только в определенном контексте).

Поиск и классификация аномалий чрезвычайно важны: большие данные и временные ряды используются в множестве окружающих нас систем – от роботов-пылесосов и смарт-часов до эквайринговых платформ и транспортных систем. Наличие определен-

ных видов аномалий может свидетельствовать о внутренней неисправности в определенном блоке такой системы или о внешнем инциденте, который породил выбивающиеся значения и требует немедленного вмешательства. Игнорирование аномалий может стоить огромных экономических или человеческих потерь, поэтому им нужно уделять особое внимание.

Анализ выбросов затрудняется тем, что какие-то из них нам необходимо принять во внимание и обработать, а какие-то являются просто неточностями, ошибками, и нам не нужно, чтобы наша модель учитывала их. Для разных случаев разработано множество различных способов выявления аномалий.

Итак, в своей статье автор представляет глубокое теоретическое объяснение термина «аномалия», объясняет свойства аномалий и представляет довольно полную их классификацию, основанную на их природе и важности. Исследователь также классифицирует аномалии в зависимости от типа данных (количественные/качественные признаки, одномерные/многомерные) и приводит разнообразные продвинутые примеры для каждой группы.

Следовательно, преимуществом полученного результата можно назвать глубокую теоретическую проработку темы, снабженную качественными примерами. Однако данной статье не хватает практического блока: возможно, для каждого класса аномалий необходимо приложить способы их определения и нивелирования.

4 Метрики для оценки моделей машинного обучения

В данном литобзоре будут упоминаться некоторые метрики, описывающие качество моделей машинного обучения – они подробно объяснены в статье “Performance Metrics for Machine Learning Models” ученых B. J. Erickson, F. Kitamura [4].

В своей работе авторы объясняют следующие понятия на примере бинарной классификации (для простоты возьмем классы A и B, и пусть для нас исход True – это

принадлежность объекта к классу A):

- True Positive (TP) – количество объектов, которые принадлежат классу A и были верно определены моделью в этот класс
- True Negative (TN) - количество объектов, которые принадлежат классу B и были верно определены моделью в этот класс
- False Positive (FP) – количество объектов, которые принадлежат классу B, но были ошибочно определены моделью в класс A
- False Negative (FN) - количество объектов, которые принадлежат классу A, но были ошибочно определены моделью в класс B

Тогда базовые метрики выражаются следующим образом:

- точность (accuracy): $\frac{TP+TN}{TP+TN+FP+FN}$
- чувствительность (sensitivity): $\frac{TP}{TP+FN}$
- специфичность (specificity): $\frac{NP}{TN+FP}$
- False Positive Rate (FPR): $\frac{FP}{TN+FP}$
- True Positive Rate (TPR): $\frac{TP}{TP+FN}$

Заметим, что последние две метрики рассчитываются для определенного уровня порога x, разграничивающего принадлежность объектов к одному или другому классу, то есть формально $FPR = FPR(x), TPR = TPR(x)$. Тогда ROC-кривая – это $TPR(FPR^{-1})$, а AUC – площадь под ROC-кривой

Таким образом, в данной статье предложено емкое и довольно полное описание основных метрик для оценки моделей машинного обучения. Рассмотрены как базовые показатели, так и более продвинутые, приведены графические иллюстрации, смысл

индикаторов объяснен на довольно глубоком уровне. Тем не менее, недостатком данной статьи можно назвать отсутствие практических примеров подсчета и применения данных метрик для оценки и сравнения реальных моделей.

Далее в данном обзоре будет рассмотрена литература по различным методам выявления аномалий, применимым ко временным рядам, притом мы пойдем от самых простых и неточных к более сложным и эффективным.

5 Z-score (z-value)

z-score – это один из основных статистических показателей. Он рассчитывается по следующей формуле: $z = \frac{x_i - E(x)}{\sqrt{Var(x)}}$, где $E(x)$ - среднее значение измерения, $Var(x)$ - дисперсия измерения. Z-score отражает, насколько далеко данное наблюдение находится от среднего значения по всей выборке. Иначе говоря, z-score показывает, на сколько стандартных отклонений данное измерение отстоит от среднего. Данный показатель используется в аналитике, финансах, A/B тестах и в иных областях.

