М.Ю. Чесноков

Поиск аномалий во временных рядах на основе ансамблей алгоритмов DBSCAN

Аннотация. В статье предложен метод построения ансамбля на основе алгоритма DBSCAN, использующий внутреннюю структуру временных рядов для адаптивного подбора входных параметров, экспериментально показывающий меньший разброс и высокие значения качества выявления аномалий на реальных и синтетических данных по сравнению с рядом популярных подходов.

Ключевые слова: поиск аномалий, временной ряд, ансамбль алгоритмов, DBSCAN.

Введение

Поиском аномалий (выбросов) называется проблема обнаружения объектов не похожих на остальные данные, не подчиняющихся ожидаемым, «нормальным» паттернам поведения [1]. Важность рассматриваемой задачи определяется тем, что аномалия несет в себе существенную информацию о состоянии системы, может быть сигналом к действию, например, к проведению дополнительной проверки [2]. Примерами практических задач обнаружения аномалий являются: выявление атак на информационные системы (Intrusion Detection), отклонений медицинских показателей пациентов (Medical Anomaly Detection), финансового мошенничества (Fraud Detection), неисправностей и повреждений устройств, моторов (Fault Detection) и пр. [1, 3, 4]. В перечисленных задачах существуют паттерны нормальных и аномальных областей данных, на основании которых требуется задать оценку аномальности каждого наблюдения. Она может определяться как непрерывная величина, бинарная метка и ранг наблюдения по оценке аномальности.

Существуют особенности задачи, усложняющие ее и затрудняющие применение готовых методов на разных постановках [1,4]:

а) корректное задание всевозможных областей «нормальности» и их точных границ;

б) изменение понятия «нормы» со временем;

в) определение аномалии и постановка задачи зависят от предметной области, типа данных (пространственные, последовательные, графы);

г) доступность разметки выбросов и нормальных объектов (сложно сделать автоматически, создание вручную требует временных и денежных затрат).

Существует множество исследований, посвященных проблеме обнаружения аномалий в рамках разнообразных научных и прикладных задач. При этом в большинстве практических приложений разметка нормальных и аномальных классов данных отсутствует, в связи с чем проблема выявления аномалий рассматривается как задача обучения без учителя [1]. Стоит отметить особенности методов поиска аномалий: их результаты работы существенно зависят от входных параметров [5, 6]; структуры данных, природы их происхождения [2, 6, 7]. В связи с отсутствием разметки становится затруднительным проводить отбор моделей и проверку качества их работы (с помощью кросс-валидации или тестировании на отложенной выборке) [6, 7]. На практике необходимо выявление аномалий среди большого разнородного набора данных, что в совокупности с вышеописанными причинами приводит к существенному разбросу качества и слабой практической применимости отдельных методов.

Изложенные проблемы можно решить за счет построения ансамблей алгоритмов – объединения результатов работы нескольких методов, совершающих разные ошибки, которые взаимно корректируются при их комбинации. Ансамбли улучшают качество, снижая зависимость модели от исследуемых данных и входных параметров, повышая стабильность результатов [2, 3, 6, 7]. В данной работе предлагается метод ансамблирования для поиска аномалий на основе алгоритма DBSCAN [8], использующий внутреннюю структуру временных рядов для адаптивного подбора параметров базового алгоритма, экспериментально показывающий уменьшение разброса результатов и повышение качества на реальных и синтетических данных по сравнению с рядом популярных подходов.

Основная часть статьи организованна следующим образом: в первом разделе рассматриваются подходы к обнаружению аномалий; второй раздел посвящен построению ансамблей для поиска аномалий; третий раздел – предложенному методу поиска выбросов во временных рядах на основе DBSCAN. В четвертом разделе обсуждаются результаты экспериментов и сравнения методов на реальных и синтетических данных.

1. Методы поиска аномалий

Обзоры и обсуждение алгоритмов обнаружения аномалий представлены в [1, 3], отдельно для последовательностей и временных рядов – в [4, 9]. Методы поиска аномалий для обучения без учителя можно категоризировать на основании используемых концепций: ближайшие соседи, кластеризация, статистические техники, классификация.

Методы на основе ближайших соседей используют предположение о том, что нормальные объекты расположены в «плотных» областях данных, в отличие от аномалий, лежащих в «разреженных» областях (подобные методы задают меру схожести/близости двух точек). Популярным методом данной группы является LOF (Local Outlier Factor) [10], использующий идею сравнения локальной плотности точки со средней локальной плотностью ее k-ближайших соседей.

