Гультяев А.А. Методы обнаружения аномалий в потоковых данных высокой размерности// Международный журнал информационных технологий и энергоэффективности. – 2025. – Т. 10 № 1(51) с. 28–35

Международный журнал информационных технологий и энергоэффективности

Сайт журнала: http://www.openaccessscience.ru/index.php/ijcse/

УДК 004.421

МЕТОДЫ ОБНАРУЖЕНИЯ АНОМАЛИЙ В ПОТОКОВЫХ ДАННЫХ ВЫСОКОЙ РАЗМЕРНОСТИ

Гультяев А.А.
ФГАОУ ВО "НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЯДЕРНЫЙ УНИВЕРСИТЕТ "МИФИ", Москва, Россия, (115409, город Москва, Каширское ш., д.31), e-mail: angultiaev@gmail.com

В статье рассматривается применимость алгоритмов машинного обучения, использующихся для решения задачи обнаружения аномалий, к непрерывным потокам данных высокой размерности, таким как показания датчиков и сенсоров или векторные представления последовательных данных, таких как части видеоряда. Ключевыми аспектами применимости алгоритмов являются возможности «холодного старта», онлайн-обучения, коррекции ответов, путем взаимодействия с оператором, а также вычислительная сложность и производительность. В статье рассмотрены детали реализации алгоритмов для возможности обработки данных высокой размерности, а также приведен сравнительных анализ их качества по нескольким метрикам машинного обучения.

Ключевые слова: Машинное обучение, искусственный интеллект, нейронная сеть, векторное представление, обнаружение аномалий.

ANOMALY DETECTION METHODS IN HIGH-DIMENSIONAL STREAMING DATA

Gultyaev A.A.
NATIONAL RESEARCH NUCLEAR UNIVERSITY MEPHI, Moscow, Russia, (115409, Moscow, Kashirskoye shosse, 31), e-mail: angultiaev@gmail.com

This paper addresses the applicability of machine learning algorithms used to solve the anomaly detection task to high dimensional continuous data streams such as sensor and sensor readings or vector representations of sequential data such as parts of a video sequence. Key aspects of the applicability of the algorithms are cold-start capabilities, online learning, correction of responses, through operator feedback, and computational complexity and performance. The paper discusses the implementation details of the algorithms to be able to process high dimensional data, and provides a comparative analysis of their quality on several machine learning metrics.

Keywords: Machine learning, artificial intelligence, neural network, vector embedding, anomaly detection.

Введение

Задача обнаружения аномалий, также известная как задача обнаружения выбросов – это задача выявления редких элементов, событий или наблюдений в данных, которые значительно отклоняются от подавляющего большинства.

Термин «аномалия» стал набирать популярность в XX веке, когда анализ данных стал включать в себя различные приложения, такие как обнаружение мошенничества и контроль качества. С появлением информатики в середине XX века акцент сместился на автоматизацию этих процессов. К концу 1990-х - началу 2000-х годов доступность больших массивов данных и достижения в области машинного обучения позволили разработать сложные методы обнаружения аномалий.

Существует три основных типа аномалий, встречающихся в данных:

1. Точечные аномалии. Один экземпляр данных, значительно отличается от остальной части набора данных. Примером точечной аномалии является транзакция на сумму 100 000 рублей в наборе данных, состоящем из типичных транзакций на сумму от 100 до 1000 рублей.
2. Контекстные аномалии. Экземпляр данных, является аномальным в определенном контексте, но не в других случаях. Контекстные аномалии характерны для временных рядов и пространственных данных. Примером контекстной аномалии является всплеск потребления электроэнергии в полночь, который необычен по сравнению с обычным полуночным использованием.
3. Коллективные аномалии. Некоторое количество данных в целом отклоняется от нормы, даже если отдельные точки данных не являются аномальными. Примером такой аномалии является последовательность сетевых запросов, происходящая при кибератаке.

Методы выявления аномалий широко используются в финансовой сфере, здравоохранении, информационной безопасности, маркетинге и продажах, промышленности и т.д.