В своей работе “Anomaly detection by robust statistics” авторы P. J. Rousseeuw, M. Hubert [9] рассмотрели z-score, оценили его эффективность и сравнили с иными методами поиска выбросов. Они проанализировали целую совокупность методов: статистическй подход, оценки ковариации, линейную регрессию, метод главных компонент, кластеризацию, классификацию и др. Алгоритм использования конкретно z-score состоит в том, чтобы подсчитать данный индикатор для всех измерений и затем установить некоторое пороговое значение: все значения в пределах порога считаются нормальными, а все, что останется – аномалии.

Исследователи пришли к выводу, что преимуществами z-score является то, что данная оценка безразмерна, т.е. нам не нужно учитывать единицы измерения и делать на них поправку. Кроме того, она устойчива к изменению порядков данных: нам не нужно думать, работаем мы с единицами или с миллионами, что упрощает анализ.

Наконец, z-score – это некоторая стандартизированная мера, позволяющая сравнивать наборы данных с разными средними и стандартными отклонениями.

Тем не менее, к минусам использования данного показателя «в чистом виде» относится то, что нам придется самостоятельно устанавливать пороговое значение при анализе аномалий, что порождает новую трудоемкую задачу. Кроме того, даже если мы каким-то образом найдем это пороговое значение, эффективность z-score при поиске аномалий, близких по значению к нормальным измерениям, стоит под вопросом.

Исследователи не делают никаких предположений о качестве, характере или структуре данных. Надо заметить, что данные статьи в целом являются скорее обзорными, чем глубокими. То есть в них аккумулируются и сравниваются некоторые инструменты, созданные и описанные другими учеными, причем авторы не углубляются в детали и глубинный смысл того или иного механизма, а скорее сравнивают их эффективность на реальных данных. Плюсом данной работы является рассмотрение и сравнение множества различных инструментов. Однако авторы тестируют их на довольно немногочисленных и необъемных датасетах, что приуменьшает репрезентативность полученных результатов – это минус.

6 Межквартильный размах

Квартиль порядка q – это некоторое число из множества доступных значений, меньше которого находятся ровно $100 * q\%$ всех измерений.

IQR (Interquartile Range, межквартильный размах) – это разница между квартилем порядка 0,75 ($q_{0.75}$, больше данного значения находится лишь 25% всех значений и выбросы) и квартилем порядка 0,25 ($q_{0.25}$, меньше данного значения находится 25% всех измерений и выбросы).

Винзоризация (winsorizing) – это процесс замены всех значений, превышающих заданный порог, на данный порог. Т.е. если нижней границей является $q_{0.25}$, то все

значения, меньшие данного квартиля, заменяются на него.

В исследовании "An outliers detection and elimination framework in classification task of data mining" ученые K. D. Sanjeev, A. K. Behera, S. Dehuri, A. Ghosh [2] используют эти два метода выявления аномалий совместно. Авторы воспринимают все значения, меньшие $q_{0.25} - 1.5 * IQR$ или большие $q_{0.75} + 1.5 * IQR$ как выбросы. Они опираются на предположение о том, что данные одномерны и аномалии существенно отличаются от «нормальных» значений.

К преимуществам этого подхода относятся простота подсчета и понятность интерпретации. Более того, авторы показали, что модель, обученная на данных, из которых выбросы были удалены с помощью $IQR + winsorizing$, имеет более высокую предсказательную способность, чем модели, обученные на иных датасетах (предобработанных или нет). Т.е. в данном исследовании $IQR + windsorizing$ довольно эффективно справились с обработкой выбросов.

Однако минусом этого показателя является то, что он основан всего на 2 значениях из датасета и ничего не говорит о распределении данных. Более того, межквартильный размах довольно грубый: если нормальные данные сосредоточены вокруг медианы, то аномалии вполне могут находиться внутри квартилей 0.25-0.75. Наконец, IQR – не самый удобный и эффективный способ для работы с многомерными измерениями.