Методы кластеризации выделяют аномалии как наблюдения, которые либо не принадлежат ни одному кластеру, либо принадлежат маленьким или разреженным кластерам. Популярным методом данной группы является DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [8] – плотностный алгоритм кластеризации, выделяющий «внутренние», «пограничные» и «шумовые» точки в данных (точки, не вошедшие ни в один кластер). В основе алгоритма лежит идея о том, что плотность точек внутри каждого кластера существенно выше плотности точек снаружи, в том числе выше плотности шумовых областей, т.е. окрестность радиуса ε любой внутренней точки каждого кластера должна включать более минимального количества точек MinPts – превышать заданный порог плотности.

Существует предположение, что нормальные данные возникают с высокой вероятностью в отличие от аномалий. Популярным методом данной группы является 3σ [11], предполагающий, что данные распределены нормально и аномалиями объявляются объекты, лежащие на расстоянии больше трех стандартных отклонений от среднего. Более устойчивая альтернатива по сравнению с вышеприведенным методом – медианное абсолютное отклонение от медианы (Median Absolute Deviation, MAD) – выборочная медиана гораздо менее чувствительна к выбросам, нежели выборочное среднее [12].

Некоторые методы классификации могут быть адаптированы для задачи поиска аномалий в предположении, что классификатор способен разделить классы аномальных и нормальных объектов. Примером является одноклассовый метод опорных векторов (One Class SVM, OCSVM) [13], выделяющий при обучении область, в которой расположены данные, а затем для каждого тестового объекта определяет попадает ли он в эту область или нет. Другой интересной техникой является изолирующий лес (Isolation Forest, IF), строящий ансамбль «изолирующих» деревьев для отделения каждого объекта от остальных [14]. Оценкой аномальности точки служит средняя глубина листьев, в которые она попала при разбиении – чем меньше (ближе к «корню»), тем более вероятно, что это аномалия.

Стоит отметить, что в задаче поиска аномалий на двух наборах данных конкретный алгоритм может давать различное качество работы – зависимость модели от данных [2, 6, 7]. Кроме этого методы обнаружения аномалий чувствительны к используемым входным параметрам (одно и то же количество соседей k в LOF может быть эффективно на одних данных и неэффективно на других) – зависимость качества от

входных параметров [5, 6]. На практике зачастую возникает необходимость обнаружения аномалий среди большого набора данных с разнообразной внутренней структурой, что в совокупности с вышеописанными факторами приводит к значительному разбросу результатов и слабой практической применимости отдельных методов. Конструирование ансамблей алгоритмов позволяет снижать зависимость модели от данных и входных параметров, повышать устойчивость итоговых результатов [5-7].

2. Ансамбли алгоритмов поиска аномалий

Детальные обзоры проблематики построения ансамблей алгоритмов для выявления аномалий представлены в [2, 5-7, 15]. Теоретические обоснования применения ансамблей для выявления аномалий в рамках используемого в задачах обучения с учителем понятия компромисса между смещением и разбросом (bias-variance tradeoff) рассмотрены в [5, 6]. В случае отсутствия разметки («истинных» оценок аномальности) смещение и разброс могут быть определены исходя из предположения, что истинные значения оценки аномальности существуют, но не наблюдаемы.

Согласно [6, 7] типичный процесс конструирования ансамбля состоит из трех этапов: 1) выбор базовых алгоритмов; 2) нормализация оценок аномальности; 3) комбинирование результатов работы детекторов для получения итоговых оценок. При построении ансамбля следует учитывать следующие точность каждой компоненты и разнородность компонент в ансамбле (детекторы должны совершать разные ошибки, которые могут быть скомпенсированы корректной работой других компонент при комбинировании). Очевидно, что точность отдельной компоненты ансамбля концептуально сложно измерить в рамках задачи поиска аномалий в силу отсутствия разметки, качество может оцениваться на основе самих данных [2, 15]. В случае наличия разметки популярным подходом к измерению качества выявления выбросов является площадь под ROC-кривой (AUC ROC) [2]. Для увеличения разнородности среди компонент ансамбля предложен ряд подходов: использование различных подпространств признаков объектов (feature bagging) [16]; жадного алгоритма отбора компонент на основе корреляции результатов их работы [15]; одного метода с различными входными параметрами [10]; техники сэмплирования данных (разбиение на подвыборки) [5, 16].

