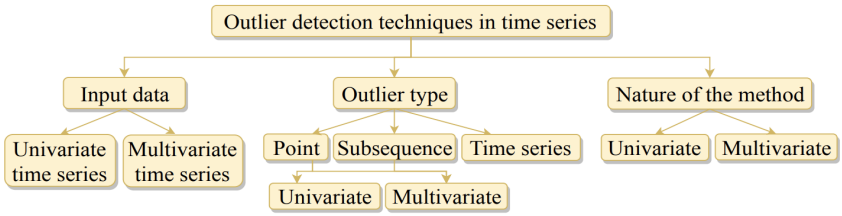
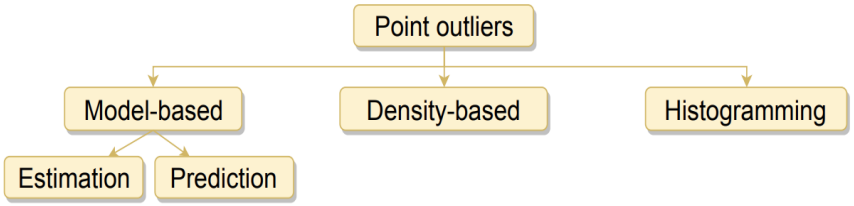
Обзор методов обнаружения аномалий во временных рядах

Всего бывает три вида аномалий во временных рядах: точечные аномалии, аномалии-подпоследовательности и аномалии-ряды (когда существует множество рядов, предположительно похожих). Так же временные ряды бывают как одномерными, так и многомерными. Ниже предоставлен краткий обзор существующих методов. [1]



1 Point outliers

1.1 Одномерный временной ряд.



Самый популярный способ определения точечного выброса это признать точку выбросом, если значение в ней сильно отличается от ожидаемого (прогнозируемого) значения:

$$|x_t - \hat{x_t}| > \tau,$$

где x_t - наблюдаемое значение, $\hat{x_t}$ - ожидаемое значение, τ - порог.

1.1.1 Estimation model-based методы

Оценивание происходит с использованием прошлых, настоящих и будущих данных.

Самое простое оценивание состоит в константных моделях на всех данных либо же на определенном интервале (к примеру, медиана). Так же используют более сложную модель: определяются сегменты различной длины с помощью какой-либо сегментации и используют среднее значение в каждом сегменте для оценивания значения, порог адаптивно выбирается как

$$\tau_i = \alpha\sigma_i$$

где α - выбираемый параметр, σ_i - стандартное отклонение на сегменте i.

Так же используют сглаженные оценки, такие как экспоненциальное взвешенное скользящее среднее, ядерные сглаживания. Другие методы предназначены для выявления маловероятных точек при предположениях о распределении данных. Так можно предположить, что распределение данных без учета выбросов нормально, использовать Gaussian Mixture Models.

Когда модель построена или распределение задано, используют неравенство (выше) для определения выброса.

Другие методы анализируют остатки, полученные из различных моделей (STL decomposition, ARIMA, линейная регрессия и так далее, обучение производится как по прошлым, так и по будущим данным. Хотя эти модели можно использовать и для прогнозирования, выбросы обнаруживаются с учетом и прошлых, и будущих измерений). Далее происходит проверка гипотез, примененная к остаткам. Предполагая, что распределение остатков известно, максимальные и минимальные значения проверяются на каждой итерации. Нулевая гипотеза: данная точка является выбросом, альтернативная: не является. Выбросы удаляются, все повторяется до тех пор, пока не перестанут обнаруживаться выбросы.

1.1.2 Prediction model-based методы

Оценивание происходит только на прошлых данных (прогнозирование), подходит для онлайн поиска аномалий.

Применяются разные модели, к примеру, DeepAnT (на сверточных сетях), авторегрессионные модели, ARIMA. Некоторые модели строят доверительные интервалы, следовательно неявно задают порог τ. Для определения выбросов в онлайне приходится каждый раз переобучать модели.

Еще один подход: теория экстремальных значений. Основанные на этой теории алгоритмы SPOT и DSPOT [2].

Есть методы, для которых не нужно переобучать модель. К примеру, Hierarchical Temporal Memory (HTM) сети.

1.1.3 Плотностной подход

Если во всех методах выше для определения выброса использовалось неравенство, то здесь x_t - выброс тогда и только тогда, когда

$$|\{x \in X | d(x, x_t) \leq R\}| < \tau$$

где d - метрика (Евклидово расстояние), x_t - точка для анализа в момент времени t, X - набор всех точек, $R \in R^+$. Чтобы учесть порядок во временном ряде, используют не все данные для поиска соседей, а только данные в окнах меньшего размера. Если в каком-то окне точка имеет τ соседей, она не является выбросом.