7 k-Nearest Neighbours (KNN)

В научной работе "A Review of Anomaly Detection Techniques Based on Nearest Neighbor" исследователи M. Zhao, J. Chen, Y. Li [12] рассматривают такой метод выявления аномалий, как k ближайших соседей. Это один из базовых методов обучения без учителя, суть которого состоит в идентификации объекта на основе свойств его k ближайших соседей. Алгоритм таков:

1. выбираем линейно независимые признаки, чтобы не хранить и не обрабатывать лишние объемы информации: для этого используются filter techniques (признаки выбираются до начала обучения алгоритма) и wrapper techniques (признаки выбираются во время обучения модели, основываясь на получаемой точности)

2. нормируем все измерения: KNN применяется в т.ч. для многомерных измерений, из-за чего разные признаки могут иметь разный масштаб, поэтому для обеспечения равного влияния признаков их нужно привести к одной шкале (обычно это делается с помощью min-max нормализации или замены значения на его z-score)

3. выбираем метрику для вычисления близости между измерениями (обычно это евклидова метрика, но может быть и иная)

4. выбираем k – число соседей, которых будем рассматривать - и задаем пороговое значение расстояния

5. вычисляем оценку аномалии: это может быть расстояние от данного измерения до k его ближайших соседей; количество соседей измерения в шаре заданного радиуса; расстояние до k-того соседа в отсортированном по расстоянию списке и т.д.; если оценка аномалии не превосходит заданный порог, то это нормальное значение, иначе - аномалия

В данной статье авторы предполагают, что плотность распределения аномалий низкая, и аномалии значительно отличаются по своим признакам от нормальных измерений. Исследователи начинают с определения оценки аномалии, далее они описывают модификации KNN, в т.ч. ORCA (Outlier Detection and Robust Clustering), LOF (Local Outlier Factor), COF (Connectivity-based Outlier Factor) и др. Затем они проводят теоретическое сравнение KNN с иными методами.

Ученые приходят к выводу, что преимуществами KNN являются способность работать с разными типами данных (при задании хорошей метрики для них), возможность

работать с данными вне зависимости от их распределения, простота интерпретации. К недостаткам, в свою очередь, относятся относительно немалые траты вычислительных мощностей, низкая способность выявлять аномалии, близкие по значению к нормальным измерениям, сложности задания метрики на сложных структурах данных.

Данная работа – это чисто теоретический обзор KNN и сопряженных методов. Авторы приводят емкое и качественное описание метода, его свойств, модификаций, сильных и слабых сторон, однако не предлагают собственных нововведений, не сравнивают эффективность KNN и других методов на реальных данных, т.е. какой-либо практический блок отсутствует.

Метод KNN был впоследствии развит многими исследователями. Так, к примеру, в работе "Outlier detection: How to Select k for k-nearest-neighbors-based outlier detectors" ученые J. Yang, X. Tan, S. Rahardja [11] разработали метод согласованности окрестностей (Neighbourhood consistency). Его идея в следующем:

1. с помощью KNN некоторым образом присваиваем оценки аномалий каждому из измерений и для каждого измерения в датасете отсекаем его k ближайших соседей

2. для каждого измерения i и его соседей выбираем 2 числа: u_i (оценка аномалии самого измерения) и v_i (средня оценка аномалии среди измерения и его k соседей) и записываем их в векторы U, V

3. рассчитываем значение согласованности окрестностей как $1 - cos(U, V)$

4. из различных сгенерированных KNN присвоений оценок аномалий выбираем то, у которого значение согласованности окрестностей наибольшее; из такого присвоения мы получим наиболее точную оценку на то, какие измерения являются аномальными

В своей работе авторы делают предположение о том, что аномалии разрежены и далеки от нормальных значений. После описания алгоритма ученые тестируют его на 12 датасетах по значению AUC и приходят к выводу о том, что предложенный ими алгоритм превосходит MOD (Mean-shift Outlier Detector), LOF (Local Outlier Detector), обычный KNN, ODIN (Outlier Detection using Indegree of Nodes) по рассматриваемой метрике.