Очевидным шагом при построении ансамбля является определение количества компонент в нем (по эмпирическому правилу «чем больше разнородных точных компонент, тем лучше»), либо задание критериев прекращения добавления компонент. Перед комбинированием оценок аномальности базовых алгоритмов необходимо помнить об их нормализации, т.е. приведению к одному масштабу, позволяющему сравнить оценки одного объекта от разных алгоритмов. После нормализации следует выбрать агрегирующую функцию для комбинирования оценок: максимум, среднее арифметическое, усреднение обрезанного множества оценок (только k наибольших), усреднение нескольких максимумов подгрупп (AOM, average of maximum). Частыми вариантами являются максимум и среднее арифметическое (наименее рискованная стратегия, в наибольшей степени сохраняющая преимущество ансамблей – взаимную корректировку ошибок компонент [2, 5-7]). Выбор подходящей агрегирующей функции существенно зависит от данных задачи [5], вследствие чего нельзя заявлять о превосходстве одного из подходов комбинирования.

Обобщая вышеприведенные исследования, можно заключить, что процесс конструирования ансамблей алгоритмов для поиска аномалий состоит из следующих этапов: 1) выбор семейства алгоритмов (на основе ближайших соседей, кластеризации и т.п.), выбор базового алгоритма (LOF, DBSCAN и т.д.), определение параметров компонент; 2) определение количества компонент в ансамбле; 3) нормализация оценок аномальности; 4) комбинирование оценок разных компонент; 5) измерение итогового качества работы всего ансамбля.

3.Метод адаптивного подбора параметров компонент ансамбля алгоритмов DBSCAN

На практике необходимо проверять большое количество разнородных выборок, что приводит к сложности определения оптимального набора входных параметров для одного алгоритма, и тем более для ансамбля. Выбор семейства алгоритмов, базовых компонент и опти-

мальных параметров вручную может занять длительное время. Упрощением данных проблем может быть использование ансамблей из одного базового алгоритма (DBSCAN, LOF и т.п.) с различными входными параметрами.

3.1. Подбор параметров DBSCAN

Существует множество исследований, посвященных подбору оптимальных параметров алгоритма DBSCAN (ε и MinPts), подтверждающих сильную зависимость результатов выявления шумовых наблюдений от параметров, а также их сложность задания априори [17, 18]. Авторы DBSCAN в оригинальной статье [8] в рамках задачи кластеризации выбирали значения ε и MinPts, позволяющие выделить наименее плотный кластер (т.е. задать пороговую плотность шума).

В работе [17] для автоматического определения ε авторы предлагают подход на основе расчета всех попарных расстояний между объектами. Другой способ автоматического определения параметров представлен в [18], где применяется алгоритм дифференциальной эволюции для нахождения оптимального ε для каждого MinPts набора входных параметров DBSCAN.

Для выявления аномалий во временных рядах следует учитывать зависимость между наблюдениями [6]. В данной работе предлагается метод построения ансамблей DBSCAN на основе адаптивного подбора параметров компонент, использующий особенность одномерных рядов – в случае отсутствия выбросов ближайшие соседи любого наблюдения это предыдущий и последующий объекты во временной оси («окно» соседей). Вводим упрощения: ось времени будем рассматривать аналогично оси значений, меру близости зададим евклидовым расстоянием между наблюдениями (в двухмерном пространстве). Рассмотрим временной ряд

$S = \{s_1, s_2, \dots, s_N,\} = \{(x_1; t_1), (x_2; t_2), \dots, (x_N; t_N)\}$,

где x_i – значение, t_i – время в точке i. Рассмотрим среднее расстояние d_i от точки i до ее ближайших соседей внутри «окна» (MinPts соседей с одной и MinPts с другой стороны по временной оси от объекта):

$$d_i = \frac{1}{2\ MinPts} \sum_{\substack{j=i-MinPts \\ j \neq i}}^{i+MinPts} dist(s_i;\ s_j),$$

$$i \in (MinPts + 1, \dots, N - MinPts),$$

где $dist$ – евклидово расстояние между точками (предварительно масштабируем обе оси для корректного нахождения расстояний – делим на стандартное отклонение, вычитаем среднее).

Расстояние d_i будет приблизительно одинаково у нормальных и существенно больше у аномальных наблюдений (Рис. 1, Рис. 2). Стоит отметить, что в случае многомерных рядов возможна модификация алгоритма [19] – рассмотрение выбросов по отдельным группам осей.

