Обзор методов обнаружения аномалий в потоках данных

В.П. Шкодырев, К.И. Ягафаров, В.А. Баштовенко, Е.Э. Ильина

***Аннотация*— Данная статья посвящена исследованию различных подходов к идентификации аномалий во временных рядах, которая заключается в обнаружении и обработке отклонений в потоках данных, получаемых во время проведения технологических процессов. Выявление аномалий в поведении системы позволяет не только повысить качество таких процессов, но и предотвращать нештатные ситуации и аварии на ранних этапах. Все это указывает на актуальность проведения исследований в данной области.**

В работе приведен обзор существующих методов и алгоритмов обнаружения аномалий с целью структуризации имеющихся данных и последующего отбора средств для разработки системы идентификации аномалий в потоках больших данных.

***Ключевые слова* — Поиск аномалий, Анализ данных, Потоки данных**

I. Введение

Интеллектуальный анализ данных, называемый также Data mining, используется для выделения новой значимой информации из большого объема данных. В условиях постоянного увеличения этих объемов, а также возрастающей значимости результатов их анализа вопрос идентификации имеющихся в них аномалий стоит особенно остро. Результаты анализа без предварительного исключения аномальных экземпляров данных могут быть значительно искажены.

Обнаружение аномалий относится к поиску непредвиденных значений (паттернов) в потоках данных. Аномалия (выброс, ошибка, отклонение или исключение) – это отклонение поведения системы от стандартного (ожидаемого). В данной статье эти термины являются эквивалентными. Они могут возникать в данных самой различной природы и структуры в результате технических сбоев, аварий, преднамеренных взломов и т.д. В настоящее время разработано множество методов и алгоритмов поиска аномалий для различных типов данных. Целью данной статьи является обзор наиболее универсальных из них.

II. Виды аномалий

Аномалии в данных могут быть отнесены к одному из трех основных типов [3].

Точечные аномалии возникают в ситуации, когда отдельный экземпляр данных может рассматриваться как аномальный по отношению к остальным данным. На рисунке 1а экземпляр A1, а также группа экземпляров A2 являются аномальными при нормальных экземплярах в группах C1 и C2. Данный вид аномалий является наиболее легко распознаваемым, большинство существующих методов создано для распознавания точечных аномалий.

Контекстуальные аномалии наблюдаются, если экземпляр данных является аномальным лишь в определенном контексте, (данный вид аномалий также называется условным). Для определения аномалий этого типа основным является выделение контекстуальных и поведенческих атрибутов.

- Контекстуальные атрибуты используются для определения контекста (или окружения) для каждого экземпляра. Во временных рядах контекстуальным

Рис.1 Виды аномалий

атрибутом является время, которое определяет положение экземпляра в целой последовательности. Контекстуальным атрибутом также может быть положение в пространстве или более сложные комбинации свойств.

- Поведенческие атрибуты определяют не контекстуальные характеристики, относящиеся к конкретному экземпляру данных.

Аномальное поведение определяется посредством значений поведенческих атрибутов исходя из конкретного контекста. Таким образом, экземпляр данных может быть контекстуальной аномалией при данных условиях, но при таких же поведенческих атрибутах считаться нормальным в другом контексте. Так, на рисунке 1б в точке A наблюдается аномалия, в отличие от точек N1 – N5, имеющих аналогичное значение. При обнаружении контекстуальных аномалий это свойство является ключевым в разделении контекстуальных и поведенческих атрибутов.

Коллективные аномалии возникают, когда последовательность связанных экземпляров данных (например, участок временного ряда) является аномальной по отношению к целому набору данных. Отдельный экземпляр данных в такой последовательности может не являться отклонением, однако совместное появление таких экземпляров является коллективной аномалией. На рисунке 1в участок A является коллективной аномалией.

Кроме того, в то время как точечные или контекстуальные аномалии могут наблюдаться в любом наборе данных, коллективные наблюдаются только в тех, где данные связаны между собой.

Стоит так же отметить, что точечные или коллективные аномалии могут в то же время являться и контекстуальными.

III. Методы обнаружения точечных аномалий

Существует несколько вариантов классификации существующих методик поиска аномалий [3]. В данной работе буду рассмотрены два вида деления: по режиму распознавания и по способу реализации.