Преимуществами разработанного алгоритма являются:

- способность работать с многомерными данными
- возможность работы с признаками вне зависимости от их распределения
- повышенная точность распознавания аномалий
- относительно низкие затраты времени

Однако в качестве недостатков нового метода можно выделить значительные затраты вычислительных мощностей, пониженную эффективность определения аномалий при «похожести» аномалий и нормальных значений.

Представленная работа дает хорошее понимание о том, что такое KNN, как данный метод применяется в поиске аномалий, какие его вариации уже созданы. Кроме того, ученые предлагают свою модификацию KNN, обосновывают ее эффективность путем обширного эксперимента. Однако, на мой взгляд, тестирование метода с помощью одного только AUC – это некоторое упущение авторов.

8 Метод опорных векторов

В статье “High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning”[8] ученые S. M. Erfani, S. Rajasegarar, S. Karunasekera, Ch.

Leckie [3] предлагают алгоритм выявления аномалий, основанный на одноклассовом методе опорных векторов (Supporting Vectors Machine, SVM).

Суть алгоритма такова: нормальные данные и аномалии рассматривают как два класса, которые нужно отделить друг от друга. Если они линейно отделимы друг от друга, то SVM; ищет «наилучшую» разделяющую гиперплоскость между ними. Наилучшая разделяющая гиперплоскость – такая, до которой расстояние от крайних точек обоих классов (они называются опорными векторами) максимально. Если же данные линейно неразделимы, то вводится новая ось (или оси), обеспечивающая явное разделение измерений.

В рассматриваемой статье авторы используют комбинацию DBN (Deep Belief Network) для преобразования признаков и уменьшения размерности, и результаты данного алгоритма уже подвергаются одноклассовому методу опорных векторов. Ученые тестируют полученный алгоритм на 8 различных датасетах и сравнивают с другими методами выявления аномалий, основанными на DBN, по AUC, времени обучения модели и времени отработки модели на тестовой выборке. Исследователи приходят к выводу о том, что разработанный алгоритм превосходит практически все представленные методы по всем метрикам.

Сильные стороны нового алгоритма таковы:

- способность работать с многомерными данными
- меньшие временные затраты
- эффективность при работе с большими датасетами
- способность работать на неразмеченных данных Однако минусом $DNB-1SV$M являются значительные траты вычислительных мощностей на больших объемах данных.

9 Кластеризация

В статье “A survey of different methods of clustering for anomaly detection” ученых S. Tripathy, L. Sahoo [10] подробно описано использование кластеризации для выявления аномалий.

Кластеризация – это метод обучения без учителя, который объединяет многомерные векторы-измерения, основываясь на их близости по определенному признаку. Таким образом алгоритм разделяет данные на группы – кластеры: все измерения в одном кластере «похожи» друг на друга, тогда как измерения из разных кластеров сильно различаются.

Как правило, реальные данные содержат линейно зависимые признаки, и совместный анализ – лишняя трата времени и вычислительных мощностей. Поэтому перед кластеризацией данные сначала избавляют от «лишних» измерений, уменьшая размерность.

Важным этапом кластеризации также является выбор метрики, по которой будет оцениваться близость измерений. Если речь идет о числовых векторах, обычно в качестве метрики используют Евклидово расстояние, однако в зависимости от специфики и предназначения данных метрику можно задать самостоятельно.

Наконец, после кластеризации аномалиями считаются либо измерения, попавшие в определенный класс, либо измерения, расстояния до которых от ближайшего кластера превышают определенный порог.

Предположение, которые авторы делают в статье, состоит в том, что аномальных данных в датасете намного меньше, чем нормальных, и выбросы значительно отличаются от аномалий.

К преимуществам данного метода относятся способность кластеризации работать на неразмеченных данных и возможность алгоритма разбивать данные на группы без каких-либо изначальных знаний о них и даже без задания количества кластеров

Однако кластеризация имеет значительные недостатки: алгоритмы кластеризации чаще всего работают итеративно, с использованием матриц и требуют постоянного пересчитывания метрики. Поэтому кластеризация требует значительных затрат времени, памяти и вычислительных мощностей.