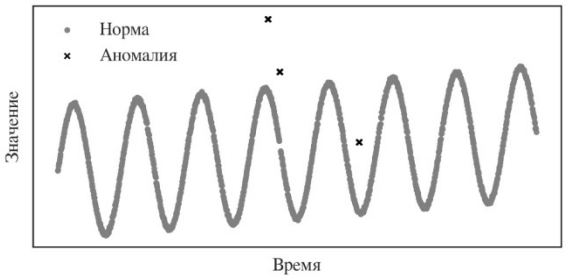

Рис. 1. Пример временного ряда с аномалиями

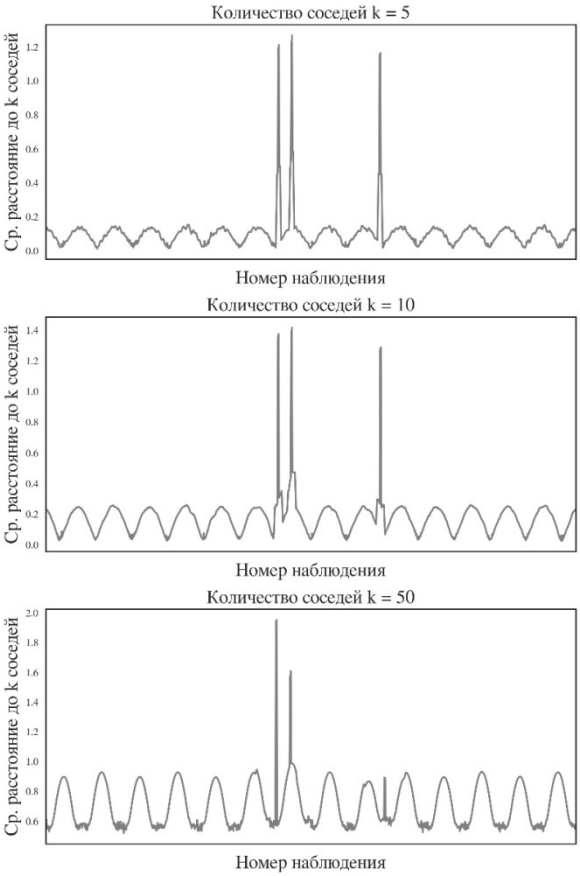

Рис. 2. Средние расстояния от наблюдения до k соседей

Для любого ряда S и заданного MinPts можно получить набор средних расстояний $D = \{d_{MinPts+1}, \dots, d_{N-MinPts}, \}$, отражающий внутреннюю структуру данных – целесообразно использовать в качестве параметра ε алгоритма DBSCAN. При этом произвольное ε при фиксированном MinPts не гарантирует определения релевантной структуры данных: чем больше ε, тем меньше точек будет отнесено к шуму, в пределе все точки будут отнесены к нормальным объектам [17].

3.2. Ансамблирование полученных компонент

Использование значений d_i из набора D в качестве параметра ε алгоритма DBSCAN позволяет выделять регионы нормальности и аномальности в данных, при этом для различных значений d_i получаются отличающиеся результаты. Вследствие чего целесообразно объединение подобных компонент в ансамбль. Количество соседей MinPts следует выбирать исходя из предположения о минимальном количестве подряд идущих аномалий в рядах.

Рассмотрим набор $M = \{MinPts_1, MinPts_2, \dots, MinPts_Q, \}$. Для каждого $MinPts_q$ получим набор средних расстояний D_q. Из D_q выберем P перцентилей набора расстояний в качестве параметров $\varepsilon_q = \{\varepsilon_{q1}, \varepsilon_{q2}, \dots, \varepsilon_{qP}\}$, получаем $Q * P$ пар входных параметров $\{MinPts_q, \varepsilon_{qp}\}$.

Далее для каждой пары запускаем DBSCAN и получаем бинарные метки аномалий $L^{qp} = \{l_1^{qp}, l_2^{qp}, \dots, l_N^{qp}\}, l_i \in \{0, 1\}$. Оценку аномальности a_i для наблюдения i получаем усреднением (наименее рискованная стратегия комбинирования) по $Q * P$ меткам:

$$a_i = \frac{1}{QP}\sum_{j=1}^{QP} l_i^j, \; i \in (0, \dots, N),$$

где $a_i \in (0; 1)$ интерпретируется как вероятность принадлежности наблюдения классу аномалий.

