**Методы поиска аномалий**

**Аннотация**: В работе проводится сравнительный анализ существующих методов поиска аномалий на неразмеченных наборах данных. Приводится описание методов улучшения существующих алгоритмов. Предложен способ усовершенствования существующих алгоритмов путем их ансамблирования.

**Ключевые слова**: Аномалия, поиск аномалий, машинное обучение.

## 0.1 Введение

Задача поиска аномалий является одной из классических задач машинного обучения. В настоящее время задачу поиска аномалий активно решают во многих областях жизнедеятельности:

а) Защита информации и безопастность

б) Социальная сфера и медицина

в) Банковская и финансовая отрасль

г) Распознавание и обработка текста, изображений, речи

д) Другие сферы деятельности(например, мониторинг неисправностей механизмов)

Количество данных в мире удваивается примерно каждые два года. Поэтому актуальной задачей является разработка новых методов и усовершенствование старых методов поиска выбросов.

## 0.2 Классификация методов обнаружений аномалий

Классическая система классификации предполагает предварительное обучение на обучающем наборе данных и последующую классификацию на основе этого набора. Данные делятся на "обучающую выборку" - данные, при помощи которых алогритм обучает классификар и, "тестовую выборку" - данные, при анилизе которых классификатор остается неизменным. Тестовая выборка нужна для того чтобы проверить корректность обучения классификатора.

Однако, в случае с поиском аномалий, возможны варианты, отличающиеся от классического. Подходящий метод классификации выбирается на основе наличия разметки данных. Выделяются три основых класса методов:

а) Обучение с учителем. Для обучения необходимо наличие полностью размеченных данные для обучения и для тестов. Классификатор обучается один раз и применяться впоследствии.В связи с тем, что для многих наборов данных заранее неизвестно что является аномалией, а что нет, применение этого метода ограничено.

б) Обучение с частичным привлечением учителя. Для обучения необходимо наличие тествого и учебного набора данных. Однако, в отличие от обучения с при-

велечением учителя, разметка данных не требуется. Все данные, представленные в выборках, считаются нормальными. На основе этих данных строится некая модель. Все данные, отклоняющиеся от этой модели, считаются аномальными.

в) Обучение без учителя. Самый гибкий способ, который не требует разметки набора данных. Идея заключается в том, что алгоритм обнаружения аномалий оценивает данные исключительно на основе внутренних свойств набора данных что является нормальным, а что является выбросом. В данной работе основное внимание будет этому именно этому способу.

**Результат метода обнаружения аномалий**

В результате работы алгоритма обнаружения аномалий с элементом данных связывается метка или оценка достоверности(показатель аномальности). Метка - показатель, который принимает нулевое значения, в случае если она связана с нормальными данными и единицу в противном случае. Оценка показывает вероятность того, что элемент является аномалией. Для разных алгоритмов используется разные шкалы оценок, поэтому приведение конкретных примеров оценок будет некорректным. В алгоритмах метода обучения с учителем зачастую используются метки как выходные данные, в алгоритмах с частичным привлечением учителя и без учителя обнаружения аномалий чаще встречаются оценки.

**Вероятностно-генеративные методы**

Основная идея генеративных методов заключается в использование вероятностного смесевого моделирования данных. Предлагается подобрать такую вероятностую модель, из которой было получены нормированные данные. Такие модели обычно называются генеративными моделями, где для каждой точки(элемента данных) можем посчитать генеративную вероятность(или вероятность правдоподобия).Т.е. задача сводится к нахождению плотности распределения p(x). Аномалиями при этом считаются точки(элементы набора данных), имеющию низкое правдоподобие. В качестве показателя аномальности выступает функция p. Для построения генеративной модели нужно решить следующую задачу:

$$\prod_{x \in X_{norm}} p(x,\theta) \rightarrow max_{\theta}$$

где $X_{norm}$ - нормальные данные представленного набора данных $p(x,\theta)|\theta \in \omega$ -семейство плотностей вероятностей, параметризованные $\theta$.

