**Анализ моделей и методов решения задач Time Series Anomaly Detection**

**1. Временные ряды и аномалии**

Обнаружение аномалий в данных временных рядов является важной задачей в самых разных областях: от производственных процессов и финансовых приложений до мониторинга здравоохранения. Аномалия может указывать на важные события, такие как производственные сбои, узкие места в доставке, системные дефекты или сбои в работе системы, и поэтому представляет осбый интерес. Поскольку временные ряды часто бывают большими и демонстрируют сложные закономерности, ученые, работающие с данными, разработали различные специализированные алгоритмы для автоматического обнаружения таких аномальных закономерностей. Количество и разнообразие алгоритмов обнаружения аномалий значительно выросло за последние несколько лет.

***Серия данных*** — это упорядоченная последовательность точек данных. Точки данных описывают некоторый объект или свойство процесса на основе непрерывной меры, такой как температура (например, физика), масса (например, химия), угол (например, астрономия), положение (например, геология) или скорость (например, машиностроение). Если порядок основан на времени, последовательность обычно называется временным рядом. Независимо от меры упорядочения, запись точек данных обычно происходит через дискретные, равноотстоящие друг от друга интервалы. По этой причине, а также поскольку большинство алгоритмов анализа рядов данных не зависят от эталонной меры, мы используем термины «ряд данных», «временной ряд» и «последовательность» как синонимы.

Точками данных записи временного ряда являются одна или несколько переменных с действительным значением. Каждая переменная моделирует один канал временного ряда. Если точки данных состоят только из одной переменной, временной ряд называется одномерным; в противном случае – многомерным. Аномалией в таком временном ряду является точка (например, выброс) или последовательность точек (например, нерегулярность), которая отклоняется от некоторой меры, модели или регулярного шаблона последовательности. В многомерных временных рядах отклонения от шаблона могут возникать как в любом отдельном канале, так и в нескольких каналах. Аномальные последовательности могут иметь разную длину и могут вновь появиться в одном и том же временном ряду.

Различные типы аномалий, шаблонные модели и свойства временных рядов привели к разработке множества различных алгоритмов обнаружения аномалий. Многие из них используют очень похожие подходы к обнаружению, но общее разнообразие подходов удивительно велико: от простого обнаружения выбросов, статистического анализа, обработки сигналов и интеллектуального анализа данных до подходов глубокого обучения. Все эти подходы имеют свои сильные и слабые стороны, Выбор подходящего алгоритма зависит от конкретной задачи.

***Обнаружение аномалий временных рядов*** — это процесс маркировки аномалий в заданном временном ряду. Двумя связанными аналитическими задачами являются ***прогнозирование временных рядов*** [65] и ***классификация временных рядов*** [87].

***Прогнозирование*** временных рядов описывает процесс прогнозирования будущего развития временных рядов. Многие подходы к обнаружению аномалий, такие как DeepLSTM [31], Torsk [60], ARIMA [65] и NumentaHTM [3], используют внутреннее прогнозирование временных рядов и определяют аномалии с использованием отклонения предсказанных на основе наблюдаемых значений.

***Классификация*** временных рядов, напротив, описывает процесс отнесения всех временных рядов к определенным классам; его часто используют в качестве этапа постобработки для классификации обнаруженных аномалий по предметно-ориентированным классам, но некоторые алгоритмы обнаружения аномалий, такие как PS-SVM [85], SR-CNN [112], COPOD [80] и NoveltySVR [86], используйте методы классификации, чтобы отнести временные ряды заранее определенным категориям.

**2. Алгоритмы обнаружения аномалий**

В настоящий момент существует множество алгоритмов обнаружения аномалий, которые словно можно классифицировать по следующим признакам (рисунок 1):

***1. По количеству признаков временных рядов:***

- для одномерных временных рядов (один признак);

- для многомерных временных рядов (несколько признаков).

***2. По способу обучения:***

- обучение с учителем;

- обучение без учителя;

- обучение условно с учителем.