1.1.4 Histogramming методы

Временной ряд можно аппроксимировать гистограммой: разбить на непересекающиеся отрезки времени, все значения ряда в пределах одного такого отрезка заменить одним значением (к примеру, средним). Ошибкой аппроксимации может быть, к примеру, MSE. Очевидно, что чем больше отрезков (бинов), тем ошибка аппрокчимации меньше.

Точки является выбросами, если при их удалении получается гистограмма с меньшей ошибкой аппроксимации, чем с ними, даже если уменьшить количество бинов на количество этих точек.

Пусть у нас B бинов, X - набор всех точек. Тогда $D \subset X$ - набор выбросов, если

$$E_X(H_B^*) > E_{X-D}(H_{B-|D|}^*)$$

где H_B^* - оптимальная гистограмма на X с D бинами, E_X - ошибка аппроксимации, $H_{B-|D|}^*$ - оптимальная гистограмма на $X - D$ с $B - |D|$ бинами.

Для нахождения лучшей гистограммы простым способом является полный перебор, что практически невозможно. В статье [3] описан метод с использованием динамического программирования для более быстрого нахождения таких оптимальных разбиений на бины.

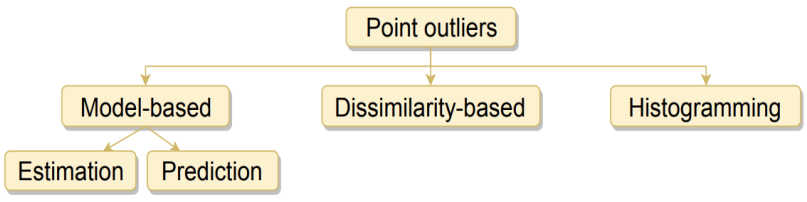
PS: Этот метод я реализовал, работает неплохо (если нет сильного тренда или очень частой периодичности). Однако по времени работает невероятно долго, для одного ряда длины 500 (что достаточно небольшой ряд) будет работать сильно больше часа.

1.2 Многомерный временной ряд.

Первым и самым очевидным подходом является применение всех методов для одномерных временных рядов отдельно к каждой компоненте многомерного временного ряда, это может хорошо работать, однако совсем не учитывает взаимосвязи между компонентами, что приводит к потере информации.

Следующий способ состоит в снижении размерности ряда, поиску меньшего количества не коррелирующих признаков. В таком случае будет уместно применение к этим признакам всех методов для одномерных временных рядов. Методы снижения размерности могут быть различными (различные проекции, PCA и так далее).

Так же можно снижать размерность до одного, тем самым получая одномерный ряд. В качестве одного из таких способов сжатия является вычисление взаимной корреляции между двумя соседними измерениями (x_{t-1} и x_t). Точечные выбросы определяются как точки, имеющие низкую корреляцию с соседними многомерными точками. Порог τ выбирается методом Отцу (алгоритм бинаризации).



1.2.1 Model-based методы

Логика работы этих методов ничем не отличается от одномерных временных рядов.

Для заданного порога τ выбросы определяются как

$$\|x_t - \hat{x_t}\| > \tau$$

где x_t - k-мерное значения временного ряда, $\hat{x_t}$ - ожидаемое значение. Точно как и в одномерном случае, бывают Estimation model-based и Prediction model-based методы.

Для Estimation самыми частыми моделями являются автокодировщики. Так как выбросы соответствуют нерепрезентативным признакам, автокодировщики плохо их восстанавливают. Существуют некоторые усложнения данного подхода (к примеру, вариационные автокодировщики с GRU).

Помимо автокодировщиков в model-based методах можно использовать любую модель для прогноза, работающую с многомерным временным рядом.

1.2.2 Dissimilarity-based методы

Данные методы основаны на вычислении попарной разницы между многомерными точками или их какими-либо представлениями без обучения моделей.

Обычно в таких методах редко используют сами данные, чаще всего используют какое-либо их представление. К примеру, можно представить как граф, узлами которого являются изначальные точки ряда, ребра - похожесть между ними. Идея состоит в применении модели случайного блуждания в графе чтобы определить вершины, которые не похожи на все остальные.

