Инженерный вестник Дона, №6 (2023)
ivdon.ru/ru/magazine/archive/n6y2023/8475

Обзор основанных на техниках машинного обучения методов обнаружения выбросов в данных

А.С. Дулесов, А.В. Байшев

Хакасский государственный университет им. Н.Ф. Катанова, Абакан

Аннотация: В статье рассмотрены методы обнаружения выбросов, основанные на различных техниках машинного обучения: контролируемые (англ. supervised), неконтролируемые (англ. unsupervised), полуконтролируемые (англ. semi-supervised). Обозначены особенности применения тех или иных методов, указаны их достоинства и ограничения. Установлено, что не существует универсального способа обнаружения выбросов подходящего для различных данных, поэтому выбор того или иного конкретного метода для реализации исследований следует производить, исходя из анализа преимуществ и ограничений присущих выбранному способу с обязательным учетом возможностей располагаемых вычислительных мощностей и характеристик имеющихся в наличии данных, в том числе, включающих их классификацию на выбросы и нормальные данные, а также объем.

Ключевые слова: выбросы, машинное обучение, обнаружение выбросов, анализ данных, интеллектуальный анализ данных, большие данные, анализ главных компонент, регрессия, изолирующий лес, машина опорных векторов.

Введение

Выбросы – это наблюдения, существенно отличающиеся от других наблюдений имеющегося набора данных [1]. По форме своего присутствия они могут представлять собой как одно значение, так и являться их последовательностью [2, 3]. В разных областях их называют также аномалиями, противоречивыми наблюдениями, исключениями и т.д. [2].

Обнаружение выбросов – одна из актуальнейших задач интеллектуального анализа данных в разных сферах деятельности человека [3]. Их выявление помогает решать проблемы обнаружения мошенничества в банковском секторе [2, 3], вести наблюдение за действиями противников в военной среде [2], идентифицировать дефекты в системах отопления [4], обнаруживать ошибки в компьютерных сетях [2, 5], выявлять неполадки космических кораблей [6] и т.д.

Методов обнаружения выбросов существует множество, поэтому систематизация информации об особенностях тех или иных из них являет

собой актуальную задачу решение которой несет потенциальную пользу исследователям из различных сфер.

Целью исследования является осуществление неисчерпывающего обзора особенностей применения методов обнаружения выбросов в данных, основанных на различных техниках машинного обучения, в том числе: контролируемых (англ. supervised), т.е. основанных на методах машинного обучения с учителем; неконтролируемых (англ. unsupervised), т.е. методы, основанных на применении техник машинного обучения без учителя; полуконтролируемых (англ. semi-supervised), т.е. методов, основанных на машинном обучении с частичным привлечением учителя.

Контролируемые методы

В методах этого типа требуется, чтобы модель, используемая для выявления выбросов, была предварительно обучена и проверена на размеченном наборе данных, где в явном виде указаны выбросы и нормальные данные [3, 7]. Соответственно, выявление выбросов в неизвестных ранее для модели данных является вопросом решения задачи классификации. Использование контролируемых методов особо актуально для сфер, где процедура разметки данных не вызывает трудностей [7].

Пользуются популярностью следующие методы контролируемого обучения для выявления выбросов: логистическая регрессия, деревья решений, наивный байесовский классификатор, машина опорных векторов (англ. Support Vector Machines – SVM), k-ближайших соседей (англ. k-nearest neighbour – kNN), нейросетевые методы [8, 9].

Логистическая регрессия, деревья решений и наивный байесовский классификатор являются простыми быстро адаптируемыми и эффективными методами, используемыми для решения задач классификации в разных областях [10, 11]. Однако, наивный байесовский классификатор и логистическая регрессия требуют независимости признаков данных, что на

практике выполняется не всегда и может снизить их эффективность [8]. В свою очередь, деревья решений хоть и являются мощным и интерпретируемым средством классификации, но требуют правильной настройки для недопущения переобучения.