Этот метод редко используется на практике, так как тяжело проверить полученную генеративную модель на адекватность, сложно убедится в правильном выборе семейства смесевых распределений. Это связано с тем, что низкое значение функции правдоподобия может означать как и аномальное значение, так и неудачно подобранную модель. Этот метод применяется с опорой на априорную информацию, в случае когда можно проверить полученную модель на адекватность.

**Линейные методы**

Основной идеей линейных методов является построение некой модели, характеризующей нормальные данные. Точки, которые значительно отклоняются от этой модели, считаются аномалиями.

Предполагается, что нормальные данные находятся в подпространсве пространства атрибутов данных(размер подпространства атрибутов данных равен размерности данных). В свою очередь, задача линейного метода - найти низкоразмерные подпространства, такие что, выборка данных этого подпространства значительно отличается от остальных точек пространства данных.

Одним из возможных вариантов решения является использование линейной регрессии. Выбирается одна из наблюдаемых переменных набора данных и относительно неё решается задача линейной регрессии оставшихся атрибутов. Итоговым ответом будет является усредненное значения показателя аномалии по всем атрибутам.

Алгоритмы, основанные на линейном подходе, требуют наличия линейной зависимости атрибутов данных.

**Метрические методы**

Мерические методы пытаются найти в данных точки, в некотором смысле изолированные от остальных[ссылка на источник]. Если в пространстве задана некоторая метрика *p(x1, x2)*, то необходимо задать следующие понятия:

— Аномалии – точки, не попадающие ни в один кластер. К данным применяется один из алгоритмов кластеризации; размер кластера, в котором оказалась точка, объявляется её показатель аномальности.

— Локальная плотность в аномальных точках низкая. Для данной точки показателем аномальности объявляется локальная плотность, которая оценивается некоторым непараметрическим способом.

— Расстояние от данной точки до ближайших соседей велико.

В качестве показателя аномальности может выступать:

— расстояние до k-го ближайшего соседа;

— среднее расстояние до k ближайших соседей;

— медиана расстояний до k ближайших соседей;

— гармоническое среднее до k ближайших соседей;

— доля из k ближайших соседей, для которых данная точка является не более чем k-ым соседом и многое другое.

Метрические методы используют в случае отсутсвия априорной информации о данных. Сложность вычисления прямо пропорциональна как размерности данных m, так и их количеству n. При росте набора данных наблюдается экспоненциальный рост сложности вычислений. Однако, эти методы хорошо проявляют себя на ограниченных наборах данных[ссылка на источник]. Следовательно такие методы как k-ближайших соседей) с нотацией ассимптотического роста $O(n^2)$ недопустимы для наборов данных с большой размерностью, если их размерность не может быть уменьшена.

### Методы улучшения алгоритмов поиска аномалий

Алгоритмы поиска аномалий можно улучшать различными методами, применяемыми в том числе для улучшения результатов работы алгоритмов и в других областях. Например, ансамблирование широко применяется для работы с нейронными сетями. Ниже рассмотрены приемы в контексте поиска аномалий.

### Семплирование

Большинство алгоритмов распознавания аномалий успешно работают на наборах данных малых размеров. Поэтому предлагается разбить начальный набор данных на несколько случайных выборок и усреднить результат. Размер этих выборок может быть как и случайным, там и фиксированного размера, но, как правило, он отличается от размеров исходного набора данных не меньше чем на порядок. Идея такого выбора заключается в том, что шумовые объекты попадут в выборки с низкой вероятностью; кластера нормальных данных будут представлены несколькими представителями, а кластера аномалий выродятся в изолированные точки. На основе этих выборок алгоритмы строят функции показателя аномальности, незначительно уступающему результату, полученному на основе анализа всех исходных данных.