В зависимости от применяемого алгоритма результатом работы системы идентификации аномалий может быть либо метка экземпляра данных как аномального, либо оценка степени вероятности того, что экземпляр является аномальным.

Процесс выявления аномалий может проводиться для данных различного формата:

- поток данных (работа в реальном времени);
- архив данных.

A. *Режимы распознавания аномалий*

Часто для решения задачи поиска аномалий требуется набор данных, описывающих систему. Каждый экземпляр в нем описывается меткой, указывающей, является ли он нормальным или аномальным. Таким образом, множество экземпляров с одинаковой меткой формируют соответствующий класс.

Создание подобной промаркированной выборки обычно проводится вручную и является трудоемким и дорогостоящим процессом. В некоторых случаях получить экземпляры аномального класса невозможно в силу отсутствия данных о возможных отклонениях в системе, в других могут отсутствовать метки обоих классов. В зависимости от того, какие классы данных используются для реализации алгоритма, методы поиска аномалий могут выполняться в одном из трех перечисленных ниже режимов:

1) Supervised anomaly detection (режим распознавания с учителем)

Данная методика требует наличия обучающей выборки, полноценно представляющей систему и включающей экземпляры данных нормального и аномального классов. Работа алгоритма происходит в два этапа: обучение и распознавание. На первом этапе строится модель, с которой в последствие сравниваются экземпляры, не имеющие метки. В большинстве случаев предполагается, что данные не меняют свои статистические характеристики, иначе возникает необходимость изменять классификатор [11].

Основной сложностью алгоритмов, работающих в режиме распознавания с учителем, является формирование данных для обучения. Часто аномальный класс представлен значительно меньшим числом экземпляров, чем нормальный, что может приводить к неточностям в полученной модели. В таких случаях применяется искусственная генерация аномалий.

2) Semi-Supervised anomaly detection (режим распознавания частично с учителем)

Исходные данные при этом подходе представляют только нормальный класс. Обучившись на одном классе, система может определять принадлежность новых данных к нему, таким образом, определяя противоположный.

Алгоритмы, работающие в режиме распознавания частично с учителем, не требуют информации об аномальном классе экземпляров, вследствие чего они шире применимы и позволяют распознавать отклонения в отсутствие заранее определенной информации о них.

3) Unsupervised anomaly detection (режим распознавания без учителя)

Применяется при отсутствии априорной информации о данных. Алгоритмы распознавания в режиме без учителя базируются на предположении о том, что аномальные экземпляры встречаются гораздо реже нормальных. Данные обрабатываются, наиболее отдаленные определяются как аномалии. Для применения этой методики должен быть доступен весь набор данных, т.е. она не может применяться в режиме реального времени.

B. *Методы распознавания аномалий*

1) Классификация

Реализация данного метода основана на предположении о том, что нормальное поведение системы может определяться одним или несколькими классами. Таким образом, экземпляр, не принадлежащий ни к одному из

классов, является отклонением. Поиск аномалий проходит в два этапа: обучение и распознавание. Классификатор обучается на массиве маркированных данных, далее определяется принадлежность к одному из известных классов. В противном случае экземпляр помечается, как аномалия.

Наиболее широко применяемыми механизмами реализации распознавания аномалий с помощью классификации являются: нейронные сети, Байесовы сети, метод опорных векторов и метод на основе правил.

• Метод обнаружения аномалий на основе нейронных сетей включает два этапа. Первый: нейронная сеть обучается распознаванию классов нормального поведения на тренировочной выборке. Второй: каждый экземпляр поступает в качестве входного сигнала нейронной сети. Система, основанная на нейронных сетях, может распознавать как один, так и несколько классов нормального поведения.

Для нахождения аномалий посредством распознавания только одного класса используются репликативные нейронные сети [7]. Получившая широкое распространение технология нейронных сетей глубинного обучения (Deep Learning) также успешно применяется для решения данной задачи [31].

• Байесовской сетью является графическая модель, отображающая вероятностные зависимости множества переменных и позволяющая проводить вероятностный вывод с помощью этих переменных. Она состоит из двух основных частей: графическая структура, которая определяет набор зависимостей и независимостей во множестве случайных величин, представляющих субъекты предметной области, и набор вероятностных распределений, определяющих силу отношений зависимости, закодированных в графической структуре. Таким образом, применение Байесовской сети при идентификации аномалий заключается в оценке вероятности наблюдения одного из нормальных или аномальных классов.

