Метод поиска аномалий на неразмеченных наборах данных

Аннотация: Проводится сравнительный анализ существующих методов поиска аномалий на неразмеченных наборах данных. Предлагается новый подход, основаный на ансамблировании нескольких метрических методов путем простого голосования показателей аномальности.В результате точность классификации повышается более чем на 30% по сравнению с классическими метрическими методами.Даются рекомендации к использованию нового подхода в зависимости от цели поиска.

Ключевые слова: аномалия, ансамблирование, машинное обучение, метрические методы, бинарная классификация.

1 Введение

Задача поиска аномалий в виде выбросов является одной из классических задач машинного обучения. В настоящее время задачу поиска аномалий активно решают в различных областях деятельности: защита информации и безопастость, социальная сфера и медицина, банковская и финансовая отрасль, распознавание и обработка текста, изображений, речи и многих других.

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

1.1 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 классификатор и, "тестовую выборку" - данные, при анилизе которых классификатор остается неизменным. Тестовая выборка нужна для того чтобы проверить корректность обучения классификатора.

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

1. Обучение с учителем. Для обучения необходимо наличие полностью размеченных данных 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.

2. Обучение с частичным привлечением учителя. Для обучения необходимо наличие тествого и учебного набора данных. Однако в отличие от обучения с привлечением 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.

3. Обучение без учителя. Не требуется разметка 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.

Разметка наборов данных - нетривиальная задача, которая требует отдельного решения. Поэтому использование алгоритмов обучения без учителя является наиболее гибким подходом, который не требует трудозатрат на предварительную разметку данных. В данной работе основное внимание будет уделено именно этому типу методов.

2 Основные классы методов поиска аномалий при обучении без учителя

Задача поиска выбросов является задачей бинарной классификации, где в результе работы алгоритма поиска каждому элементу набора данных присваивается бинарная метка. Бинарная метка - показатель, который принимает нулевое значение в том случае, если она связана с нормальными данными, и единицу в противном случае. Присвоение этой метки происходит на основе анализа оценки достоверности(показателя аномальности) каждого элемента. Оценка показывает вероятность того, что элемент является аномалией. Для разных алгоритмов используются разные шкалы оценок, поэтому приведение конкретных примеров оценок будет некорректным.

2.1 Вероятностно-генеративные методы

Основная идея генеративных методов заключается в использовании вероятностного смесевого моделирования данных. Предлагается подобрать такую вероятностную модель, из которой были получены нормальные данные. Такие модели обычно называются генеративными моделями, где для каждой точки(элемента данных) можно 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:

$$\prod_{x \in X_{norm}} p(x, \theta) \rightarrow max_{\theta} \tag{1}$$

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

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.

2.2 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.

2.3 Метрические методы

Метрические методы пытаются найти в данных точки, в некотором смысле изолированные от остальных[1]. 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 отсутствия 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[2]. 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.

2.4 Сравнение методов

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

Таблица 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
Метрический	$\geq O(nlogn)$	$\geq O(n)$	Высокая

3 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.

3.1 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). [4]

3.2 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

$$b(x) = F(b_1(x), ..., b_T(x)) = \frac{1}{T}\sum_{t=1}^{T} b_t(x) \tag{2}$$

  ,где b_i - некоторая функция.

- Взвешенное голосование

$$b(x) = F(b_1(x), ..., b_T(x)) = \frac{1}{T}\sum_{t=1}^{T} w_t b_t(x) \tag{3}$$

$$\sum_{t=1}^{T} w_t = 1, w_t \geq 0 \tag{4}$$

  ,где w_i- некоторый коэффициент.

- Смесь экспертов

$$b(x) = F(b_1(x), ..., b_T(x)) = \frac{1}{T}\sum_{t=1}^{T} w_t(x) b_t(x) \tag{5}$$

$$\sum_{t=1}^{T} w_t = 1, \forall x \in X \tag{6}$$

  ,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[5].

3.3 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.

4 Результаты работы и обсуждение

Можно улучшить алгоритм поиска аномалий путем используя один из вышеописанных подходов. Например, новый метод будет заключаться в ансамблировании нескольких метрических методов. Было выбрано три метрических метода - метод K ближайших соседей, метод компонентного коэффициента выбросов, метод локального коэффициенты выбросов[6]. 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[3]:

Таблица 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 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

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

Таблица 4: Количество истинно/ложно позитивно/негативно классифицировавшихся

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

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.

5 Заключение

Задача поиска аномалий достаточно нетривиальна, а способы её решения могут сильно различаться в зависимости от характектеристик данных и цели поиска. В отсутствии предварительно размеченных данных применяют алгоритмы обучения без учителя. Наиболее универсальным классом таких методов являются метрические методы. Их легко комбинировать между собой, а также они не требуют наличия априорной информации о данных. Т.е их использование является наиболее универсальным подходом.

Предложен новый метод поиска аномалий заключающийся в ансамблировании простым голосованием трех метрических методов: К ближайших соседей, компонентного коэффициента выбросов, локального коэффициента выбросов. Сравнение нового метода с уже существующими показало выигрыш до 30% по метрикам AUC ROC и F1 на наборах данных с количеством аномалий не более 1% от общего числа элементов и не менее чем 0.1%.

Метод имеет свои особенности, в частности тендецию к нахождению ложно позитивных результатов. Количество ложно позитивных значений может достигать 5%. Однако, количество ложно негативных значений крайне мало(менее 0.1%). Поэтому метод рекомендуется использовать в задачах, где акцент делается на нахождении позитивных значений.

Список литературы

[1] Александр Дьяконов - Поиск аномалий (Anomaly Detection)[Электронный ресурс], https://goo.gl/Z43Ne9/ , 2017.

[2] Ramaswamy, S. Rastogi and Shim K. - A survey of outlier detection methodologies. Artificial intelligence review publisher, 2004.

[3] G. O. Campos, A. Zimek, J. Sander - On the Evaluation of Unsupervised Outlier Detection: Measures, Datasets, and an Empirical Study. Data Mining and Knowledge Discovery 30(4), 2016.

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

[5] Гущин Александр - Методы ансамблирования обучающихся алгоритмов. Московский физико-технический институт, 2015.

[6] Breunig M, Kriegel H. - LOF: Identifying Density-Based Local Outliers. SIGMOD International Conference on Management of Data, 2000.