Методы обнаружения аномалий разделяются на три основных категории: методы обучения с учителем (англ. supervised), методы обучения без учителя (англ. unsupervised) и методы обучения с частичным привлечением учителя (англ. semi-supervised). Алгоритмам обнаружения аномалий с учителем требуется размеченный набор данных для «нормальных» и «аномальных» событий. Для решения такой задачи могут быть использованы классические алгоритмы классификации, такие как машины опорных векторов [1], деревья решений [2], случайные леса [3] или многослойные перцептроны. Для обнаружения аномалий без учителя не требуется размеченный набор данных, а алгоритмы полагаются на неявные зависимости в данных для выявления аномальных событий. В этом случае могут быть использованы методы кластеризации, такие как DBSCAN [4], методы, основывающиеся на плотности распределения исходных данных, такие как Local Outlier Factor (LOF) [5] или Isolation Forest [6]. Методам обнаружения аномалий с частичным привлечением учителя также требуется размеченный набор данных, но только «нормальных» событий. Аномалии идентифицируются как отклонения от «выученных» шаблонов «нормальности» событий. Примерами таких методов является одноклассовый метод опорных векторов [7] или модели гауссовский смесей (англ. Gaussian Mixture Models) [8].

Методам обнаружения аномалий зачастую требуется набор данных, на котором модель может обучаться. Если такой набор данных недоступен, или данные приходят со временем, классические подходы к распознаванию аномалий не применимы. Ситуация, в которой набор исторических наблюдений недоступен на момент запуска работы модели, называется холодным стартом. Для решения такого рода задач необходимы алгоритмы, которые могут обучаться на новых данных, которые приходят со временем. Обучение таких моделей называется онлайн-обучением.

Помимо вышеуказанных проблем, алгоритм, еще не обученный на достаточном количестве данных будет часто ошибаться при решении задачи. Для более быстрого обучения, в алгоритм может быть внедрен механизм обратной связи с оператором. При ложноположительном или ложноотрицательном ответе алгоритма, оператор может вручную указать на факт отсутствия или присутствия аномалии в текущей части данных.

В статье рассмотрены алгоритмы обнаружения аномалий, такие как:

1. онлайновый метод k-ближайших соседей (Online k-NN);
2. онлайновый метод опорных векторов (Online SVM);
3. инкрементальная модель гауссовских смесей (Incremental GMM);
4. онлайновый случайный лес;
5. автокодировщик;
6. самоорганизующиеся карты;
7. методы обучения с подкреплением;
8. инкрементальный алгоритм DBSCAN.

Все алгоритмы могут обучаться на последовательно поступающих данных, а также прямо или косвенно допускать взаимодействие с оператором.

Сравнительный анализ алгоритмов

Онлайновый метод k-ближайших соседей (Online k-NN)

Алгоритм k-NN классифицирует точки данных на основе мажоритарных классов среди их ближайших соседей в пространстве признаков. Для обнаружения аномалий может быть использован порог метрики расстояния между наблюдениями. В таком случае, если расстояние до новой точки данных превышает этот порог, объект будет классифицирован как аномальный.

Алгоритм k-NN обучается путем сохранения исходной выборки данных для последующего расчета расстояний. Онлайн-обучение такого алгоритма достигается за счет добавления новых точек данных в исходную выборку.

Метод k-ближайших соседей прост в реализации, и поддерживает добавление исходных данных без обучения заново. Однако, с ростом обучающей выборки, расчет расстояний будет становиться более вычислительно сложным процессом [9], а выбор порога расстояния является нетривиальной задачей, требующей экспериментов и тонкой настройки алгоритма. Стоит также отметить, что алгоритм k-ближайших соседей работает менее эффективно на данных высокой размерности.

Онлайновый метод опорных векторов (Online SVM)

Метод опорных векторов стремится найти гиперплоскость, наилучшим образом разделяющую два класса, максимизируя расстояние между классами [1]. Онлайновая версия метода адаптирует этот подход для потоковых данных. Обучение начинается с минимального набора размеченных данных, и, с получением размеченных новых данных (например, от обратной связи с оператором), координаты гиперплоскости изменяются, а новые объекты классифицируются исходя того, с какой стороны от построенной гиперплоскости они находятся. [10]

Данный метод эффективен в пространствах высокой размерности. Модель может обновляться путем использования таких техник, как стохастический градиентный спуск, а функции ядер позволяют улавливать нелинейные зависимости в данных. Однако метод требует большое количество размеченных данных, а следовательно, интенсивного взаимодействия с оператором.

Инкрементальная модель гауссовских смесей (Incremental GMM)

GMM моделируют распределения данных как смесь нескольких гауссовских распределений, отражающих основные закономерности исходных данных. Модель обучается на данных, представляющих «нормальные» события. При классификации новой точки данных, модель рассчитывает оценку правдоподобия принадлежности этой точки данных смеси распределений [8]. Низкая оценка правдоподобия указывает на возможное наличие аномалии. С получением новых данных или путем взаимодействия с оператором, параметры модели обновляются с использованием инкрементального EM-алгоритма [11].