Наиболее простой реализацией данного подхода является Наивный байесовский подход (Naive Bayes Approach). Описание алгоритма его работы приведено в [37].

• Метод опорных векторов (Support Vector Machine) применяется для поиска аномалий в системах, где нормальное поведение представляется только одним классом. Данный метод определяет границу региона, в котором находятся экземпляры нормальных данных. Для каждого исследуемого экземпляра определяется, находится ли он в определенном регионе. Если экземпляр оказывается вне региона, он определяется как аномальный. Описание работы метода опорных векторов приведено в [37].

• Последний метод основывается на генерации правил, которые соответствуют нормальному поведению системы. Экземпляр, который не соответствует этим правилам, распознается как аномальный. Алгоритм состоит из двух шагов. Первый: обучение правил из выборки с помощью одного из алгоритмов, таких как RIPPER, Decision Trees и т.д. Каждому правилу присваивается свое значение, которое пропорционально соотношению между числом обучающих экземпляров, классифицируемых, как правило, и общим числом обучающих экземпляров, покрываемых этим правилом. Второй шаг: поиск для каждого тестируемого экземпляра правила, которое наилучшим образом подходит к данному экземпляру. Система может распознавать как один, так и несколько классов поведения

Одним из подвидов систем на основе правил являются системы нечеткой логики. Они применяется, когда граница между нормальным и аномальным поведением системы является размытой. Каждый экземпляр является аномалией в некоторой степени удаленности от центра масс нормального интервала. Описание применения данного подхода к задаче поиска аномалий приведено в [40].

2) Кластеризация

Данная методика предполагает группировку похожих экземпляров в кластеры и не требует знаний о свойствах возможных отклонений. Выявление аномалий может строиться на следующем предположении:

– Нормальные экземпляры данных относятся к кластеру данных, в то время как аномалии не принадлежат ни к одному из кластеров.

Однако при такой формулировке может возникнуть проблема определения точных границ кластеров. Отсюда следует другое предположение:

– Нормальные данные ближе к центру кластера, а аномальные – значительно дальше.

В случае, когда аномальные экземпляры не являются единичными, они также могут образовывать кластеры. Таким образом, их выявление строится на следующем предположении:

– Нормальные данные образуют большие плотные кластеры, а аномальные – маленькие и разрозненные.

Одной из простейших реализаций подхода на основе кластеризации является алгоритм k-means, описанный в работе [39]. Методология применения подхода на основе кластеризации также приведена в [25].

3) Статистический анализ

При использовании этого подхода исследуется процесс, строится его профиль (модель), который затем сравнивается с реальным поведением. Если разница в реальном и предполагаем поведении системы, определяемая заданной функцией аномальности, выше установленного порога, делается вывод о наличии отклонений. Применяется предположение том, что нормальное поведение системы будет находиться в зоне высокой вероятности, в то время как выбросы – в зоне низкой.

Данный класс методов удобен тем, что не требует заранее определенных знаний о виде аномалии. Однако сложности могут возникать в определении точного статистического распределения и порога [2].

Методы статистического анализа подразделяются на

две основные группы:

• Параметрические методы. Предполагают, что нормальные данные генерируются параметрическим распределением с параметрами θ и функцией плотности вероятности P(x, θ), где x – наблюдение. Аномалия является обратной функцией распределения. Эти методы часто основываются на Гауссовой или регрессионной модели, а также их комбинации. Подробное описание параметрических методов приведено в [30].

• Не параметрические методы. Предполагается, что структура модели не определена априорно, вместо этого она определяется из предоставленных данных. Включает методы на основе гистограмм или функций ядра.

Базовый алгоритм поиска аномалий с применением гистограмм включает два этапа. На первом этапе происходит построение гистограммы на основе различных значений выбранной характеристики для экземпляров тренировочных данных. На втором этапе для каждого из исследуемых экземпляров определяется принадлежность к одному из столбцов гистограммы. Не принадлежащие ни к одному из столбцов экземпляры помечаются как аномальные. Подробный алгоритм, основанный на применении гистограмм, описан в [13].

