Time-Series Anomaly Detection Service at Microsoft

2022年10月6日 21:40

《微软公司的时间序列异常检测服务》

摘要：

大公司需要实时监控其应用程序和服务的各种指标。微软开发了一个时间序列异常检测服务，帮客户持续监测时间序列，及时提醒潜在的事件。

本异常检测方法的流水线由三个模块组成：数据提取、实验平台和在线计算。

提出了一种基于光谱残差（SR）和卷积神经网络（CNN）的新型算法。

首次尝试从视觉显著性检测领域借用SR模型来进行时间序列异常检测。

创新性地把SR和CNN结合在一起，以提高SR模型的性能。

SR应该是频谱残差？

一.简介：

异常检测时数据挖掘的一个重要研究领域。准确的异常检测可以触发及时的故障排除。

雅虎发布了EGADS，对雅虎不同属性的数百万时间序列进行自动监测。

设计一个用于时间序列异常检测的工业服务有许多挑战。

1.缺少标签：处理数百万的时间序列，没有简单的方法来手动标记每个时间序列。并且，时间序列的数据分布是不断变化的，可能有的异常情况以前没有出现过，但是系统仍然需要识别。因此监督学习的模型不太合适。

2.泛化：不同业务场景下各种类型的时间序列都需要监测。现有的方法对不同模式没有足够的通用性。

3.效率：在商业应用当中，监控系统必须接近实时的情况下处理数百万甚至数十亿的时间序列。因此效率非常重要，及时时间复杂度大的模型在精度上很好，但是实际情况也不能采纳。

需要开发一种准确、高效和通用的异常检测方法。

传统的统计模型准确度不足，监督模型缺乏标记数据，无监督模型过于耗时&参数敏感。

目标是在无监督的情况下开发一种更有竞争力的方法，同时有利于准确性、效率和通用性。

从视觉显著性检测领域借用了光谱残差模型用于我们的异常检测应用。

频谱/光谱残差是一种高效的无监督算法，首次将该想法用于时间序列异常检测。动机是，时间序列异常检测任务于视觉显著性检测问题，本质上是相似的。显著性是照片或者场景中“突出”的东西（使我们的眼脑快速聚焦），而异常现象出现在时间序列曲线中时，它们也是视觉中最突出的部分。

