# 异常检测文档

贾纪东

August 2022

## 1 综述

异常检测的英文是 anomaly detection。这篇文档所写的方法都是针对时间序列的异常检测，具体来说是对服务器 KPI 中的异常点检测所用的方法做一个梳理总结。

### 1.1 服务器 KPI

本文主要检测的服务器 KPI 有 cpu 使用率、内存使用率、发送字节速率和接收字节速率。

对于 KPI 中的数据特征有以下几点，对于稳定运行的服务器来说，首先 KPI 具有周期性；其次 KPI 在同一周期内的波动与用户的使用习惯有着很强的相关性，即和时间有相关性；另外在对于两个周期内的曲线也具有波动性，其波动的模式应该遵循某一分布；第四由于各种原因，周期性也可能被打破，比如节假日或者机器故障或者用户习惯的改变形成了新的周期性；最后采样可能引入噪声，模型中应该避免拟合到这种噪声。

对于 KPI 中异常点检测的难点主要有缺少对于异常点的专家评判标准，难以获得大量异常点数据（大部分数据均为正常数据），不同的波形形态差距较大，很难用一个统一的模型去建模。因此现在大部分无监督的研究都是假设数据集里均为正常点，并去找到正常的曲线模式。

### 1.2 检测主要方法

检测方法主要有传统的基于统计的方法、无监督方法和有监督方法。由于 KPI 异常点标注较为困难，现在大量的研究方向都集中在无监督方法上。

## 2 数据预处理

1. 填补缺失数据点。
2. 时间戳转换，若要将时间作为一个特征则可以利用 one-hot 编码。
3. 序列特征提取：可以对序列加窗；也可以提取序列的一阶差分、滑窗平均以及数据点与前 n 个序列点平均值的差值等。

## 3 统计方法

统计方法简单来说就是针对所要检测的时间序列进行建模，假设时间序列中的值满足某种分布，然后利用所拥有的数据去进行统计估计模型的参数。

这种做法的好处是统计好了以后就可以直接拿来用，进行实时检测比较简单，但是坏处就是检测结果很依赖于模型的假设，这就使得很难在广泛的数据集上得到普遍较好的效果。

接下来简单介绍两种模型：1）GMM；2）facebook：prophet（trend，seasonality，holiday，noise）。第一种是假设 KPI 的值符合高斯分布或者是高斯混合分布，然后进行参数估计，最后认为在 $3\sigma$ 范围之外的点为异常点。这种假设针对波动性不大的序列能够取得比较好的效果，比如说一台运行常规任务的服务器的 cpu。第二个是 facebook 采用的一种算法叫做 prophet，其想法是将 KPI 的波动分解为四种波形，分别是增长的趋势分量、周期性分量、节假日分量（即特殊原因造成的波形）以及噪声分量。算法里对每一种分量都采用了一种模型，然后拿已有数据利用机器学习的方法进行拟合。

## 4 无监督方法

因为对于异常点的判断标准比较困难，并且难以对大规模的数据进行异常点的标注，因此无监督算法成为了现在研究的主流方向。无监督的方法主要分为两类，一种是传统的机器学习中的聚类方法，另外一种是深度学习的方法。

## 4.1 聚类方法

聚类方法的思想是假设了异常的数据点与正常点的数据在特征空间内的分布能够被分开。并且假设了一堆数据中正常的数据占据了绝大多数或者全部都是正常的数据，那么我们如果能划定一个范围将这些数据中比较集中的数据点圈起来，那么在圈外的数据点就认为是异常点。

针对这种方法检测的难点在于特征的选取，究竟选取什么样的特征很重要。这种方法的好处是召回率和准确率都会比较好，缺点也是没有考虑不同时间对于数据点的影响。

可以选取的特征非常多样，并不仅仅局限于原序列的数值。可以是序列的一阶差分、前你个值与当前值的差值、上一个周期同一时刻与现在的差值等。

选取的聚类算法可以有以下几种,one-class-svm,孤立森林,local-outlier-factors。其中 one-class-svm 的问题是会造成误报率升高。

## 4.2 深度生成模型

深度生成模型假设的是神经网络对于正常数据的分布能够重建，而由于训练时并没有很多的异常数据，因此对于异常数据，生成模型的重建误差会比较大。所以利用预测时产生的重建误差，我们可以找到一个重建误差的阈值，认为超过这一阈值的点是异常点。

生成模型主要有考虑时间特征的 bagel 算法，这一算法利用了 CVAE 模型，没有考虑时间特征的 donut 算法，这一算法利用了 VAE 模型。还有利用 RNN 对未来时间序列进行预测并以此作为异常点检测标准的算法。除了这些之外，也有利用 GAN 来进行无监督学习的算法，但是基本思想还是与之前类似。

总的来说基本上大部分的深度生成算法都运用了编码器和解码器的结构。这一结构的根本思想就是认为训练后的神经网络能重建正常数据的分布，但是无法重建异常数据的分布从而导致重建误差的升高。

### 4.2.1 VAE-donut

假设训练集数据均为正常数据，并利用 VAE 去学习到正常数据的分布。

#### 4.2.2 CVAE-bagel

比 VAE 多的一点是将时间也作为输入，时间被编码为 one-hot 编码，前七位代表周几，然后 24 位代表第几个小时，再然后是 60 位代表第几分钟。

#### 4.2.3 RNN

输入输出部分运用了 RNN，这样考虑到了同一滑动窗中的时序特征。

## 5 有监督方法

有监督算法有 Opprentice，主要利用了有标签的数据对模型进行训练。但是由于现实中很难找到大规模的有标签数据，因此这一方案暂时无法投入使用。

## 6 参考资料

[1] https://github.com/yzhao062/anomaly-detection-resources
[2] https://github.com/NetManAIOps

## 7 总结

思想是：具体分析所要监控的序列，获取关于序列波动性，周期性等先验知识。