Этот метод помогает значительно сократить вычислительную сложность, а так же уменьшить вероятность "подгона" алгоритма под конкретный набор данных. В силу особенностей задачи, необходимое условие - отсутствия параметризации алгоритмов - зачастую означает их детерминированность(в отсутсвии стохастичности,

показатель аномальности однозначно определяется по заданной выборке). В общем случае при добавлении новых данных в общий набор данных, можно не пересчитывать заново показатель аномальности для всего набора данных, а добавить запуски алгоритма на новых данных в ансамбль(так называемый warm start[ссылка на источник])

**Ансамблирование голосованием**

Ансамблированием в задаче поиска аномалий называют использование нескольких различных алгоритмов с последующи усреднением их показателя аномальности. При использовании различных классов алгоритмов можно столкнуться с проблемой того, что показатель аномальности выглядит по-разному в различных алгоритмах и сравнивать напрямую эти показатели некорректно. Поэтому традиционное приведение показателей значений различных функций к одному диапазону, например,к [0,1], будет некорректным. Существует несколько наиболее известных видов ансамблирования:

— Простое голосование

$$b(x) = F(b_1(x),...,b_T(x)) = \frac{1}{T}\sum_{t=1}^{T} b_t(x)$$

,где $b_i$ - некоторая функция.

— Взвешенное голосование

$$b(x) = F(b_1(x),...,b_T(x)) = \frac{1}{T}\sum_{t=1}^{T} w_t b_t(x)$$

$$\sum_{t=1}^{T} w_t = 1, w_t \geqslant 0$$

,где $w_i$- некоторый коэффициент.

— Смесь экспертов

$$b(x) = F(b_1(x),...,b_T(x)) = \frac{1}{T}\sum_{t=1}^{T} w_t(x) b_t(x)$$

$$\sum_{t=1}^{T} w_t = 1, \forall x \in X$$

,где $w_i(x)$- некоторая функция коэффициента.

Простое голосование - это частный случай взвешенного голосования, а взвешенное голосование является частным случаем смеси экспертов.

Различные методы ансамблирования такие как беггинг, бустинг, стекинг и другие применяются для улучшения работы алгоритмов обучения с учителем . Для

алгоритмов обучения без учителя применяется простое голосование, т.к. задача изменения весов голосования нетривиальна в задаче обучения без учителя[ссылка на источник].

### Итеративный отбор

Итеративный отбор основан на идее многократного применения алгоритмов ансамблирования. Преположим, построена некоторая модель, описывающая нормальные данные. Эта модель построена на основе всех имеющихся данных, но точность этой модели невелика, она умеет определять только явные аномалии. Отсортировав все точки по показателю аномальности, можно выбрать k самых аномальных объекта в данных и исключить из данных. После этого можно перестроить модель и повторить вышеуказанные действия несколько раз, пока не будут достигнуты некоторые условия. При каждой итерации точность модели будет увеличиваться.

Идея итеративного отбора может быть обобщена различными способами. Результат работы одного алгоритма может быть использован для отсеивания явных аномалий и настройки нового алгоритма, не обязательно совпадающего с предыдущим, на оставшихся данных. Возможна и противоположная механика: по результатам работы одного алгоритма отбираются явные, гарантированные представители нормальных данных, и исключительно на них строится модель, их описывающая.

### Сравнение методов

Таблица 0.1 — Сравнение алгоритмов поиска аномалий

| Класс методов | Временная сложность | Расход памяти | Универсальность |
|---|---|---|---|
| Вероятностно-ген. | O(1) | O(n) | Очень низкая |
| Линейный | $\geq O(n^2)$ | $\geq O(n^2)$ | Низкая |
| Параметрический | O(1) | O(n) | Низкая |
| Метрический | $\geq O(nlogn)$ | $\geq O(n)$ | Высокая |

Под универсальностью понимается возможность применять алгоритмы к различным набором данных, не обладающих специфическими характеристиками, и, не обладая априорной информацией, получать высокую точность классификации.