SVM обладает плюсами в выявлении выбросов в данных даже большой размерности и подходит, в том числе, для их обнаружения в реальном времени [8, 9]. Плюсом SVM является возможность выбора ядра, что дает ему некоторую гибкость. Ядро – это специальная функция, определяющая то, как будут преобразованы данные обучающей выборки из исходного пространства в более многомерное пространство признаков, для которого строится гиперплоскость, разделяющая данные на классы. Недостатками SVM являются: чувствительность к шумам, необходимость в стандартизации данных, плохая результативность в случае наличия перекрывающихся классов, а также отсутствие четкого подхода к определению ядра [12].

kNN определяет к какому классу принадлежат данные на основании классов ближайших соседей [7, 13]. Основа метода состоит в том, что обнаруживаемые выбросы должны быть ближе к известным ранее [8].

kNN прост в реализации, нечувствителен к наличию выбросов, интерпретируем, гибок [12]. Однако он вычислительно сложен, и для обработки больших объемов данных требует наличия соответствующих мощностей.

Применение нейросетей для решения задач классификации также весьма популярно в решении задач выявления выбросов. Они подходят для многоклассового выявления выбросов, устойчиво работают с данными, обладающими различными видами распределения, хорошо работают даже при их применении для обработки больших объемов данных. Однако процесс обучения моделей такого класса может быть весьма длительным из-за

вычислительной сложности, потребности в обработке больших объемов данных, а также подборе правильных гиперпараметров [14–16].

В целом, к недостатку контролируемых методов обнаружения выбросов можно отнести то, что количество выбросов относительно нормальных данных в обучающем наборе часто бывает несбалансированным, в результате чего модель смещает результат оценки принадлежности данных в сторону более объемного класса [8]. В некоторых случаях такая ситуация может быть разрешена настройкой того или иного гиперпараметра модели, однако, это не всегда помогает.

Неконтролируемые методы

Эти методы не требуют наличия размеченных данных и основаны на предположении о том, что нормальные данные встречаются в обучающем наборе гораздо чаще выбросов. Соответственно, если это предположение в реальности оказывается неверно, то такие методы будут работать некорректно [8].

Среди множества неконтролируемых методов можно выделить класс основанных на кластеризации. Процесс кластеризации позволяет разделять массу данных на группы(кластеры), внутри которых выделенные объекты более схожи между собой, чем с какими-либо прочими из другого кластера [17]. Основа этих методов заключается в том, что нормальные данные должны быть сгруппированы в кластерах с большим количеством наблюдений, в свою очередь выбросы не должны принадлежать ни одному кластеру или принадлежать кластерам с малым числом наблюдений [8]. Способы обнаружения выбросов на основе кластеризации предполагают использование различных алгоритмов: к-средних (англ. k-means), основанная на плотности пространственная кластеризация для приложений с шумами (англ. density-based spatial clustering of applications with noise – DBSCAN), упорядочение точек для обнаружения кластерной структуры (англ. ordering

points to identify the clustering structure – OPTICS) и прочие. Проблемы их применения зависят как от выбранного алгоритма, так и от объема и качества исследуемых данных. Так могут проявляться в разной степени проблемы: трудная интерпретация результатов, вычислительная сложность, необходимость задания количества кластеров, чувствительность к выбросам [17].

Изолирующий лес – один из популярных неконтролируемых методов обнаружения выбросов. В его основе лежит предположение о том, что выбросы склонны быть удаленными(изолированными) от нормальных данных. Эта парадигма отличает изолирующий лес от большинства прочих существующих методов обнаружения выбросов, формирующих профиль нормальных данных и затем использующих его для выявления аномальных. Изолирующий лес обнаруживает выбросы на основе множества случайно выстраиваемых бинарных деревьев, с помощью которых он последовательно разделяет данные и идентифицирует как выбросы те из них, которые быстрее изолируются в деревьях по сравнению с большинством остальных данных [18]. Алгоритм обладает линейной временной сложностью и отличается малыми требованиями к памяти, поэтому хорошо работает даже с большими объемами данных [8, 18]. К недостаткам можно отнести необходимость в правильном подборе параметров деревьев и выборки данных.