***3. По семейству методов:***

- статистические;

- стохастическое обучение;

- обнаружения выбросов;

- анализ сигналов;

- интеллектуальный анализ данных;

- машинное обучение;

- глубокое обучение.

***4. По способам обнаружения аномалий:***

- прогнозирование;

- восстановление;

- измерение расстояния;

- кодирование;

- распределение;

- деревья.

РИСУНОК С КЛАССИФИКАЦИЕЙ

Как видно из рисунка алгоритмы обнаружений аномалий можно разделить

*Таблица 1. Алгоритмы обнаружения аномалий в одномерных временных рядах*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **№ п/п** | **Метод** | **Категория** | **Семейство** | **Способ обучения** |
|  | NoveltySVR | Classic ML | Distance | Unsupervised |
|  | PS-SVM | Classic ML | Distance |
|  | Ensemble GI | Data Mining | Encoding |
|  | HOT SAX | Data Mining | Distance |
|  | TSBitmap | Data Mining | Encoding |
|  | NormA-SJ | Data Mining | Distance |
|  | SAND | Data Mining | Distance |
|  | Series2Graph | Data Mining | Encoding |
|  | Left STAMPi | Data Mining | Distance |
|  | SSA | Data Mining | Distance |
|  | NumentaHTM | Deep Learning | Forecasting |
|  | Sub-LOF | Outlier Detection | Distance |
|  | Sub-IF | Outlier Detection | Trees |
|  | DWT-MLEAD | Signal Analysis | Distribution |
|  | FFT | Signal Analysis | Reconstruction |
|  | SR | Signal Analysis | Reconstruction |
|  | DSPOT | Statistics | Distribution |
|  | ARIMA | Statistics | Forecasting |
|  | MedianMethod | Statistics | Forecasting |
|  | SARIMA | Statistics | Forecasting |
|  | Triple ES | Statistics | Forecasting |
|  | PCI | Statistics | Reconstruction |
|  | RForest | Classic ML | Forecasting | Semi-supervised |
|  | XGBoosting | Classic ML | Forecasting |
|  | TARZAN | Data Mining | Encoding |
|  | HealthESN | Deep Learning | Forecasting |
|  | OceanWNN | Deep Learning | Forecasting |
|  | Bagel | Deep Learning | Reconstruction |
|  | Donut | Deep Learning | Reconstruction |
|  | IE-CAE | Deep Learning | Reconstruction |
|  | SR-CNN | Deep Learning | Reconstruction |
|  | Sub-Fast-MCD | Statistics | Distribution |

*Таблица 2. Алгоритмы обнаружения аномалий в многомерных временных рядах*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **№ п/п** | **Метод** | **Категория** | **Семейство** | **Способ обучения** |
|  | PCC | Classic ML | Reconstruction | Unsupervised |
|  | HBOS | Classic ML | Distance |
|  | k-Means | Classic ML | Distance |
|  | KNN | Classic ML | Distance |
|  | EIF | Classic ML | Trees |
|  | Torsk | Deep Learning | Forecasting |
|  | CBLOF | Outlier Detection | Distance |
|  | COF | Outlier Detection | Distance |
|  | DBStream | Outlier Detection | Distance |
|  | LOF | Outlier Detection | Distance |
|  | COPOD | Outlier Detection | Distribution |
|  | IF-LOF | Outlier Detection | Trees |
|  | iForest | Outlier Detection | Trees |
|  | RobustPCA | Classic ML | Reconstruction | Semi-supervised |
|  | RBForest | Classic ML | Forecasting |
|  | Hybrid KNN | Deep Learning | Distance |
|  | DeepAnT | Deep Learning | Forecasting |
|  | DeepNAP | Deep Learning | Forecasting |
|  | LSTM-AD | Deep Learning | Forecasting |
|  | MTAD-GAT | Deep Learning | Forecasting |
|  | Telemanom | Deep Learning | Forecasting |
|  | MSCRED | Deep Learning | Reconstruction |
|  | AE | Deep Learning | Reconstruction |
|  | DAE | Deep Learning | Reconstruction |
|  | EncDec-AD | Deep Learning | Reconstruction |
|  | LSTM-VAE | Deep Learning | Reconstruction |
|  | OmniAnomaly | Deep Learning | Reconstruction |
|  | TAnoGan | Deep Learning | Reconstruction |
|  | Fast-MCD | Statistics | Distribution |
|  | LaserDBN | Stochastic Learn. | Encoding |
|  | NF | Deep Learning | Distribution | Supervised |
|  | HIF | Outlier Detection | Trees |
|  | MultiHMM | Stochastic Learn. | Encoding |