В результате получаем следующий метод **DBSCAN-a**:

Вход - временной ряд
$S = \{ (x_1; t_1), (x_2; t_2), \dots, (x_N; t_N)\}$,
набор $M = \{MinPts_1, \dots, MinPts_Q, \}$.
Выход-оценки аномальности
$a_i, \forall\, i \in (0, 1, \dots, N)$.

Шаг 1. Масштабируем ряд S по обоим осям.
Шаг 2. Для набора M рассчитываем расстояния $\{D_1, D_2, \dots, D_Q, \}$.
Шаг 3. Вычисляем $\varepsilon_q = \{\varepsilon_{q1}, \varepsilon_{q2}, \dots, \varepsilon_{qP}\}$ как P перцентилей D_q.
Шаг 4. Применяем DBSCAN для всех полученных пар $\{MinPts_q, \varepsilon_{qp}\}$, получаем $Q * P$ бинарных меток L^{qp} по каждому наблюдению.
Шаг 5. Рассчитываем оценку аномальности $a_i, \forall\, i \in (0, 1, \dots, N)$.

4. Эксперименты и сравнения

4.1. Данные, алгоритмы для сравнения

Следует отметить, что для задачи поиска аномалий во временных рядах существует мало публично доступных размеченных наборов данных [2]. Одним из таких примеров является набор рядов от YAHOO [20], состоящий из синтетических и реальных временных рядов с размеченными аномалиями (порядка 1400 наблюдений в каждом – Табл. 1). Реальные данные представляют собой обфусцированные показатели сервисов YAHOO (Рис. 3).

Сравнение предложенного метода (DBSCAN-a) произведено на каждом из временных рядов (всего 140 шт.) в режиме обучения без учителя со следующими ансамблями популярных алгоритмов: медианное отклонение от медианы (MAD); одноклассовый метод опорных векторов

Табл. 1. Описание данных

Тип данных	Кол-во рядов	Диапазон кол-ва точек	Диапазон кол-ва аномалий
Синтетические	100	1421 - 1421	1 - 9
Реальные	40	741 - 1461	1 - 227

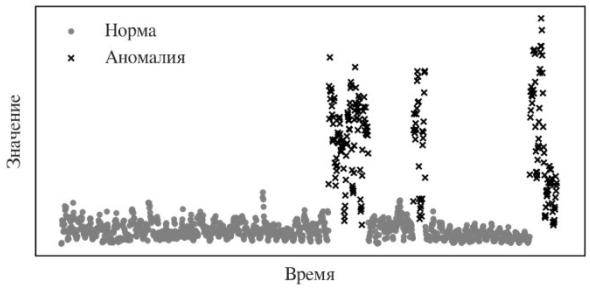

Рис. 3. Пример реальных данных

Табл. 2. Диапазоны входных параметров базовых алгоритмов

Базовый алгоритм	**Параметр**	**Диапазон значений**
MAD	Порог	1.0 - 3.1
OCSVM	Порог кол-ва ошибок	0.01 - 0.3
	Коэф. ядра	0.3 - 0.5
IF	Кол-во деревьев	{100, 150, 250}
	Доля выбросов	0.01 - 0.3
LOF	Кол-во соседей	{5, 10, 50}
	Доля выбросов	0.01 - 0.3
DBSCAN	ε	0.3 - 2.3
	MinPts	{5, 10, 50}
DBSCAN-a	MinPts	{5, 10, 50}
	Перцентили расстояний	{5, 10, 30, 50, 70, 90, 95}

(OCSVM); изолирующий лес (IF); фактор локальной аномальности (LOF); DBSCAN с фиксированным набором ε и MinPts. Использовались реализации алгоритмов из python библиотеки scikit-learn. Каждый ансамбль включал одинаковое количество базовых алгоритмов с разными входными параметрами (Табл. 2). Точность обнаружения аномалий для каждого ряда в экспериментах измерялось площадью под ROC кривой (AUC ROC), а разброс на множестве рядов интерквартильным размахом (IQR) и стандартным отклонением – AUC ROC.