Известен также и такой алгоритм обнаружения выбросов как фактор локальных выбросов (англ. local outlier factor – LOF). Он использует плотность точек данных в распределении в качестве ключевого фактора для обнаружения выбросов. LOF, как и kNN, использует оценку расстояний до k-ближайших соседей для определения выбросов. Он предусматривает вычисление показателя LOF на основе как локальной плотности используемой выборки, так и ее k-ближайших соседей. LOF неплохо идентифицирует локальные выбросы, однако, значимым недостатком метода

является его малая пригодность для работы с большими объемами данных, из-за его вычислительной сложности, а также необходимость в подборе гиперпарамтеров [18].

Полуконтролируемые методы

Полуконтролируемые методы – это пограничный класс методов между неконтролируемыми и контролируемыми методами выявления выбросов и обладают сходством с каждым из них. Например, многие полуконтролируемые методы могут быть адаптированы для работы в неконтролируемом режиме [8].

Эти методы, как и в случае контролируемых, используют размеченные данные, в которых в явном виде обозначены две категории данных: выбросы и нормальные. Однако, в отличие от контролируемых методов, для построения моделей используют преимущественно одну категорию данных, что делает их применяемыми для выявления выбросов даже в несбалансированных наборах данных [8].

Одним из популярных методов рассматриваемого вида является одноклассовый метод опорных векторов (англ. one-class support vector machines – OSVM). Его используют при решении задач не только выявления выбросов, но также для классификации и регрессии. Алгоритм OSVM, также, как и стандартный SVM, зависим от настроек параметров используемого в модели ядра, а также от настроек гиперпараметров [8, 9].

В вопросах обнаружения выбросов хорошо зарекомендовали себя также автоэнкодеры [19, 20]. Автоэнкодер – это нейронная сеть, которая копирует входные данные на выход. Обнаружение аномалий ими осуществляется путем сжатия обычных данных в скрытое пространство меньшей, чем исходные данные, размерности с помощью автоэнкодера, а затем восстановления их с помощью декодера и сравнения разницы между исходными и восстановленными данными.

Автоэнкодер обучается путем минимизации разницы между исходными и восстановленными данными, при этом для процесса их обучения используют только нормальные, не содержащие выбросов данные. В результате реализации такого подхода, автоэнкодеры при их применении на реальных данных сжимают и восстанавливают нормальные данные с меньшей ошибкой, чем в случаях обработки аномальных данных. На основе анализа ошибки и происходит процесс идентификации выбросов.

К недостаткам применения автоэнкодеров можно отнести тот факт, что в некоторых случаях они могут воспроизводить с малой ошибкой даже данные с аномалиями [20]. Кроме того, для создания модели автоэнкодера часто требуется большое количество данных и времени для его обучения и настройки гиперпараметров .

По схожему с методами, основанными на автоэнкодерах, принципу работают и методы, основанные на широко известном способе анализа главных компонент, в которых также используется снижение размерности и их восстановление для выявления выбросов [21]. Однако, как правило, такие методы обладают высокой вычислительной сложностью [21].

Заключение

На основании вышеизложенного, можно сделать вывод о том, что на сегодняшний день существуют различные методы обнаружения выбросов в данных, основанные на различных техниках машинного обучения, каждый из которых обладает своими достоинствами и недостатками. Учитывая тот факт, что универсального метода не существует, выбор того или иного для реализации собственного исследования, следует производить, исходя из анализа преимуществ и ограничений, присущих выбранному способу, с учетом возможностей вычислительных мощностей и характеристик имеющихся в наличии данных, в том числе, включающих их классификацию на выбросы и нормальные данные, а также объем.