Итак, в рассматриваемой статье дано очень полное и подробное описание кластеризации. Объяснено, какие плюсы и минусы присущи кластеризации, какие ограничения она имеет, как и зачем происходит выбор признаков и их редуцирование, описаны основные группы алгоритмов кластеризации – от метода k-средних до ко-кластеризации. Однако в статье не представлены результаты работы данных алгоритмов на различных датасетах, их точность и эффектвность.

Кластеризация также зарекомендовала себя как удачный метод для поиска аномалий в данных, данный подход был в дальнейшем продолжен и улучшен.

Так, например, в статье “A Mixed Clustering Approach for Real Time Anomaly Detection” исследователей F. A. Mazarbhuiya, M. Shenify [8] предложен алгоритм, сочетающий в себе кластеризацию k-средних и агломеративную иерархическую кластеризацию. Разработанный ими алгоритм применяется к многомерным измерениям, каждому из которых приписано время его совершения. Он состоит из следующих шагов:

1. разбиваем все измерения на k начальных кластеров, основываясь на их близости к центроидам (фактически, средним) кластера

2. для каждого кластера обновляем средние и «жизненный период» (промежуток между наименьшим и наибольшим временем измерений кластера) и смотрим, какие векторы должны изменить свой класс

3. если такие векторы есть, то повторяем шаг (2), пока таких векторов не останется

4. объединяем кластеры на основе их жизненных периодов: если пересечение временных промежутков двух кластеров больше определенного порога, то объеди-

няем их и пересчитываем жизненные периоды; если кластеров для объединения не осталось, то алгоритм закончен

5. измерения, которые не принадлежат ни одному из кластеров, принадлежат разреженному кластеру или принадлежат малому количеству кластеров (суммарно, с учетом объединений), рассматриваются как аномалии

Предположение авторов таково: аномалии разрежены, т.е. они не сконцентрированы, не образуют собственный класс.

Чтобы протестировать разработанную модель, ученые провели эксперимент на 2 объемных датасетах с большим количеством параметров. Исследователи сравнивали новый алгоритм с уже существующими (k-means, PCM, ACA, IF, OnCAD) по таким характеристикам, как точность, чувствительность, специфичность, и во всех случаях он показал наилучшие результаты.

Таким образом, плюсы смешанного алгоритма таковы:

- хорошо работает с численными, категориальными и временнЫми признаками
- при изначально заданном числе кластеров сам определяет оптимальное число кластеров данных
- обладает более высокой точностью по сравнению с «предшественниками»
- временные затраты меньше, чем у известных алгоритмов

Однако у данного подходя есть и минусы: аномалии могут повлиять на установление центроидов, что исказит реальные классы нормальных данных. Кроме того, аномалии могут «собраться» в собственный кластер, и этот случай будет обработан с ошибками.

Подводя итоги, данная работа – это пример глубокого исследования с обширным теоретическим блоком, представлением, подробным описанием и тестированием собственного нововведения.

10 Изолирующий лес

Исследователи F. T. Liu, K. M. Ting, Z.-H. Zhou преставили метод изолирующего леса в своей одноименной статье “Isolation Forest” [7].

Изолирующий (разделяющий) лес – это алгоритм обучения без учителя, который позволяет выявлять аномалии в данных с помощью построения ряда случайных двоичных деревьев решений. В каждом узле дерева случайным образом выбираются очередной признак и его пороговое значение; все измерения, которые не превышают данный порог, отправляются в левый дочерний узел, остальные – в правый. Далее алгоритм рекурсивно повторяется до тех пор, пока каждое измерение не окажется в своем листе (конечном узле), или пока не будет достигнута максимальная заданная высота дерева.