4.2. Результаты экспериментов и обсуждение

Полученные результаты экспериментов отражены в Табл. 3 и Табл. 4. На Рис. 4 и Рис. 5 представлены графики («ящики с усами») результатов вычислений шести типов ансамблей на двух группах данных – реальных (40 шт.) и синтетических (100 шт.), каждое наблюдение – значение AUC ROC на одном из рядов. Как видно из графиков DBSCAN-a позволяет уменьшить разброс результатов по сравнению с другими ансамблями на реальных и синтетических

Табл. 3. Результаты на синтетических данных

Базовый алгоритм	**Средний AUC ROC**	**Стандартное отклонение**	**Медиана AUC ROC**	**IQR**
MAD	0.912	0.187	**1.0**	0.025
OCSVM	0.894	0.213	0.989	0.017
IF	0.896	0.222	0.996	0.007
LOF	0.943	0.148	0.997	0.030
DBSCAN	0.930	0.142	**1.0**	0.012
DBSCAN-a	**0.953**	**0.121**	**1.0**	**0.001**

Табл. 4. Результаты на реальных данных

Базовый алгоритм	**Средний AUC ROC**	**Стандартное отклонение**	**Медиана AUC ROC**	**IQR**
MAD	0.957	0.096	0.989	0.031
OCSVM	0.966	0.057	0.988	0.014
IF	0.975	**0.054**	0.995	0.007
LOF	0.927	0.100	0.971	0.108
DBSCAN	0.968	0.087	**0.999**	0.027
DBSCAN-a	**0.977**	0.055	**0.999**	**0.006**

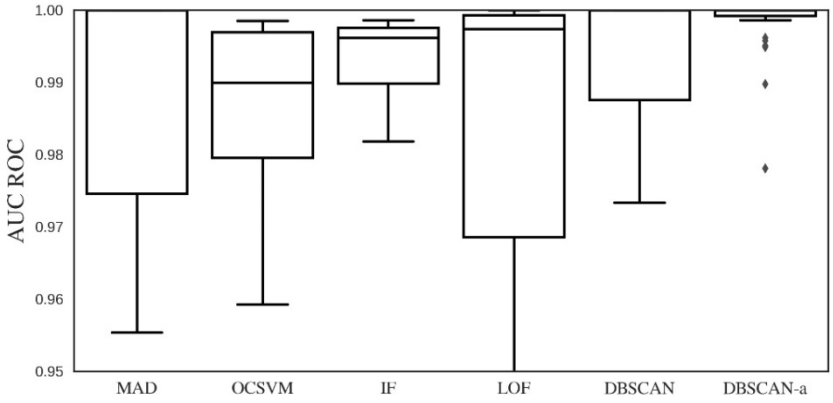

Рис. 4. Результаты на синтетических данных

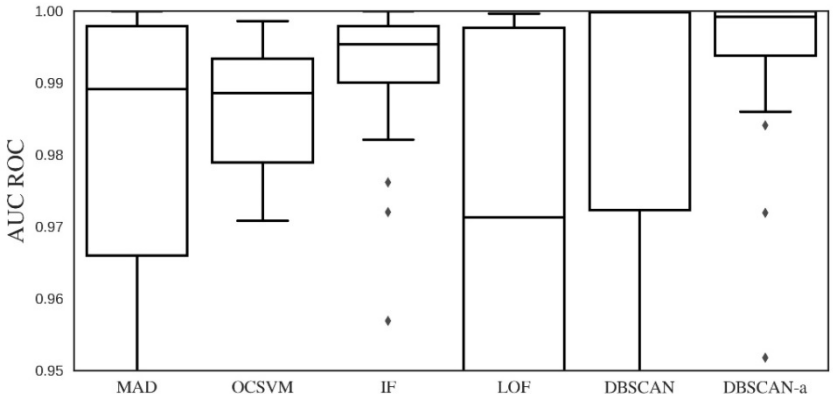

Рис. 5. Результаты на реальных данных

данных, что существенно в задаче обучения без учителя, когда в отсутствие разметки необходимо быть уверенным в хорошей работе метода без оптимизации параметров на множестве наборов данных. В случае измерения разброса стандартным отклонением на реальных данных чуть лучше оказывается IF, что произошло в силу чувствительности стандартного отклонения к наблюдениям, лежащим за границей усов. Стоит отметить, что стандартное отклонение на синтетических данных выше, чем на реальных, так как количество рядов в группе больше (100 шт. против 40 шт.), соответственно она более разнородна (графики для наглядности построены с 0.95 до 1.0, при этом существуют значения меньше 0.95). При этом медианы результатов AUC ROC выше на синтетических рядах, вследствие чего можно заключить, что они «проще», имеют более очевидные аномалии по сравнению с реальными данными. Следует отметить, что близость AUC ROC к 1 также объясняется чувствительностью метрики к несбалансированным выборкам (на несколько аномалий приходится тысяча нормальных наблюдений).