### Улучшенный метод поиска аномалий

Исходя из характеристик вышеописанных методов, можно сделать вывод о том, что метрические методы поиска аномалий обладают наибольшей универсальностью. Также метрические методы легко совмещать за счёт единой методики измере-

ния показателя аномальности. Можно предложить новый метод поиска аномалий - ансамблирование нескольких метрических методов. Было выбрано три метрических метода - метод K ближайших соседей, метод компонентного коэффициента выбросов[ссылка на литературу по методам]. Для проверки работоспособности алгоритмов поиска аномалий на неразмеченных данных, эти алгоритмы проверялись на размеченных данных. Для этого работа алгоритма поиска аномалий была протестирована на двух наборах данных[ссылка на описание наборов данных]:

Таблица 0.2 — Характеристики датасетов, метрики полноты и точности

| Набор данных | Кол-во элем. | Кол-во атриб. | Полнота | Точн. | Кол-во аном. |
|---|---|---|---|---|---|
| WBC | 453 | 9 | 0.99 | 0.94 | 10 |
| KDDCUP99 | 60853 | 41 | 0.93 | 0.06 | 246 |

Таблица 0.3 — Сравнение алгоритмов поиска аномалий

| Алгоритм | AUC ROC WBC | F1 WBC | AUC ROC KDD | F1 KDD |
|---|---|---|---|---|
| LoOp | 0.98 | 0.72 | 0.68 | 0.05 |
| ODIN | 0.62 | 0.80 | 0.80 | 0.06 |
| KDEOS | 0.25 | 0.64 | 0.61 | 0.05 |
| LDOF | 0.64 | 0.96 | 0.88 | 0.07 |
| INFLO | 0.99 | 0.9 | 0.98 | 0.29 |
| Разр. алгоритм | 0.92 | 0.97 | 0.93 | 0.06 |

Таблица 0.4 — Количество истинно/ложно позитивно/негативно классифицировавшихся

| Набор данных | ИП | ЛП | ИН | ЛН |
|---|---|---|---|---|
| WBC | 10 | 18 | 425 | 0 |
| KDDCUP99 | 230 | 3603 | 57004 | 16 |

Проведем сравнения с другими алгоритмами поиска аномалий. Как можно увидеть из результатов метрик AUC ROC и F1, алгоритмы по-разному классифицируют разные наборы данных.Например, алгоритм LoOP показывает высокий AUC ROC на первом наборе данных, но на втором наборе данных его показали значительно снижаются. В свою очередь, алгоритм ODIN показывает низкие результаты,

по сравнению с остальными алгоритмами, на первом наборе данных, но на втором наборе данных его AUC ROC высок. Разработанной алгоритм показывает средние значения AUC ROC, но высокие значения показателя F1, что позволяет утверждать, что этот алгоритм жизнеспособен и возможно его применение на определенных наборах данных.

### Заключение

Задача поиска аномалий достаточно нетривиальна, а способы её решения могут сильно различаться в зависимости от характектеристик данных и цели поиска. Был предложен новый метод поиска аномалий. Исходя из количества истинно позитивных результатов и ложно позитивных результатов, можно рекомендовать использовать данный метод в задачах где акцент делается на нахождении истинно позитивных значений, пренебрегая некоторым количеством полученных ложно позитивных значений.

## Список использованных источников


1. *MachineLearning.ru Профессиональный информационно-аналитический ресурс, посвященный машинному обучению.* Выборка [Электронный ресурс]. — 2016. URL:http://www.machinelearning.ru/wiki/index.php?title=Vyuborka, дата обращения: 06.04.2018.

2. *Информационно-справочная система онлайн доступа к полному собранию технических нормативно-правовых актов РФ.* ГОСТ 20886-85: Организация данных в системах обработки данных. Термины и определения[Электронный ресурс]. — 2015. URL:http://www.gostrf.com/normadata/1/4294832/4294832686.pdf, дата обращения: 16.04.2018.

3. *Википедия.* Data set [Электронный ресурс]. — 2013. URL:https://en.wikipedia.org/wiki/Dataset, дата обращения: 25.02.2018.

4. *Википедия.* Statistical classification[Электронный ресурс]. — 2010. URL:https://en.wikipedia.org/wiki/Statistical_classification, дата обращения: 25.03.2018.

5. *Панченко, Т.В.* Генетические алгоритмы / Т.В. Панченко. — Издательский дом «Астраханский университет», 2007. — ст. 6-20.

6. *Яминов, Булат.* Генетические алгоритмы[Электронный ресурс]. — 2008. URL:http://rain.ifmo.ru/cat/view.php/theory/unsorted/genetic-2005, дата обращения: 24.03.2018.