Рассмотрим подробнее такой классификационный признак как способ обнаружения аномалий.

**Методы прогнозирования**

Методы прогнозирования используют (постоянно) обучаемую модель для прогнозирования количества временных шагов на основе текущего плавающего окна. Значения прогнозируемых точек данных зависят исключительно от точек данных временных рядов в предыдущем плавающем окне и ранее изученной модели. Затем прогнозируемые точки сравниваются с наблюдаемыми значениями в исходном временном ряду, чтобы определить, насколько аномальны наблюдаемые значения. Большинство методов прогнозирования используют скользящее окно с шагом 1 для построения контекстного окна и прогнозирования по одной точке за раз. Методы этого семейства больше всего различаются типом модели прогнозирования, которую они используют (т. е. типом модели нормального поведения), способом построения этой модели (т. е. подходом к обучению) и метрикой расчета оценок аномалий (т. е. метрика расстояния для прогнозируемых и наблюдаемых значений). Представителями этой категории, которые мы рассматриваем в нашей оценке, являются: AD-LTI [148], ARIMA [65], DeepAnT [94], DeepNAP [72], HealthESN [32], LSTM-AD [89], MedianMethod [10], MTAD-GAT [161], NumentaHTM [3], NoveltySVR [86], OceanWNN [143], RBForest [165], RForest [21], SARIMA [52], Telemanom [64], Torsk [60], Triple ES [ 1] и XGBoosting [34].

Обычно методы прогнозирования обучаются условно с учителем: для изучения нормальной модели данных используется обучающий временной ряд без аномалий. Отклонение от этого нормального ожидаемого поведения в тестовом временном ряду, т.е., значительная разница в наблюдаемых и прогнозируемых точках данных, расценивается как аномальная. Это относится к RBForest, RForest, OceanWNN, XGBoosting, AD-LTI, DeepAnT, DeepNAP, HealthESN, LSTM-AD (который может прогнозировать несколько точек), MTAD-GAT и Telemanom. Однако NumentaHTM, NoveltySVR, Torsk (который также может прогнозировать несколько точек), ARIMA и SARIMA строят свою нормальную модель непосредственно на тестовом наборе данных, используя некоторое количество начальных точек (без учителя). Для этих начальных точек методы предполагают нормальность и не рассчитывают оценку аномалии. Модель периодически перестраивается для адаптации к изменениям данных. Обновления модели доводятся до крайности в MedianMethod и Triple ES. Оба алгоритма восстанавливают всю нормальную модель из каждого контекстного окна: MedianMethod использует в качестве прогноза медиану контекстного окна, а Triple ES подгоняет модель тройного экспоненциального сглаживания к каждому контекстному окну для прогнозирования одной последующей точки.

**Методы реконструкции**

Методы реконструкции строят модель нормального поведения путем кодирования подпоследовательностей нормального обучающего временного ряда в (маломерном) скрытом пространстве. Чтобы обнаружить аномалии во временном ряде, подпоследовательности из тестового ряда реконструируются из скрытого пространства, а затем значения реконструированных подпоследовательностей сравниваются с исходными, наблюдаемыми значениями ряда. Входными данными для процесса реконструкции являются обучающие окна (обычно создаваемые с использованием скользящего окна с шагом 1), которые предоставляют модели временной контекст. Поскольку модель построена только на нормальных данных (условно с учителем), аномальные подпоследовательности в серии испытаний не могут быть восстановлены с помощью модели. Следовательно, показатель аномалии можно рассчитать по разнице между исходной и реконструированной подпоследовательностями. Представителями этой категории в нашей оценке являются: AutoEncoder (AE) [117], Bagel [79], DenoisingAutoEncoder (DAE) [117], Donut [150], EncDec-AD [88], FFT [111], Image-embedding. CAE (IE-CAE) [44], LSTM-VAE [106], MSCRED [159], OmniAnomaly [125], PCI [157], PCC [121], RobustPCA [101], SR-CNN [112] и TAnoGan [8].