Распознавание аномалий на основе функции ядра происходит аналогично параметрическим методам за исключением способа оценки плотности вероятности. Сравнение результатов работы данного метода с параметрическим методом на основе Гауссовой модели приведено в [16].

4) Алгоритм ближайшего соседа

Для использования данной методики необходимо определить понятие расстояния (меры похожести) между объектами. Примером может быть Евклидово расстояние.

Два основных подхода основываются на следующих предположениях:

• Расстояние до k-го ближайшего соседа. Для реализации этого подхода расстояние до ближайшего объекта определяется для каждого тестируемого экземпляра класса. Экземпляр, являющийся выбросом, наиболее отдален от ближайшего соседа.

• Использование относительной плотности основано на оценке плотности окрестности каждого экземпляра данных. Экземпляр, который находится в окрестности с низкой плотностью, оценивается как аномальный, в то время как экземпляр в окрестности с высокой плотностью оценивается как нормальный. Для данного экземпляра данных расстояние до его k-го ближайшего соседа эквивалентно радиусу гиперсферы с центром в данном экземпляре и содержащей k остальных экземпляров.

5) Спектральные методы

Спектральные методы находят аппроксимацию данных, используя комбинацию атрибутов, которые передают большую часть вариативности в данных.

Эта методика основана на следующем предположении: данные могут быть вложены в подпространство меньшей размерности, в котором нормальное состояние и аномалии проявляются иначе. Спектральные методы часто применяются совместно с другими алгоритмами для предобработки данных.

Исследование модификаций спектрального метода приведено в [24].

6) Гибридные методы

Гибридные методики распознавания аномалий, позволяют сочетать преимущества различных подходов. При этом различные техники могут применяться как последовательно, так и параллельно для достижения усредненных результатов.

Примерами гибридных систем распознавания аномалий могут служить следующие исследования:

• Совмещение кластеризации и алгоритма ближайшего соседа в работе [20].

• Параллельное использование совмещенных алгоритмов Байесовых сетей и решающих деревьев, а также алгоритма ближайшего соседа с классификацией на основе правил в работе [22].

• Совмещение метода опорных векторов и нейронной сети глубинного обучения в работе [6].

Обзор публикаций с описанием конкретных алгоритмов, реализующих рассмотренные выше методы, приведен в таблице 1.

Сравнительный анализ методов приведен в таблице 2.

ТАБЛИЦА I. ОБЗОР ПУБЛИКАЦИЙ

Метод	**Публикации**
Классификация на основе репликационных нейронных сетей	Dau, Ciesielski [2014]
Классификация на основе нейроных сетей глубинного обучения	Xu et al. [2015], Yan et al. [2015]
Классифкация на основе Байесовых сетей	Hill et al. [2009], Heard [2010]
Классификация на основе правил	Li et al.[2007], Nasr et al. [2016],
Классификация на основе систем нечеткой логики	Ghosh et al. [2017]
Классификация на основе метода опорных векторов	Amer et al. [2013], Zhang et al. [2015]
Кластеризация	Portnoy et al.[2001], Кокорева с соавт.[2015], Kiss et al. [2014]
Параметрические методы статистического анализа	Thatte et al. [2010]
Статистический анализ на основе гистограмм	Kind et al. [2009]
Статистический анализ на основе функции ядра	Latecki et al. [2007], Zhang et al. [2015], Sharma et al. [2016]
Алгоритм ближайшго соседа	Liao, Vemuri [2002], Su [2011]
Спектральные методы	Денисова, Мясников [2014],

ТАБЛИЦА II. СРАВНЕНИЕ МЕТОДОВ

Метод	Результат	Режим распознавания	Определение класса аномалий	Работа без предварительного обучения
Классификация	Метка	Supervised, semi-supervised	Да	Нет
Кластеризация	Метка	Unsupervised, semi-supervised	Нет	Нет
Статистический анализ	Степень	Semi-supervised	Нет	Нет
Алгоритм ближайшего соседа	Степень	Unsupervised	Нет	Да
Спектральные методы	Метка	Unsupervised, Semi-supervised	Нет	Да

C. *Распознавание аномалий в потоках данных*

Выявление аномалий в режиме реального времени может потребовать дополнительной модификации методов. Наиболее простым в реализации является алгоритм скользящего окна.