На практике часто возникает ситуация одновременной проверки большого количества временных рядов с различной внутренней структурой. В силу хороших экспериментальных результатов применения метода DBSCAN с адаптивным подбором параметров на большом наборе данных целесообразно его использование в соответствующих практических приложениях.

Заключение

В данной работе предложен метод поиска аномалий на основе ансамблей алгоритмов DBSCAN, использующих внутреннюю структуру временных рядов. Проведено экспериментальное сравнение предложенного метода на реальных и синтетических данных с ансамблями популярных техник: медианное отклонение от медианы (MAD), одноклассовый метод опорных векторов (OCSVM), изолирующий лес (IF), фактор локальной аномальности (LOF),

DBSCAN с фиксированными параметрами. Метод экспериментально показывает меньший разброс качества при одновременном применении на большом количестве временных рядов по сравнению с рассмотренными подходами. В дальнейшем планируется проведение исследований по следующим направлениям: автоматический выбор параметра MinPts; апробация других механизмов выбора алгоритмов и отбора компонент; использование альтернативных способов комбинирования компонент.

Литература

1. Chandola V., Banerjee A., Kumar V. Anomaly detection: A survey //ACM computing surveys (CSUR). – 2009. – Т. 41.
2. Zimek A., Campello R. J. G. B., Sander J. Ensembles for unsupervised outlier detection: challenges and research questions a position paper //Acm Sigkdd Explorations Newsletter. – 2014. – Т. 15. – №1. – С. 11-22.
3. Aggarwal C. C. Outlier Analysis Second Edition. – 2016.
4. Gupta M. et al. Outlier detection for temporal data: A survey //IEEE Transactions on Knowledge and Data Engineering. – 2014. – Т. 26. – № 9. – С. 2250-2267.
5. Aggarwal C. C., Sathe S. Theoretical foundations and algorithms for outlier ensembles //ACM SIGKDD Explorations Newsletter. – 2015. – Т. 17. – №1. – С. 24-47.
6. Aggarwal C. C., Sathe S. Outlier Ensembles: An Introduction. – Springer, 2017.
7. Aggarwal C. C. Outlier ensembles: position paper //ACM SIGKDD Explorations Newsletter. – 2013. – Т. 14. – № 2. – С. 49-58.
8. Ester M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise //Kdd. – 1996. – Т. 96. – №. 34. – С. 226-231.
9. Chandola V., Banerjee A., Kumar V. Anomaly detection for discrete sequences: A survey //IEEE Transactions on Knowledge and Data Engineering. – 2012. – Т. 24. – №5. – С. 823-839.
10. Breunig M. M. et al. LOF: identifying density-based local outliers //ACM sigmod record. – ACM.– 2000. – Т. 29. – № 2. – С. 93-104.
11. Shewhart W. A. Economic control of quality of manufactured product. – ASQ Quality Press. –1931.
12. Leys C. et al. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median //Journal of Experimental Social Psychology. – 2013. – Т. 49. – № 4. – С. 764-766.
13. Vapnik V. The nature of statistical learning theory. – Springer science & business media, 2013.
14. Liu F. T., Ting K. M., Zhou Z. H. Isolation forest //Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on. – IEEE.– 2008. – С. 413-422.
15. Schubert E. et al. On evaluation of outlier rankings and outlier scores //Proceedings of the 2012 SIAM International Conference on Data Mining. – Society for Industrial and Applied Mathematics. – 2012. – С. 1047-1058.
16. Lazarevic A., Kumar V. Feature bagging for outlier detection //Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. – ACM.– 2005. – С. 157-166.
17. Zhou H., Wang P., Li H. Research on adaptive parameter determination in DBSCAN algorithm [J] //Journal of Xi'an University of Technology. – 2012. – Т. 28. – № 3. – С. 289-292.
18. Karami A., Johansson R. Choosing dbscan parameters automatically using differential evolution //International Journal of Computer Applications. – 2014. – Т. 91. – № 7.
19. Kut A., Birant D. Spatio-temporal outlier detection in large databases //CIT. Journal of computing and information technology. – 2006. – Т. 14. – № 4. – С. 291-297.
20. Yahoo! labs. Webscope dataset ydata-labeled-time-series-anomalies-v1_0 [Online]. URL: http://webscope.sandbox.yahoo.com/ (дата обращения 25.08.2017).