Исключением из полуконтролируемого обучения этих методов являются четыре неконтролируемых метода FFT, SR, PCC и PCI: они кодируют входные подпоследовательности тестовых серий в заранее определенное скрытое пространство и, таким образом, намеренно теряют информацию, т.е. е., точность, которая необходима для выявления аномалий. В процессе реконструкции не все детали исходных подпоследовательностей могут быть воссозданы. Следовательно, о различиях между реконструированными подпоследовательностями и исходными можно сообщить как оценки аномалий.

**Методы кодирования**

Методы кодирования аналогичны методам восстановления тем, что они также кодируют подпоследовательности временного ряда в скрытом пространстве низкой размерности. Однако они не пытаются реконструировать подпоследовательности из скрытого пространства, а вычисляют оценку аномалии непосредственно из представлений скрытого пространства. Более конкретно, оценки аномалий присваиваются точкам, которые соответствуют закодированным подпоследовательностям в скрытом пространстве. В данной оценке мы рассматриваем следующих представителей этой категории: Ensemble GI [43], GrammarViz [120], LaserDBN [100], MultiHMM [78], PST [128], Series2Graph [16], TARZAN [71] и TSBitmap [144].

GrammarViz и его преемник Ensemble GI дискретизируют подпоследовательности, чтобы затем вывести иерархические грамматические правила; оба алгоритма считают трудно сжимаемые подпоследовательности (низкое покрытие грамматических правил) аномальными. Аналогично, TSBitmap кодирует дискретизированные подпоследовательности как растровые изображения, сохраняющие частоту подпоследовательностей; расстояния между растровыми изображениями ведущего и запаздывающего окна затем используются в качестве оценки аномалии. TARZAN также кодирует частоты дискретизированных подпоследовательностей, но использует суффиксные деревья как для обучающих, так и для тестовых временных рядов (полуконтролируемый); разница между ожидаемой частотой (в результате обучения) подпоследовательности и наблюдаемой частотой используется как показатель аномалии. LaserDBN, PST и MultiHMM строят вероятностные модели и используют логарифмическое правдоподобие подпоследовательностей в качестве оценки аномалии; в то время как MultiHMM строит модель на основе обычных обучающих временных рядов (полуконтролируемых), LaserDBN и PST учитывают только тестовые временные ряды. Series2Graph преобразует подпоследовательности тестовых временных рядов в пространство меньшей размерности, из которого подход строит ориентированный циклический граф. Ребра графа представляют собой переходы между группами подпоследовательностей. Чем чаще временной ряд пересекает ребро, тем выше становится его оценка. Таким образом, ребра с низкими оценками являются более аномальными.

**Дистанционные методы**

Дистанционные методы используют специализированные метрики расстояния для сравнения точек или подпоследовательностей временного ряда друг с другом. Ожидается, что аномальные подпоследовательности будут иметь большие расстояния до других подпоследовательностей, чем подпоследовательности с нормальным поведением. Для расчета расстояний алгоритмы этого семейства могут использовать либо все остальные подпоследовательности, только некоторых ближайших соседей, либо определенные центроиды кластера в качестве опорных точек расстояния. Некоторые подходы перед вычислением расстояний выполняют отображение подпоследовательностей в многомерное пространство. Методы расстояний на основе кластеров группируют похожие подпоследовательности вместе, а затем вычисляют расстояния до плотных областей. Большинство методов в этой категории создают подпоследовательности посредством скользящего окна с шагом 1 во временном ряду теста. Дистанционные методы обычно не требуют обучающих данных и, следовательно, не контролируются. Представителями этой категории в нашей оценке являются: CBLOF [59], COF [130], DBStream [55], HOT SAX [70], Hybrid KNN [124], k-Means [151], KNN [110], LOF [ 22], NormA-SJ [15], PhaseSpace-SVM (PS-SVM) [85], SAND [17], SSA [155], STAMP [156], STOMP [164], Sub-LOF [22], VALMOD [82] и левый STAMPi [156].