Инкрементальная модель гауссовский смесей позволяет получать оценку правдоподобия для задачи распознавания аномалий, которую можно интерпретировать как вероятность аномалии. Тем не менее, модели такого типа склонны переобучаться, а использование инкрементального EM-алгоритма является вычислительно сложной задачей.

Онлайновый случайный лес (Online Random Forest)

Случайный лес является ансамблевым алгоритмом, использующим несколько решающих деревьев для повышения эффективности классификации. Вместо классических деревьев решений в алгоритме использованы деревья VFDT (Very Fast Decision Tree), также известные как деревья Хеффдинга [12], которые позволяют обучаться на потоковых данных. Аномальность точки данных может быть определена в зависимости от того, насколько глубоко в дереве оказалась данная точка.

Алгоритм случайного леса в целом позволяет добиваться большей точности классификации, уменьшая разброс предсказаний, а также может предоставить информацию о том, какие признаки в обучающей выборке более важны. Несмотря на то, что данный способ подходит для большого количества данных, алгоритм VFDT достаточно сложен в реализации по сравнению с классическими деревьями решений, такими как CART [2], а также не может быстро адаптироваться к новым зависимостям, полученным из данных.

Автокодировщик (AutoEncoder)

Автокодировщики являются нейронными сетями, обучаемыми для реконструирования исходных данных по неявным зависимостям в них. Автокодировщики обычно состоят из двух частей: кодировщика и декодера. Первый представляет исходные данные в виде векторов меньшей размерности, а второй на основе данного вектора пытается восстановить исходные данные [13]. В задаче распознавания аномалий автокодировщик обучается на «нормальных» примерах. При получении новой точки данных, рассчитывается ошибка реконструирования (например, среднеквадратичная ошибка) данной точки. Поскольку, алгоритм не обучался на аномальных примерах, высокие значения такой ошибки могут являться индикаторами аномалий.

Автокодировщики являются мощными инструментами, т.к. могут улавливать сложные нелинейные зависимости, а также позволяют оперировать данными больших размерностей. Однако важно понимать, что обучение нейронных сетей достаточно трудоемкий процесс, а выбор архитектуры и гиперпараметров модели требует значительного опыта от разработчика.

Самоорганизующаяся карта (Self-Organizing Map, SOM)

Самоорганизующаяся карта является нейронной сетью, обучающейся без учителя. Данный алгоритм снижает размерность признакового пространства, при этом сохраняя

топологические характеристики исходных данных. Нейронная сеть обучается организовывать похожие точки данных ближе друг к другу в признаковом пространстве [14]. В данном случае, аномалиями будут являться точки данных, попавшие в области, в которых небольшое число экземпляров из обучающей выборки, или их нет совсем.

Самоорганизующиеся карты поддерживают онлайн обучение на новых примерах, однако могут терять свою эффективность на данных очень высокой размерности.

Методы обучения с подкреплением (Reinforcement Learning)

Обучение с подкреплением включает в себя обучение политике максимизации совокупного вознаграждения при взаимодействии с некой средой [15]. Для задачи обнаружения аномалий, точки данных представляют собой состояния среды, действиями модели являются решения об аномальности новой точки данных, модель получает вознаграждение, если оператор согласен с ее решением.

Данный подход, несмотря на свою эффективность при обучении сложными зависимостям и возможность прямого взаимодействия с оператором, сложен в реализации, а также требует большого количества контактов с оператором для эффективного обучения.

Инкрементальный алгоритм DBSCAN

Алгоритм DBSCAN использует предположение о достижимости одной точки исходных данных из других точек на основе выбранного порога метрики расстояния. Если точка недостижима из других точек обучающей выборки, она является выбросом [4]. В задаче выявления аномалий, индикатор выброса может служить индикатором аномального события.

Данный алгоритм поддерживает онлайн обучение по мере получения новых обучающих данных, и естественным образом подходит для задачи идентификации выбросов (аномалий), однако алгоритм очень чувствителен к выбору гиперпараметров, таких как метрика расстояния, количество точек или порог расстояния. Помимо этого, с ростом размера обучающей выборки, вычислительная сложность алгоритма также растет.

В таблице ниже приведена оценка каждого из рассмотренных методов по трем критериям: эффективность работы на больших выборках, эффективность работы на данных высокой размерности, а также субъективная оценка сложности реализации и настройки алгоритма.