После запуска алгоритма высчитываются длины путей от корня до каждого листа. Данный алгоритм во многом опирается на рандомизацию, поэтому обычно повторяется несколько раз, и затем полученные длины путей усредняются. Утверждается, что при такой обработке данных аномалии будут быстро оказываться в «крайних» нодах, из-за чего средняя длина пути от корня до них будет значительно меньше, чем средняя длина пути от корня до листьев, содержащих нормальные значения.

Предположения, сделанные учеными, таковы: аномалии в данных немногочисленны; выбросы существенно отличаются от остальных данных; данные различны, т.е. измерения уникальны или же повторяющихся значений очень немного.

Авторы создали полный алгоритм и протестировали его на 12 различных датасетах, включающих естественные и синтетически сгенерированные данные, данные без выбросов, данные с различной плотностью распределения аномалий. Ими был введен индикатор «оценка аномалии», равный $2^{-\frac{E(h)}{c(n)}}$, где $E(h)$ - средняя длина пути от корня до данного измерения в дереве; $c(n)$ - средняя длина пути от корня до любого измерения при n запусках алгорита. Исследователи также сравнивали данный алгоритм с

LOF, RF, ORCA по времени отработки и AUC.

Явными преимуществами IF являются:

- низкая вычислительная сложность
- малые временные затраты (благодаря линейной сложности)
- эффективность при работе с большими массивами данных
- способность обнаруживать как точечные аномалии, так и их скопления
- малые затраты на хранение
- способен работать с многомерными данными с множеством «лишних» параметров
- хорошо обучается даже на маленьких датасетах
- не основывается на предположении о некотором модельном распределении данных

При всех преимуществах изолирующего леса, он все же имеет несколько недостатков. Во-первых, разделяющий лес – это модель обучения без учителя, что лишает пользователя возможности управлять процессом обучения модели и обработки данных. В ситуациях работы со специфическими датасетами или данными, где точность играет важную роль, это может быть существенным недостатком. Во-вторых, значения во временных рядах, как правило, находятся в определенном диапазоне, т.е. вероятность существования повторяющихся или очень близких данных велика. Это неизбежно скажется на работе изолирующего леса, поскольку он лучше всего работает на наборе из уникальных значений. Наконец, не всегда аномалии сильно отличаются от нормальных показателей, иногда они крайне близки по значению. Точность работы изолирующего дерева на таких датасетах находится под вопросом.

Идея использования изолирующего леса для выявления аномалий получила широкое распространение и развитие. В частности, ученые S. Hariri, M. C. Kind, R. J. Brunner в своей статье "Extended Isolation Forest"[6] представили модель расширенного изолирующего леса.

Исследователи указали, что поскольку в каждой ноде признаком является случайное значение из интервала доступных значений рассматриваемого признака, в многомерном случае IF как бы делает разрез пространства, параллельный одной из осей. Это вызывает смещение, или предвзятость, модели при работе с тестовым датасетом. Иными словами, алгоритм повторяет эти «вертикально-горизонтальные» паттерны при поиске аномалий на любых новых данных – даже там, где этих паттернов нет. В качестве визуализации недостатков обычного изолирующего леса исследователи построили так называемые «тепловые карты», отображающие оценку аномалии для каждой точки плоскости, и сравнили их с реальным распределением измерений.

Чтобы решить данную проблему, ученые предложили строить гиперплоскости со случайным наклоном: для этого требуется случайно выбранная точка пространства и само значение наклона, выбранное рандомно. Соответственно, предположение данное работы состоит в том, что в каждом из полученных полупространств находится некоторое ненулевое количество измерений.

Исследователи провели эксперименты на 5 датасетах, содержащих данные различного характера, в том числе естественных и синтетических. Они сравнили работу обычного изолирующего леса, повернутого изолирующего леса (еще одна модификация IF) и расширенного изолирующего леса. Были проанализированы такие метрики как дисперсия, AUC, сходимость оценки аномалии и «тепловые карты» оценок аномалий.