Методы ближайшего соседа определяют оценки аномалий путем вычисления расстояния точек (KNN, COF, LOF) или подпоследовательностей (STAMP, STOMP, VALMOD, Left STAMPi, Sub-LOF, HOT SAX, Hybrid KNN) до их ближайших соседей. Нечастые, необычные подпоследовательности расположены на больших расстояниях от своих соседей и поэтому оцениваются как аномальные. Отдельный подкласс этой группы (STAMP, STOMP, VALMOD, Left STAMPi) эффективно вычисляет профиль матрицы, который записывает расстояние каждой подпоследовательности до ее ближайшего несобственного соседа [156, 164]. DBStream и k-Means группируют подпоследовательности, а затем используют расстояния между подпоследовательностями и соответствующими центроидами кластера в качестве оценок аномалий. Аналогично, CBLOF использует для кластеризации многомерные точки вместо подпоследовательностей. NormA-SJ, SAND и SSA создают эталонную модель нормального поведения, с которой сравниваются подпоследовательности. Расстояние между подпоследовательностью и эталонной моделью используется в качестве оценки аномалии. PS-SVM адаптирует SVM одного класса к преобразованному представлению подпоследовательностей и использует обратное расстояние до границы решения в качестве оценки аномалии. Единственный полуконтролируемый подход в этом семействе — гибридный KNN; для построения модели нормальности, с которой сравниваются новые подпоследовательности, требуются нормальные обучающие данные без аномалий.

**Методы распределения**

Методы распределения оценивают распределение данных или адаптируют к данным модель распределения. Распределения рассчитываются либо по точкам данных, либо по подпоследовательностям, полученным с помощью оконного анализа. Хотя сходство точек и подпоследовательностей может быть фактором для подбора распределения (очень похожие шаблоны считаются равными), в этом семействе алгоритмов аномальность оценивается по частоте, а не по расстоянию. Оценки аномалий обычно измеряются с использованием вероятностей, правдоподобий или расстояний между точками или подпоследовательностями т. е. ранее рассчитанные распределения. В общем, это неконтролируемый подход, поскольку аномалии можно обнаружить в крайних точках/хвостах распределений. В полуконтролируемом случае распределение оценивается по обучающему временному ряду, который содержит только нормальное поведение, а затем точки или подпоследовательности тестового временного ряда проверяются на соответствие ранее изученному распределению. Представителями методов распределения в нашей оценке являются: COPOD [80], DWT-MLEAD [134], Fast-MCD [115], HBOS [47], Normalizing Flows (NF) [116], S-H-ESD [62], DSPOT. [122] и Sub-Fast-MCD [115].