Таблица 1.- Оценки рассмотренных алгоритмов

Алгоритм\Критерий	Эффективность для больших выборок	Эффективность на данных высокой размерности	Сложность реализации и настройки
Online k-NN	Низкая	Низкая	Низкая
Online SVM	Средняя	Средняя	Средняя
Incremental GMM	Средняя	Высокая	Высокая
Online Random Forest	Высокая	Средняя	Средняя
Autoencoder	Высокая	Высокая	Средняя
Self-Organizing Map	Средняя	Низкая	Низкая
Reinforcement Learning	Высокая	Высокая	Высокая
Incremental DBSCAN	Средняя	Средняя	Средняя

Результаты экспериментов

Рассмотренные алгоритмы протестированы при решения задачи обнаружения аномалий. Размер выборки данных для решения задачи составляет 17 280 наблюдений, каждое наблюдение представляет собой вектор размерности 1024, а баланс классов составляет 99% «нормальных» событий и 1% аномалий. Процесс обучения каждой из моделей начинается с необученной модели, а данные передаются в модель последовательно.

В Таблице 2 ниже приведены оценки точности, полноты и f1-метрики рассмотренных алгоритмов при решении задачи распознавания аномалий на потоковых данных.

Таблица 2. - Метрики качества решения задачи обнаружения аномалий

Алгоритм\Метрика	Precision	Recall	f1-score
Online k-NN	0.28	0.3	0.29
Online SVM	0.54	0.68	0.60
Incremental GMM	0.64	0.72	0.68
Online Random Forest	0.36	0.78	0.49
Autoencoder	0.76	0.88	0.81
Self-Organizing Map	0.29	0.67	0.40
Reinforcement Learning	0.67	0.71	0.69
Incremental DBSCAN	0.43	0.52	0.47

Наихудшее качество показал алгоритм k-ближайших соседей, однако такой результат ожидаем в ситуациях, когда данные имеют высокую размерность.

Алгоритм онлайнового случайного леса и самоорганизующаяся карта показали достаточно высокую полноту, но плохую точность. Такой результат связан с тем, что по мере обучения случайного леса, деревья в нем становятся все глубже, и процесс обнаружения аномалий, опирающийся на глубину нахождения конкретной вершины осложняется. Процесс обнаружения аномалий для самоорганизующейся карты заключается в выявлении точек данных, попавших в области, в которых находятся мало примеров из обучающей выборки, или они отсутствуют вовсе. В данном случае, по мере обучения самоорганизующейся карты, она теряет способность распознавать аномалии из-за увеличения размера обучающей выборки, о чем и свидетельствуют результаты эксперимента.

Наиболее высокие значения метрик показали модели автокодировщика, гауссовских смесей и обучения с подкреплением. Однако, учитывая сложность модели обучения с подкреплением, для более высокого качества ей требуется больший размер обучающей выборки. Автокодировщик, учитывая его относительную простоту в реализации и в обучении, является лидирующим алгоритмом в решении задачи выявления аномалий в потоковых данных высокой размерности.

Заключение

Классические методы обнаружения аномалий имеют ограниченную область применения из-за необходимости наличия размеченных обучающих данных. В статье рассмотрены алгоритмы, которые позволяют эффективно бороться нехваткой данных и проблемой холодного старта, а также работать с последовательно поступающими данными.

Необходимыми критериями при выборе алгоритмов являлись поддержка онлайн-обучения и возможность прямого или косвенного взаимодействия с оператором.

Каждый из рассмотренных алгоритмов имеет свою область применения, преимущества и недостатки. Например, метод k-ближайших соседей и самоорганизующиеся карты хуже работают с данными высокой размерности, что подтверждается экспериментами. По результатам тестирования выявлено, что наилучшим образом с задачей выявления аномалий в потоковых данных высокой размерности справляются автокодировщики. Также с решением данной задачи хорошо справляются модели гауссовских смесей. Методы обучения с подкреплением также показывают перспективные результаты, однако из-за сложности реализации и обучения модели, а также необходимости большего количества данных для обучения, хуже подходят для решения поставленной задачи.

Стоит также отметить, что разные задачи могут иметь разную специфику, и использование алгоритмов машинного обучения может быть излишним при наличии в исходных данных сильных статистических зависимостей, а при реализации методов для решения задач выявления аномалий, помимо возможностей алгоритма следует также учитывать специфики конкретной задачи и обучающих данных.