К сильным сторонам расширенного изолирующего леса относятся:

- более точная классификация измерений как аномалия/нормальное значение
- более низкая дисперсия оценок аномалий

- AUC, сходимость аномалии не ниже, чем у двух других видов изолирующих лесов
- все те же плюсы, которые имеет обычный изолирующий лес

Тем не менее, у расширенного изолирующего леса есть существенный недостаток. В силу случайности выбора точки и наклона прямой, он генерирует довольно много пустых полупространств, т.е. лишних ветвей. Это увеличивает сложность алгоритма, не привнося при этом никакой дополнительной отдачи, точности или эффективности.

Две вышеуказанные работы представляют собой высококачественные научные статьи, написанные с использованием продвинутого математического аппарата. В них рассматриваемые инструменты анализируются с различных сторон, есть многочисленные ссылки на исследования предшественников в данной сфере, и вклад исследователей четко отделен от того, что было изобретено и создано ранее.

11 Заключение

Итак, в данном литературном обзоре были рассмотрены различные методы выявления аномалий во временных рядах. Современные подходы к поиску выбросов в данных включают в себя статистические методы, метод К ближайших соседей, метод опорных векторов, кластеризацию и изолирующий лес. Каждый из указанных методов имеет многочисленные модификации, а также свои сильные и слабые стороны. Следовательно, наиболее оптимальной "стратегией"при решении задачи поиска выбросов является выбор метода с учетом специфики данных, ресурсов и задач конкретного исследования.

Важной проблемой в сфере обнаружения аномалий является то, что не существует "идеального"алгоритма: если модель превосходит по скорости, то она уступает по памяти; если превосходит по памяти и скорости, то уступает по точности и затрачиваемым вычислительным ресурсам и т.п. Поэтому одно из возможных направлений для исследования в данной области - это конструирование и тестирование моделей и алгоритмов, сочетающих в себе рассмотренные базовые подходы к обнаружению аномалий. Грамотная комбинация позволит взять лучшие черты каждого из алгоритмов и построить модель, которая будет лучше "предшественников"по каждому из желаемых параметров.

Список литературы

[1] Ratnadip Adhikari and Ramesh K Agrawal. An introductory study on time series modeling and forecasting. *arXiv preprint arXiv:1302.6613*, 2013.

[2] Ch Sanjeev Kumar Dash, Ajit Kumar Behera, Satchidananda Dehuri, and Ashish Ghosh. An outliers detection and elimination framework in classification task of data mining. *Decision Analytics Journal*, 6:100164, 2023.

[3] Sarah M Erfani, Sutharshan Rajasegarar, Shanika Karunasekera, and Christopher Leckie. High-dimensional and large-scale anomaly detection using a linear one-class svm with deep learning. *Pattern Recognition*, 58:121–134, 2016.

[4] Bradley J Erickson and Felipe Kitamura. Performance metrics for machine learning models. *Radiology*, 3:1–7, 2021.

[5] Ralph Foorthuis. On the nature and types of anomalies: a review of deviations in data. *International journal of data science and analytics*, 12(4):297–331, 2021.

[6] Sahand Hariri, Matias Carrasco Kind, and Robert J Brunner. Extended isolation forest. *IEEE transactions on knowledge and data engineering*, 33(4):1479–1489, 2019.

[7] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE, 2008.

[8] Fokrul Alom Mazarbhuiya and Mohamed Shenify. A mixed clustering approach for real-time anomaly detection. *Applied Sciences*, 13(7):4151, 2023.

[9] Peter J Rousseeuw and Mia Hubert. Anomaly detection by robust statistics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(2):e1236, 2018.

[10] Sarita Tripathy and PL Sahoo. A survey of different methods of clustering for anomaly detection. *International Journal of Science and Engineering Research*, 6(1), 2015.

[11] Jiawei Yang, Xu Tan, and Sylwan Rahardja. Outlier detection: How to select k for k-nearest-neighbors-based outlier detectors. *Pattern Recognition Letters*, 174:112–117, 2023.

[12] Ming Zhao, Jingchao Chen, and Yang Li. A review of anomaly detection techniques based on nearest neighbor. In *2018 International Conference on Computer Modeling, Simulation and Algorithm (CMSA 2018)*, pages 290–292. Atlantis Press, 2018.