DWT-MLEAD, Fast-MCD и Sub-Fast-MCD оценивают распределение Гаусса по временному ряду. После этого аномалия точек или подпоследовательностей измеряется по их расстоянию до среднего значения распределения. Разница в этих подходах заключается в том, что DWT-MLEAD использует дискретное вейвлет-преобразование (DWT) в качестве этапа предварительной обработки и логарифмическое правдоподобие подпоследовательностей в качестве оценки аномалии; С другой стороны, Fast-MCD и Sub-Fast-MCD рассчитывают оценку как расстояние Махаланобиса между точками (Fast-MCD) или подпоследовательностями (Sub-Fast-MCD) и расчетным распределением Гаусса обычного временного ряда обучения. (полуконтролируемый). Другой алгоритм, HBOS, оценивает общее распределение вероятностей подпоследовательностей с помощью гистограмм; затем он использует обратную плотность ячеек гистограммы подпоследовательностей в качестве оценки аномалии. Алгоритм COPOD строит эмпирическую многомерную кумулятивную функцию распределения (связку) для оценки хвостовой вероятности точек; эти хвостовые вероятности затем преобразуются в оценки аномалий. S-H-ESD — это детектор точечных аномалий, который сначала выполняет разложение временных рядов STL, а затем применяет к остаткам тест Граббса, который предполагает распределение Гаусса, чтобы отметить точки-выбросы. Другой детектор точечных аномалий, DSPOT, оценивает обобщенное распределение Парето экстремальных значений временного ряда и применяет пороговое значение к хвостам распределения для обозначения аномальных точек. Наконец, NF — это контролируемый метод обнаружения аномалий последовательности, который преобразует произвольное априорное распределение в распределение Гаусса с помощью нейронной сети; затем он помечает подпоследовательности, попадающие в хвосты распределения, как аномальные.

**Методы изолированного дерева**

Методы изолированного дерева создают ансамбль случайных деревьев, которые разделяют выборки (точки или подпоследовательности) тестовых временных рядов. Для построения дерева методы рекурсивно выбирают случайные признаки и случайные значения разделения в качестве узлов дерева, чтобы в конечном итоге изолировать образцы в листьях дерева. Количество разбиений, необходимых для выделения выборки, представляет собой меру, описываемую средней длиной пути по всем случайным деревьям в ансамбле. Поскольку аномальные образцы легче отделить, чем обычные, они в среднем находятся ближе к корню дерева и имеют заметно более короткие пути. По этой причине длины путей характеризуют нормальность выборок и, следовательно, их обратное значение переводится в оценки аномалий. Представителями этой категории в нашей оценке являются: Extended Isolation Foresсс (EIF) [58], Hybrid Isolation Forest (HIF) [91], Isolation Forest - Local Outlier Factor (IF-LOF) [36], Isolation Forest (iForest) [ 83] и Sub-IF [83].

Описанный выше общий метод дерева изоляции, который также является методом, на котором основаны все алгоритмы этого семейства, представляет собой алгоритм iForest. Контролируемыми вариантами этого подхода являются EIF и HIF. Алгоритм Sub-IF является расширением для обнаружения аномалий подпоследовательностей, т.е. алгоритм, который может обрабатывать последовательности вместо точек, а IF-LOF представляет собой комбинацию iForest [83] и LOF [22].

На рисунке 4 средние значения AUC-ROC алгоритмов связаны с процентом успешно обработанных наборов данных. Следовательно, координата Y точки указывает на надежность оценки AUC-ROC точки. Алгоритмы имеют цветовую маркировку в зависимости от их семейства, а наиболее важные точки отмечены.

Большинство алгоритмов (87%) успешно обработали более 70% наборов данных, а многие алгоритмы (35%) даже обработали более 99% наборов данных. Таким образом, подавляющее большинство сообщаемых измерений качества являются надежными. Например, DWT-MLEAD имеет средний показатель AUC-ROC 0,83 при надежности 100 % и, скорее всего, будет хорошо работать и с другими наборами данных. Напротив, RobustPCA имеет средний показатель AUC-ROC всего 0,54 при надежности 100% и, следовательно, скорее всего, будет плохо работать и с другими наборами данных.

Только 8 алгоритмов имеют надежность ниже 52%, среди которых MultiHMM, S-H-ESD и TAnoGan показали себя плохо, DBStream и EncDec-AD показали посредственные результаты, а HealthESN, NF и LSTM-AD показали отличные результаты. Таким образом, качественные характеристики не особенно репрезентативны, но демонстрируют значительные трудности с настройкой и выполнением: MultiHMM, S-H-ESD и DBStream обнаружили множество внутренних ошибок, которые мы не смогли исправить; NF, HealthESN, EncDec-AD и TAnoGan боролись с ограничением по времени в 4 ℎ; и LSTM-AD часто превышал лимит памяти в 3 𝐺𝐵.

**Список источников**