Литература

1. Золотова Т.В., Волкова Д.А. Методы интеллектуальной обработки данных для коррекции атипичных значений котировок акций // Статистика и экономика. 2022. Т. 19, № 2. С. 4-13.

2. Chandola V., Banerjee A., Kumar V. Anomaly detection: A survey. ACM Computing Surveys, 2009, vol. 41(3), pp. 1-58.

3. Zimek A., Filzmoser P. There and back again: Outlier detection between statistical reasoning and data mining algorithms. WIREs Data Mining and Knowledge Discovery, 2018, vol. 8, no. 6, pp. 1-26.

4. Farouq S., Byttner S., Bouguelia M.-R., Nord N., Gadd H. Large-scale monitoring of operationally diverse district heating substations: A reference-group based approach. Engineering Applications of Artificial Intelligence, 2020, vol. 90, p. 103492.

5. Tartakovsky A.G., Polunchenko A.S., Sokolov G. Efficient computer network anomaly detection by changepoint detection methods. IEEE Journal of Selected Topics in Signal Processing, 2013, vol. 7(1), pp. 4-11.

6. Fujimaki R., Yairi T., Machida K. An approach to spacecraft anomaly detection problem using kernel feature space. Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, KDD ’05, New York, NY: ACM, 2005, pp. 401-410.

7. Fernandez A., Bella J., Dorronsoro J.R. Supervised outlier detection for classification and regression. Neurocomputing, 2022, vol. 486, pp. 77-92.

8. Arunraj N., Hable R., Fernandes M., Leidl K., Heigl M. Comparison of Supervised, Semi-supervised and Unsupervised Learning Methods in Network Intrusion Detection System (NIDS) Application. Anwendungen und Konzepte der Wirtschaftsinformatik, 2017, vol. 6, pp. 10-19.

9. Manandhar P., Aung Z. Towards Practical Anomaly-Based Intrusion Detection by Outlier Mining on TCP Packets, Database and Expert Systems Applications, 2014, pp. 164-173.

10. Харламов А.А., Ермоленко Т.В., Жонин А.А. Моделирование динамики процессов на основе анализа последовательности текстовых выборок // Инженерный вестник Дона. 2013. №4. URL: ivdon.ru/ru/magazine/archive/n4y2013/2047.

11. Красников И.А., Никуличев Н.Н. Гибридный алгоритм классификации текстовых документов на основе анализа внутренней связности текста // Инженерный вестник Дона. 2013. №3. URL: ivdon.ru/ru/magazine/archive/n3y2013/1773.

12. Линдигрин А.Н. Сравнительный анализ методов машинного обучения в задачах обнаружения сетевых аномалий // Известия ТулГУ. Технические науки. 2019. №12, С. 400-404.

13. Daneshpazhouh A., Sami A. Semi-Supervised Outlier Detection with Only Positive and Unlabeled Data Based on Fuzzy Clustering. International Journal on Artificial Intelligence Tools, 2015, vol. 24(03), p. 1550003.

14. Dau H.A., Ciesielski V., Song A., Anomaly Detection Using Replicator Neural Networks Trained on Examples of One Class. Simulated Evolution and Learning, 2014, pp. 311-322.

15. Yan W., Yu L., On Accurate and Reliable Anomaly Detection for Gas Turbine Combustors: A Deep Learning Approach, Annual Conference of the Prognostics and Health Management Society, 2015, vol. 6, pp. 1-8.

16. Колюцкий К.Н. Нейросетевой подход выявления аномалий в информационной системе // Вестник МФЮА. 2012. №1. С. 49-52.

17. Пестунов И.А., Синявский Ю.Н. Алгоритмы кластеризации в задачах сегментации спутниковых изображений // Вестник КемГУ. 2012. Т2, №4(52). С. 110-125.