Данная методика используется для временных рядов, которые разбивается на некоторое число подпоследовательностей – окон (рис.2). Необходимо выбрать окно фиксированной длины, меньшей чем длина самого ряда, чтобы захватить аномалию в процессе скольжения. Поиск аномальной подпоследовательности осуществляется при помощи скольжения окна по всему ряду с шагом, меньшим длины окна.

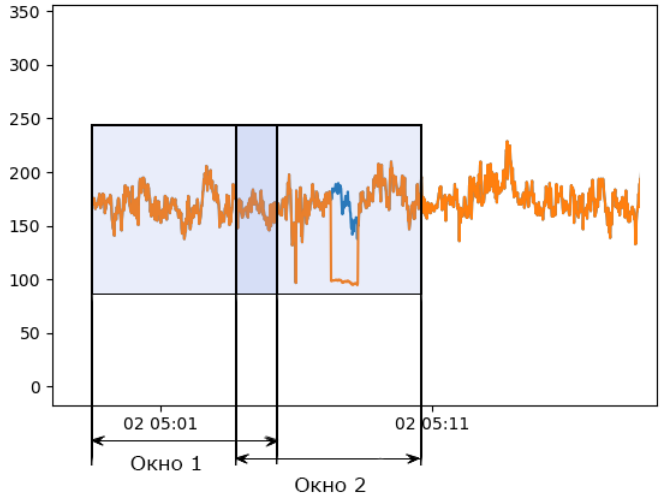

Рис.2 Применение алгоритма скользящего окна

Стоит отметить, что для методов, требующих наличия всего объема данных (функционирующих в режиме распознавания без учителя) применение данной техники может привести к повышенной неточности результатов, так как вычисления будут проводиться только для экземпляров в пределах окна.

В случае применения алгоритмов, основанных на предварительном построении модели с помощью классификации, существенных модификаций системы не требуется, поскольку на этапе распознавания каждый экземпляр обрабатывается отдельно.

Варианты построения систем распознавания аномалий в потоках данных приведены в [17], [28], [34].

IV. Способы выявления контекстуальных и коллективных аномалий

Все описанные выше методики применяются для поиска точечных аномалий, однако они также могут быть применены для распознавания коллективных и контекстуальных аномалий, при сведении их к точечным.

При поиске контекстуальных аномалий данный подход реализуется с помощью определения контекстуальных атрибутов и преобразования данных на их основе [8]. После этого к преобразованным данным можно применить один из методов идентификации точечных аномалий. Альтернативой данному методу является моделирование временных рядов на основе авторегрессии (например, построение модели ARIMA) [23] или преобразование их к символьным последовательностям [26], [35].

При поиске коллективных аномалий возможно определение подпоследовательностей фиксированной длины, как единичных объектов, однако при этом делается предположение, что все участки, являющиеся коллективными аномалиями, имеют одинаковую длину.

Описание комплексной методики выявления коллективных контекстуальных аномалий приведено в исследовании [12].

V. Заключение

Данная работа посвящена рассмотрению видов аномалий в потоках данных, а также обзору существующих методов и подходов к их поиску. Были проведены классификация и сравнение наиболее распространенных групп методов по основным критериям, приведены краткие описания алгоритмов. Кроме того, был осуществлен обзор конкретных реализаций и модификаций данных методов в публикациях последних лет.

Литература

[1] S. Agrawal, J. Agrawal, “Survey on Anomaly Detection using Data Mining Techniques”, Procedia Computer Science, vol. 60, 2015, pp. 708-713.

[2] M. Amer, M. Goldstein, S. Abdennadher, “Enhancing one-class support vector machines for unsupervised anomaly detection”, CM SIGKDD Workshop on Outlier Detection and Description, 2013, pp. 8-15

[3] V. Chandola, A. Banerjee, V. Kumar, “Anomaly detection: A survey”, ACM Computing Surveys, vol. 41(3), 2009, pp. 1–58.

[4] H. Dau, V. Ciesielski, A. Song, “Anomaly Detection Using Replicator Neural Networks Trained on Examples of One Class”, Simulated Evolution and Learning. Lecture Notes in Computer Science, vol 8886, 2014.

[5] S. Ghosh, A. Pal, A. Nag, S. Sadhu and R. Pati, “Network anomaly detection using a fuzzy rule-based classifier”, Computer, Communication and Electrical Technology, 2017 , pp. 61 -65.