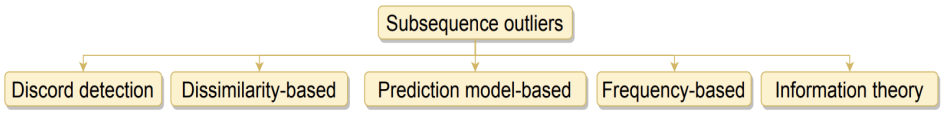
1.2.3 Histogramming методы

Идея точно такая же, как в одномерном случае. Поиск точек, при удалении которых уменьшается ошибка аппроксимации в гистограмме (только теперь в многомерной).

2 Subsequence outliers

Задача - определение подряд идущих точек, поведение которых как-либо отличается от поведения всех остальных. Одной из проблем является то, что у такой последовательности выбросов фиксированная длина. Многие методы могут искать только последовательности с заранее заданной длиной, однако существуют и методы без задания такого гиперпараметра. Так как сравнивать последовательности сильно сложнее (как минимум, по времени), чем точки, то многие методы вместо сравнения подпоследовательностей сравнивают их какие-либо представления.

2.1 Одномерный временной ряд.



Самый очевидный способ состоит в сравнении каждой подпоследовательности с каждой (Discord detection). Таким образом, подпоследовательность D - это выброс в ряде X, если

$$\forall S \in A \min_{D' \in A, D \cap D' = \emptyset} (d(D, D')) > \min_{S' \in A, S \cap S' = \emptyset} (d(S, S'))$$

где A - множество всех подпоследовательностей X (подряд идущих), d - Евклидово расстояние между двумя последовательностями одинаковой длины.

Самый простой поиск - полный перебор, но это очень долго. Но существует эвристический поиск с выбрасыванием ненужных вычислений (HOT-SAX алгоритм).

2.1.1 Dissimilarity-based методы

Идея состоит в сравнении последовательностей с 'нормальной' последовательностью.

Для заданного порога τ последовательность является выбросом, если

$$s(S, \hat{S}) > \tau$$

где S - анализируемая последовательность или ее представление, $\hat{S}$ - ожидаемое 'нормальное' значение.

Для определения этой самой 'нормальности' часто используют кластеризацию, группируя в один кластер все похожие последовательности. Таким образом для каждого объекта кластера его 'нормальное' значение будет центроидом этого кластера. Так же используют Fuzzy C-means кластеризацию [4], которая позволяет объектам принадлежать более чем 1 кластеру.

Для 'нормальной' последовательности можно использовать прошлые данные (если в них нет выбросов), а так же какие-либо внешние данные, если они были сгенерированы с такого же процесса (если в них нет выбросов).

Другой метод позволяет не задавать длину последовательности пользователю. Предполагается, что данные поступают батчами одинаковой длины (B_i это i-ый батч). Уже существующие первые L батчей делятся на M непересекающихся подпоследовательностей не обязательно одинаковой длины: $B_i = S_{i1} \cup S_{i2} \cup ... \cup S_{iM}$ (все батчи разбиваются в одинаковых точках). Кластеризация применяется к наборам $S_{1j}, S_{2j}, ..., S_{Lj}$ для каждого j отдельно. Каждый новый батч разбивается аналогично, используя динамическую кластеризацию кластеризуется, после чего принимается решение, является ли этот кусок выбросом или нет в зависимости от расстояния до центроиды.

Существуют методы без кластеризации. Аналогично написанному выше, можно построить граф и применить случайное блуждание для определения выбросов (теперь узлами будут не точки, а последовательности). В частности, близость последовательностей можно считать с использованием Piecewise Aggregate Pattern Representation (PAPR) [5], что является матричным представлением, в котором заключена статистическая информация о последовательностях.

2.1.2 Prediction model-based методы

Интуиция аналогична Prediction model-based методам из точечных выбросов.

Последовательность $S = x_p, ..., x_{p+n-1}$ - выброс, если

$$\sum_{i=p}^{p+n-1} |x_i - \hat{x_i}| > \tau$$

где $\hat{S}$ - прогнозируемое значение. Все многообразие этого метода заключается в выборе модели, способной прогнозировать на заданный горизонт.