Список литературы

1. Bishop C. M., Nasrabadi N. M. Pattern recognition and machine learning. – New York: Springer, 2006. – Т. 4. – №. 4. – С. 338-339.
2. Breiman L. Classification and regression trees. – Routledge, 2017.
3. Breiman L. Random forests //Machine learning. – 2001. – Т. 45. – С. 5-32.
4. 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.
5. Breunig M. M. et al. LOF: identifying density-based local outliers //Proceedings of the 2000 ACM SIGMOD international conference on Management of data. – 2000. – С. 93-104.
6. Liu F. T., Ting K. M., Zhou Z. H. Isolation forest //2008 eighth ieee international conference on data mining. – IEEE, 2008. – С. 413-422.
7. Schölkopf B. et al. Estimating the support of a high-dimensional distribution //Neural computation. – 2001. – Т. 13. – №. 7. – С. 1443-1471.
8. Reynolds D. A. et al. Gaussian mixture models //Encyclopedia of biometrics. – 2009. – Т. 741. – С. 659-663.
9. Andoni A., Indyk P., Razenshteyn I. Approximate nearest neighbor search in high dimensions //Proceedings of the International Congress of Mathematicians: Rio de Janeiro 2018. – 2018. – С. 3287-3318.
10. Crammer K. et al. Online passive-aggressive algorithms //Journal of Machine Learning Research. – 2006. – Т. 7. – №. 3.
11. Neal R. M., Hinton G. E. A view of the EM algorithm that justifies incremental, sparse, and other variants //Learning in graphical models. – Dordrecht : Springer Netherlands, 1998. – С. 355-368.
12. Hulten G., Spencer L., Domingos P. Mining time-changing data streams //Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. – 2001. – С. 97-106.

13. Zhai J. et al. Autoencoder and its various variants //2018 IEEE international conference on systems, man, and cybernetics (SMC). – IEEE, 2018. – С. 415-419.
14. Kohonen T. Self-organized formation of topologically correct feature maps //Biological cybernetics. – 1982. – Т. 43. – №. 1. – С. 59-69.
15. Sutton R. S., Barto A. G. Reinforcement learning: An introduction. – MIT press, 2018.

References

1. Bishop C. M., Nasrabadi N. M. Pattern recognition and machine learning. – New York: Springer, 2006. – Vol. 4. – No. 4. – pp. 338-339.
2. Breiman L. Classification and regression trees. – Routledge, 2017.
3. Breiman L. Random forests //Machine learning. – 2001. – Vol. 45. – pp. 5-32.
4. Ester M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise //kdd. – 1996. – Vol. 96. – No 34. – pp. 226-231.
5. Breunig M. M. et al. LOF: identifying density-based local outliers //Proceedings of the 2000 ACM SIGMOD international conference on Management of data. – 2000. – pp. 93-104.
6. Liu F. T., Ting K. M., Zhou Z. H. Isolation forest //2008 eighth ieee international conference on data mining. – IEEE, 2008. – pp. 413-422.
7. Schölkopf B. et al. Estimating the support of a high-dimensional distribution //Neural computation. – 2001. – Vol. 13. – No. 7. – pp. 1443-1471.
8. Reynolds D. A. et al. Gaussian mixture models //Encyclopedia of biometrics. – 2009. – Vol. 741. – pp. 659-663.
9. Andoni A., Indyk P., Razenshteyn I. Approximate nearest neighbor search in high dimensions //Proceedings of the International Congress of Mathematicians: Rio de Janeiro 2018. – 2018. – pp. 3287-3318.
10. Crammer K. et al. Online passive-aggressive algorithms //Journal of Machine Learning Research. – 2006. – Vol. 7. – No. 3.
11. Neal R. M., Hinton G. E. A view of the EM algorithm that justifies incremental, sparse, and other variants //Learning in graphical models. – Dordrecht : Springer Netherlands, 1998. – pp. 355-368.
12. Hulten G., Spencer L., Domingos P. Mining time-changing data streams //Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining. – 2001. – pp. 97-106.
13. Zhai J. et al. Autoencoder and its various variants //2018 IEEE international conference on systems, man, and cybernetics (SMC). – IEEE, 2018. – pp. 415-419.
14. Kohonen T. Self-organized formation of topologically correct feature maps //Biological cybernetics. – 1982. – Vol. 43. – No. 1. – pp. 59-69.
15. Sutton R. S., Barto A. G. Reinforcement learning: An introduction. – MIT press, 2018.