[6] S. Erfani, Sarah, M. Baktashmotlagh, S. Rajasegarar, S. Karunasekera and C. Leckie, “A randomised nonlinear approach to large-scale

anomaly detection", 29th AAAI Conference on Artificial Intelligence, 2015, pp. 25–30, Hyatt Regency in Austin, Texas.
[7] S. Hawkins, H. He, G. J. Williams and R. A. Baxter, "Outlier detection using replicator neural networks", 4th International Conference on Data Warehousing and Knowledge Discovery. Springer-Verlag, 2002, pp. 170 – 180.
[8] M. Hayes, M.Capretz, "Contextual anomaly detection framework for big sensor data", Journal of Big Data, vol. 2(2), 2015.
[9] N.A. Heard, D.J. Weston, K. Platanioti, D.J. Hand, "Bayesian anomaly detection methods for social networks", Ann. Appl. Stat. 4 vol. 2, 2010, pp. 645 – 662.
[10] D.J. Hill, B. S. Minsker, and E. Amir, "Real-time Bayesian anomaly detection in streaming environmental data", Water Resour. Res., 45, 2009.
[11] H. Huang, "Rank Based Anomaly Detection Algorithms" Electrical Engineering and Computer Science – Dissertations, 2013, 331.
[12] Y. Jiang, C. Zeng, J. Xu and T. Li. "Real time contextual collective anomaly detection over multiple data streams", 2014.
[13] A. Kind, M. P. Stoecklin and X. Dimitropoulos, "Histogram-based traffic anomaly detection," IEEE Transactions on Network and Service Management, vol. 6(2), 2009, pp. 110-121.
[14] I. Kiss, B. Genge, P. Haller, G. Sebestyén, "Data clustering-based anomaly detection in industrial control systems", IEEE 10th International Conference on Intelligent Computer Communication and Processing (ICCP), 2014, pp. 275-281.
[15] L.J. Latecki, A. Lazarevic, D. Pokrajac, "Outlier Detection with Kernel Density Functions", Machine Learning and Data Mining in Pattern Recognition. Lecture Notes in Computer Science, vol 4571. Springer, Berlin, Heidelberg, 2007.
[16] R. Laxhammar, G. Falkman and E. Sviestins, "Anomaly detection in sea traffic - A comparison of the Gaussian Mixture Model and the Kernel Density Estimator," 12th International Conference on Information Fusion, Seattle, WA, 2009, pp. 756-763.
[17] W. Lee, S. J. Stolfo, P. K. Chan, E. Eskin, W. Fan, et al., "Real Time Data Mining-based Intrusion Detection", DARPA Information Survivability Conference & Exposition II, vol. 1, 2001.
[18] X. Li , J. Han , S. Kim , H. Gonzalez, "Roam: Rule- and motif-based anomaly detection in massive moving object data sets", 7th SIAM International Conference on Data Mining, 2007.
[19] Y. Liao, , V.R, Vemuri, "Use of K-Nearest Neighbor classifier for intrusion detection", Computers & Security, vol. 21(5), 2002, pp. 439-448
[20] W-C. Lin, S-W. Ke, C-F. Tsai, "An intrusion detection system based on combining cluster centers and nearest neighbors", Knowledge-Based Systems, vol. 78, 2015, pp. 13-21.
[21] A. A. Nasr, M. Z. Abdulmaged, "A Learnable Anomaly Detection System using Attributional Rules", International Journal of Computer Network and Information Security, vol. 8(11), 2016.
[22] M. Panda, A. Abraham, M. Patra, "Hybrid intelligent systems for detecting network intrusions". Security Comm. Networks, vol. 8, 2012, pp. 2741–2749
[23] E. H. M. Pena, M. V. O. de Assis and M. L. Proença, "Anomaly Detection Using Forecasting Methods ARIMA and HWDS," *2013* 32nd International Conference of the Chilean Computer Science Society (SCCC), Temuco, 2013, pp. 63-66.
[24] J. Piñeyro, A. Klempnow, V. Lescano, "Effectiveness of new spectral tools in the anomaly detection of rolling element bearings", Journal of Alloys and Compounds, vol. 310(1–2), 2000, pp. 276-279.
[25] L. Portnoy, E. Eskin, S. J. Stolfo, "Intrusion Detection with Unlabeled Data Using Clustering", Columbia University, New York, 2001.