18. Chabchoub Y., Togbe M.U., Boly A., Chiky R. An In-Depth Study and Improvement of Isolation Forest, IEEE Access, 2022, vol. 10, pp. 10219-10237.

19. Sakurada M., Yairi T., Anomaly Detection Using Autoencoders with Nonlinear Dimensionality Reduction, Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis - MLSDA'14, 2014, pp. 1-8.

20. Zhisheng X., Yan Q., Amit Y. Likelihood Regret: An Out-of-Distribution Detection Score For Variational Autoencoder. Advances in Neural Information Processing Systems. 33, 2020, pp. 1-12.

21. Nalisnick E., Matsukawa A, Teh Y.W., Gorur D., Lakshminarayanan B. Do Deep Generative Models Know What They Don't Know?, International Conference on Learning Representations, 2019, pp. 1-18.

References

1. Zolotova T.V., Volkova D.A. Statistika i ekonomika, 2022, vol. 19, no. 2, pp. 4-13.

2. Chandola V., Banerjee A., Kumar V. ACM Computing Surveys, 2009, vol.41(3), pp. 1-58.

3. Zimek A., Filzmoser P. WIREs Data Mining and Knowledge Discovery, 2018, vol. 8, no. 6, pp. 1-26.

4. Farouq S., Byttner S., Bouguelia M.-R., Nord N., Gadd H. Engineering Applications of Artificial Intelligence, 2020, vol. 90, p. 103492.

5. Tartakovsky A.G., Polunchenko A.S., Sokolov G. IEEE Journal of Selected Topics in Signal Processing, 2013, vol. 7(1), pp. 4-11.

6. Fujimaki R., Yairi T., Machida K. Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, KDD '05, New York, NY: ACM, 2005, pp. 401-410.

7. Fernandez A., Bella J., Dorronsoro J.R. Neurocomputing, 2022, vol. 486, pp. 77-92.

8. Arunraj N., Hable R., Fernandes M., Leidl K., Heigl M. Anwendungen und Konzepte der Wirtschaftsinformatik, 2017, vol. 6, pp. 10-19.

9. Manandhar P., Aung Z. Database and Expert Systems Applications, 2014, pp. 164-173.

10. Harlamov A.A., Ermolenko T.V., Zhonin A.A. Inzhenernyj vestnik Dona, 2013, no. 4(27). URL: ivdon.ru/ru/magazine/archive/n4y2013/2047.

11. Krasnikov I.A., Nikulichev N.N. Inzhenernyj vestnik Dona, 2013, no. 3(26). URL: ivdon.ru/ru/magazine/archive/n3y2013/1773.

12. Lindigrin A.N. Izvestiya TulGU. Tekhnicheskie nauki, 2019, no. 12, pp. 400-404.

13. Daneshpazhouh A., Sami A. International Journal on Artificial Intelligence Tools, 2015, vol. 24(03), p. 1550003.

14. Dau H.A., Ciesielski V., Song A. Simulated Evolution and Learning, 2014, pp. 311-322.

15. Yan W., Yu L. Annual Conference of the Prognostics and Health Management Society, 2015, vol. 6, pp. 1-8.

16. Kolyutsky K.N. Vestnik MFYUA, 2012, no. 1, pp. 49-52.

17. Pestunov I.A., Sinyavskiy Yu.N. Vestnik KemGU, 2012, vol. 2, no. 4(52), pp. 110-125.

18. Chabchoub Y., Togbe M. U., Boly A., Chiky R. IEEE Access, 2022, vol. 10, pp. 10219-10237.

19. Sakurada M., Yairi T. Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis - MLSDA'14, 2014, pp. 1-8.

20. Zhisheng X., Yan Q., Amit Y. Advances in Neural Information Processing Systems. 33, 2020, pp.1-12.

21. Nalisnick E., Matsukawa A, Teh Y.W., Gorur D. International Conference on Learning Representations, 2019, pp. 1-18.