7. *Википедия.* JSON[Электронный ресурс]. — 2011. URL:https://en.wikipedia.org/wiki/JSON, дата обращения: 22.04.2018.

8. *А.И Кибзун, Е.Р. Горианова.* Теория вероятности и математическая статистика / Е.Р. Горианова А.И Кибзун. — ФИЗМАЛИТ, 2002. — ст.41.

9. *Entefy.* Data in the digital universe doubles in size every 2 year[Электронный ресурс]. — 2015. URL:https://goo.gl/RbBaE8, дата обращения: 25.03.2018.

10. *Дьяконов, Александр.* Поиск аномалий (Anomaly Detection)[Электронный ресурс] / Александр Дьяконов. — 2017. URL:https://goo.gl/Z43Ne9, дата обращения: 16.04.2018.

11. *F.E., Grubbs.* Procedures for Detecting Outlying Observations in Samples. Technometrics / Grubbs F.E. — American Statistical Association and American Society for Quality, 1969.

12. *Moya M.M., Hush D.R.* Network Constraints and Multi-objective Optimization for One-class Classification. / Hush D.R. Moya M.M. — Journal Neural Networks Volume 9 Issue 3, April 1996, 1996. — p.463-474.

13. *Chandola V Banerjee A, Kumar V.* Anomaly Detection: A Survey / Kumar V. Chandola V, Banerjee A. — ACM Computing, 2009.

14. *Goldstein M, Uchida S.* Behavior Analysis Using Unsupervised Anomaly Detection / Uchida S. Goldstein M. — The 10th Joint Workshop on Machine Perception and Robotics, 2014.

15. *Андрей, Гахов.* Интеллектуальный анализ данных / Гахов Андрей. — Харьковский национальный университет имени В.Н. Карамзина, 2014.

16. *Hodge V., Austin J.* A survey of outlier detection methodologies / Austin J. Hodge V. — Artificial intelligence review, 2004.

17. *Knox, Edwin M.* Algorithms for Mining Distance-Based Outliers in Large Datasets / Edwin M. Knox, Raymond T. Ng. — University of British Columbia, 1998.

18. *S. Bayers, A.Raftery.* Nearest Neighbor Clutter Removal for Estimating Features in Spatial Point Processes / A.Raftery S. Bayers. — Journal of the American Statistical Association June 93, 442;, 1998.

19. *Hautam¨aki, Ville.* Outlier Detection Using k-Nearest Neighbour Graph / Ville Hautam¨aki. — University of Joensuu, Department of Computer Science Joensuu, Finland, 2004. — p. 1-4.

20. *Callahan, P. B.* A decomposition of multidimensional point sets with applications to k-nearestneighbors and n-body potential fie / P. B. Callahan, S. R. Kosara. — s. Journal of the Association for Computing Machin, 1995. — p. 67-90.

21. *Goldstein, Markus.* A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data / Markus Goldstein. — Seiichi Uchida, 2016.

22. *M., Goldstein.* Anomaly Detection in Large Datasets / Goldstein M. — University of Kaiserslauterna, 2014.

23. *Kriegel, Hans-Peter.* LoOP: Local Outlier Probabilities / Hans-Peter Kriegel. — Institut für Informatik, Ludwig-Maximilians Universität München, 2009. — p. 1-3.

24. *Rousseeuw, Leroy.* Robust Regression and Outlier Detection / Leroy Rousseeuw. — John Wiley and Sons, 1996.

25. *B.Chu Chia-Hua Ho, Cheng-Hao Tsa.* Warm Start for Parameter Selection of Linear Classifiers / Cheng-Hao Tsa B.Chu, Chia-Hua Ho. — National Taiwan University, 2015.

26. *Александр, Гущин.* Методы ансамблирования обучающихся алгоритмов / Гущин Александр. — Московский физико-технический институт, 2015.

27. *Лабинцев, Егор.* Метрики в задачах машинного обучения[Электронный ресурс]. — 2011. URL:https://habr.com/company/ods/blog/328372/, дата обращения: 25.03.2018.