[26] S. Sarkar, K. G Lore, S. Sarkar, V. Ramanan, S. R Chakravarthyet al., "Early Detection of Combustion Instability from Hi-speed Flame Images via Deep Learning and Symbolic Time Series Analysis", Annual Conference of the Prognostics and Health Management Society, vol.6, 2015.
[27] M. Sharma, K. Das, M. Bilgic, B. Matthews, D. Nielsen, N. Oza, "Active Learning with Rationales for Identifying Operationally Significant Anomalies in Aviation", Machine Learning and Knowledge Discovery in Databases. Lecture Notes in Computer Science, vol 9853, 2016
[28] M-Y. Su, "Real-time anomaly detection systems for Denial-of-Service attacks by weighted k-nearest-neighbor classifiers", Expert Systems with Applications, vol. 38(4), 2011, pp. 3492-3498.
[29] S. T. Teoh, K. Zhang, S-M. Tseng, K-L. Ma, S. F. Wu, "Combining visual and automated data mining for near-real-time anomaly detection and analysis in BGP" ACM workshop on Visualization and data mining for computer security. ACM, New York, NY, USA, 2004, pp. 35-44.
[30] G. Thatte, U. Mitra and J. Heidemann, "Parametric Methods for Anomaly Detection in Aggregate Traffic," IEEE/ACM Transactions on Networking, vol. 19(2), 2011, pp. 512-525.
[31] W. Yan and L. Yu, "On Accurate and Reliable Anomaly Detection for Gas Turbine Combustors: A Deep Learning Approach", Annual Conference of the Prognostics and Health Management Society, vol. 6, 2015.
[32] L. Zhang, J. Lin, and R. Karim, 'Adaptive Kernel Density-based Anomaly Detection for Nonlinear Systems', 2016.
[33] M. Zhang, B. Xu and J. Gong, "An Anomaly Detection Model Based on One-Class SVM to Detect Network Intrusions," 11th International Conference on Mobile Ad-hoc and Sensor Networks (MSN), Shenzhen, 2015, pp. 102-107.
[34] S. Zhao, M. Chandrashekar, Y. Lee and D. Medhi, "Real-time network anomaly detection system using machine learning," 11th International Conference on the Design of Reliable Communication Networks (DRCN), Kansas City, MO, 2015, pp. 267-270.
[35] С. Антипов, М. Фомина, «Проблема обнаружения аномалий в наборах временных рядов», Программные продукты и системы № 2, 2012, с. 78 – 82.
[36] Д. Заварзин "К вопросу поиска аномалий во временных рядах", Инновации в науке: сб. ст. по матер. XXIX междунар. науч.-практ. конф. № 1(26). – Новосибирск: СибАК, 2014.
[37] Е. В. Зубков, В. М. Белов, «Методы интеллектуального анализа данных и обнаружение вторжений», Вестник СибГУТИ № 1, 2016.
[38] А. Денисова, В. Мясников, «Обнаружение аномалий на гиперспектральных изображениях», КО. №2, 2014.
[39] Я. Кокорева, А. Макаров, «Поэтапный процесс кластерного анализа данных на основе алгоритма кластеризации k-means», Молодой ученый, №13, 2015. с. 126-128.
[40] А. Суханов, «Интеллектуальные методы обнаружения и прогнозирования аномальных событий в темпоральных данных», Диссертация на соискание ученой степени кандидата технических наук, РГУПС, Ростов-на-Дону, 2015.

The Overview Of Anomaly Detection Methods in Data Streams

Viacheslav P. Shkodyrev, Kamil I. Yagafarov, Valentina A. Bashtovenko, Ekaterina E. Ilyina

This article is devoted to the research of different approaches to the anomaly detection in time-series data, which includes identification and processing of deviations in data streams obtained from technological process. Detection of anomalies in system behavior helps not only to increase quality of these processes, but also to avoid emergency situations and accidents at the early stages. All of these demonstrates the relevance of the topic.

Existing methods and algorithms of anomaly detection are reviewed in this paper. The aim of research lies in structuring of available techniques and providing a subsequent selection of methods for system of anomaly detection development for Big Data streams.