Чесноков Михаил Юрьевич. Аспирант Московского физико-технического института (Государственный университет, МФТИ). Окончил МФТИ в 2015 году. Область научных интересов: информационные технологии, анализ данных, машинное обучение. E-mail: mikhail.chesnokov@phystech.edu

Time series anomaly detection based on DBSCAN ensembles

M.Y. Chesnokov

The quality of anomaly detection algorithms highly depends on the input parameters and internal structure of dataset, in addition this problem usually occurred in unsupervised setting leading to the conceptual complexity of quality measurement. In practice there is a significant variance of results of anomaly detection due to the huge amount of datasets under consideration having diverse internal structure. Outlier ensemble is a kind of technique which can improve the variance of anomaly detection and increase the overall quality of identification. In this paper we investigate the problem of anomaly detection in time series in unsupervised setting, propose the method of outlier ensemble construction based on DBSCAN algorithm, which uses the time series internal structure for adaptive input parameters selection. Experiments on synthetic and real datasets show the decrease of variance and high quality of method compared to popular techniques such as Median Absolute Deviation, One Class SVM, Isolation Forest, Local Outlier Factor and simple DBSCAN.

Keywords: anomaly detection, time series, unsupervised ensembles, DBSCAN.

References

1. Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), 15.
2. Zimek, A., Campello, R. J., & Sander, J. (2014). Ensembles for unsupervised outlier detection: challenges and research questions a position paper. Acm Sigkdd Explorations Newsletter, 15(1), 11-22.
3. Aggarwal, C. C. (2016). Outlier Analysis Second Edition.
4. Gupta, M., Gao, J., Aggarwal, C. C., & Han, J. (2014). Outlier detection for temporal data: A survey. IEEE Transactions on Knowledge and Data Engineering, 26(9), 2250-2267.
5. Aggarwal, C. C., & Sathe, S. (2015). Theoretical foundations and algorithms for outlier ensembles. ACM SIGKDD Explorations Newsletter, 17(1), 24-47.
6. Aggarwal, C. C., & Sathe, S. (2017). Outlier Ensembles: An Introduction. Springer.
7. Aggarwal, C. C. (2013). Outlier ensembles: position paper. ACM SIGKDD Explorations Newsletter, 14(2), 49-58.
8. Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In Kdd (Vol. 96, No. 34, pp. 226-231).
9. Chandola, V., Banerjee, A., & Kumar, V. (2012). Anomaly detection for discrete sequences: A survey. IEEE Transactions on Knowledge and Data Engineering, 24(5), 823-839.
10. Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000, May). LOF: identifying density-based local outliers. In ACM sigmod record (Vol. 29, No. 2, pp. 93-104). ACM.
11. Shewhart, W. A. (1931). Economic control of quality of manufactured product. ASQ Quality Press.
12. Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. Journal of Experimental Social Psychology, 49(4), 764-766.
13. Vapnik, V. (2013). The nature of statistical learning theory. Springer science & business media.
14. Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008, December). Isolation forest. In Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on (pp. 413-422). IEEE.
15. Schubert, E., Wojdanowski, R., Zimek, A., & Kriegel, H. P. (2012, April). On evaluation of outlier rankings and outlier scores. In Proceedings of the 2012 SIAM International Conference on Data Mining (pp. 1047-1058). Society for Industrial and Applied Mathematics.
16. Lazarevic, A., & Kumar, V. (2005, August). Feature bagging for outlier detection. In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining (pp. 157-166). ACM.
17. Zhou, H., Wang, P., & Li, H. (2012). Research on adaptive parameter determination in DBSCAN algorithm [J]. Journal of Xi'an University of Technology, 28(3), 289-292.
18. Karami, A., & Johansson, R. (2014). Choosing DBSCAN parameters automatically using differential evolution. International Journal of Computer Applications, 91(7).
19. Kut, A., & Birant, D. (2006). Spatio-temporal outlier detection in large databases. CIT. Journal of computing and information technology, 14(4), 291-297.
20. Yahoo! labs. Webscope dataset ydata-labeled-time-series-anomalies-v1_0 [Online]. Available at: http://webscope.sandbox.yahoo.com/ (accessed August 25, 2017).

Chesnokov M.Y. Department of theoretical and applied problems of innovation, Moscow Institute of Physics and Technology (State University), Moscow, Russia, e-mail: mikhail.chesnokov@phystech.edu