Для определения последовательности-выброса предложено использовать аккумулятор: последовательно предсказывать значения, при нахождении точки-выброса (прогнозируемой точки-выброса; не факт, что она на самом деле выброс) аккумулятор увеличивается на 1, при нахождении "нормальной"точки аккумулятор уменьшается на 1. При достижении аккумулятора какого-то заранее заданного порога (который зависит от длины нужной нам последовательности-выбросов) последовательность точек объявляется выбросом (Важно: эта последовательность не обязательно будет подряд идущей, смысл этого найти какое-то количество выбросов за короткий промежуток времени). Авторы заметили, что в рядах с явно выраженным пиком во времени часто находятся ложные точки (false positive), для этого случая предложено уменьшать аккумулятор на 3 после пиков в данных, а так же возможность аккумулятору уходить в отрицательные значения. [6]

2.1.3 Frequency-based методы

Последовательность S является выбросом, если появляется не так же часто, как ожидается:

$$|f(S) - \hat{f}(S)| > \tau$$

где $f(S)$ - частота встречаемости S, $\hat{f}(S)$ - ожидаемая частота встречаемости S. Так как очень сложно найти хотя бы две полностью совпадающие подпоследовательности в ряду, то применяют дискретизацию при подсчете частотности. Очевидно, чтобы этот метод хорошо работал необходим достаточно длинный временной ряд.

2.1.4 Information theory based методы

Используются для нахождения периодически повторяющихся выбросов - последовательностей. Предположение: часто встречающиеся последовательности несут мало информации, тогда как редко встречающиеся - много. Цель: найти редкую, но все равно повторяющуюся последовательность.

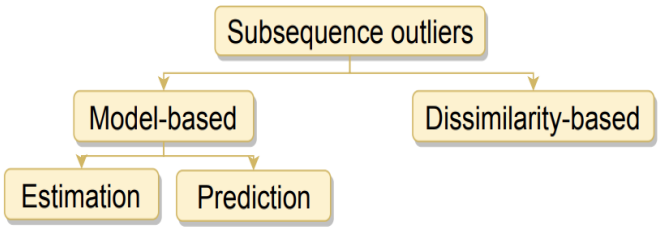
Последовательность S является выбросом, если

$$I(S) * f(S) > \tau$$

где $I(S) \geq 0$ - информация, содержащаяся в S, $f(S) \geq 1$ - количество появления последовательности S в изначальном ряду (дискретизированном). $I(S)$ считается с использованием количества раз встреч значений в S во всем ряду (тоже дискретизированном).

2.2 Многомерный временной ряд.

Аналогично точечным выбросам, можно применять одномерные подходы к каждой компоненте, а так же снижать размерность для получения некоррелирующих признаков.



2.2.1 Model-based методы

Все аналогично предыдущим model-based методам.
Последовательность $S = x_p, ..., x_{p+n-1}$ выброс, если

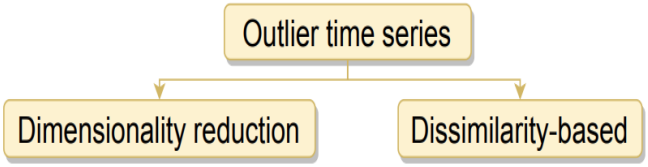
$$\sum_{i=p}^{p+n-1} \|x_i - \hat{x_i}\| > \tau$$

2.2.2 Dissimilarity-based методы

Аналогично одномерному случаю, в отличие от которого в многомерном этот способ практически не используется.

3 Outlier time series

Цель: в многомерном временном ряду найти компоненту, отличающуюся от остальных. (в литературе сильно не анализируются такие выбросы)



3.1 Dimensionality reduction методы

Цель - уменьшить размерность в набор некоррелирующих переменных. К примеру, уменьшить за счет сбора статистических признаков с каждого временного ряда и к этому набору применить PCA. Используя кластеризацию в пространстве после применения PCA можно находить выбросы по расстоянию до центроиды в их классе.

3.1.1 Dissimilarity-based методы

Использование расстояний между рядами. Чаще всего кластеризуют, после чего ищут расстояние до центроиды кластера (центроидой является какой-то ряд). K-means обычно не используют, либо его модификации, либо что-то другое.

Список литературы

[1] A review on outlier/anomaly detection in time series data, https://arxiv.org/pdf/2002.04236.pdf

[2] Anomaly Detection in Streams with Extreme Value Theory, https://www.amossys.fr/upload/anomaly_detection_with_evt.pdf

[3] Mining deviants in a time series database, http://www.vldb.org/conf/1999/P9.pdf

[4] Fuzzy C-Means clustering, https://en.wikipedia.org/wiki/Fuzzy_clustering#Fuzzy_C-means_clustering

[5] A Piecewise Aggregate pattern representation approach for anomaly detection in time series, https://www.sciencedirect.com/science/article/abs/pii/S0950705117303465?via%3Dihub

[6] Time Series Anomaly Detection, https://arxiv.org/ftp/arxiv/papers/1708/1708.03665.pdf