此外，提出一种基于SR和CNN的组合的新方法。

当有足够多的标记数据时，CNN是一种最先进的监督式显著性检测方法；SR是一种最先进的无监督方法。

通过在SR输出的基础上直接应用CNN联合两大模型。

异常识别的问题在SR模型的输出上容易很多，因此可以自动生成异常来训练CNN，实现比原始SR模型更显著的性能提升。

由于用于CNN训练的异常点是完全合成的，因此SR-CNN还算是无监督的。在没有人工标注数据的情况下建立了最先进的性能。

实验证明该算法比最先进的无监督模型更准确和普遍。

二.系统概述

整个系统由三部分组成：数据采集、实验平台和在线计算。

整个系统的流水线如下图所示。

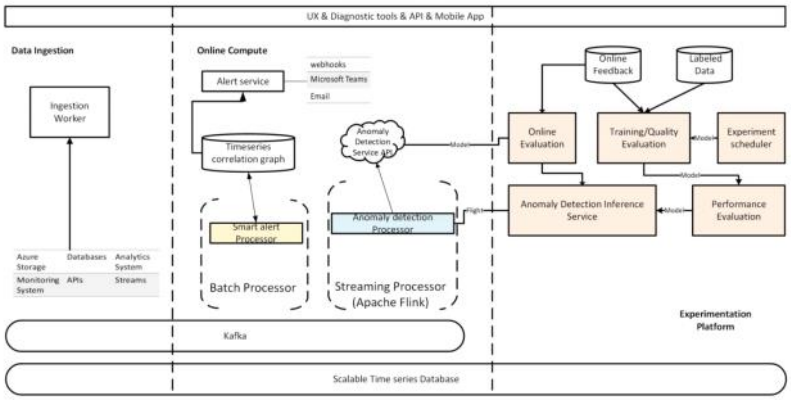


图2：系统概述

用户向系统输入时间序列生成检测任务，提取工作者根据指定粒度更新时间序列（分钟、小时、天），异常检测器在线计算传入时间序列的异常状态……（不太重要）

1.数据提取

用户通过创建一个数据源来注册一个检测任务，数据源由连接字符串和粒度组成。连接字符串用于连接用户的存储系统和异常检测服务，粒度表述数据源更新频率，最小粒度一分钟。

2.在线计算

每个数据点进入管道之后，在线计算模块开始工作，为了检测一个进入点的异常状态，需要一个时间序列数据点的滑动窗口。用Flink3管理内存中的点。

异常检测处理器检测每个单一时间序列的异常情况，但是实际情况中，单一的异常情况不足以有效诊断服务，智能警报处理器会将不同的时间序列异常情况进行关联，生成事件报告，但是本文主要讨论异常检测。

3.实验平台

建立一个实验平台评估异常检测模型的性能，部署新模型之前会进行离线实验和在线A/B测试。

三.应用

在微软，监测业务指标并在有任何超出正常模式的情况下迅速采取行动解决问题是普遍需求。

提供一个服务，以分钟级监测数据源的事件序列，一旦有异常产生就会报告，并且做相应的处理。

四.方法论

时间序列异常检测的问题定义：给定一个实值序列，即x = x1，x2，…，xn，其中xi通常是m维的，任务是输出一个序列，y = y1，y2，…，yn，yi就是0/1代表是否是异常点。实际就是一个二分类问题。
要在没有标记数据的情况下开发出一种通用高效的算法。

受视觉计算领域的启发，采用了光谱残差（SR），一种基于快速傅里叶变换（FFT）的简单而强大的方法。这种方法属于无监督学习，在视觉显著性检测应用中高效。而视觉显著性检测和时间序列异常检测任务本质是相似的，因为异常点在视觉上也是显著点。

当有足够的标记数据可用时，卷积神经网络端到端的模式又是有效的。不过大规模的标记数据还是没法实现的。

所以提出了SR-CNN，将CNN应用于SR的输出，CRR复杂学习一个判别规则，取代原始SR解决方案所采用的单一阈值。可以使用生成的异常标签训练CNN判别器。

1.SR（光谱残差）

SR算法由三个主要步骤组成：（1）傅里叶快速变换得到对数振幅频谱；（2）计算频谱残差；（3）反傅里叶变换，将序列转回空间域。

输入一个时间序列，就可以得到一个显著图。如下图所示。

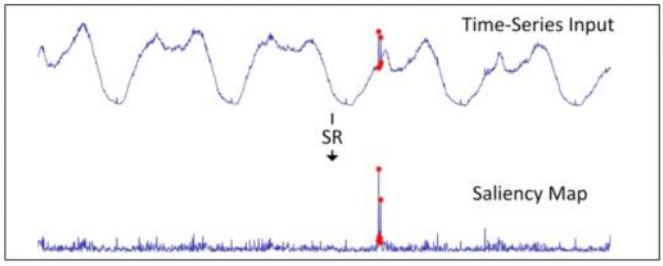


图4：SR模型的结果实例

具体的数学操作没太理解。
显著图中的异常点就比原始输入的异常点突出很多。
基于显著图，可以利用一个简单的规则正确标注异常点，即阈值t判断法。

$$O(x_i) = \begin{cases} 1, & \text{if } \frac{S(x_i) - \overline{S(x_i)}}{\overline{S(x_i)}} > \tau, \\ 0, & \text{otherwise,} \end{cases}$$

xi就是序列中的点，S（xi）就是显著图中的值，—就是xi前z个点的局部平均数（不会和整个序列比较，只是和周围的点比较）。

在实践中，FFT操作是在一个滑动的时间序列窗口内进行的。而且，希望该算法能以较低的延迟发现异常点。就是输入一个时间序列x1~xn，希望能很快地判断xn是不是异常点。但是SR方法在目标点在时间窗口的中央时效果更好。因此在将序列输入到SR模型之前，我们在xn之后添加几个预测点，预测点xn+i的值通过以下方法计算

$$\bar{g} = \frac{1}{m}\sum_{i=1}^{m} g(x_n, x_{n-i})$$

$$x_{n+1} = x_{n-m+1} + \bar{g} \cdot m$$

g（xi，xj）表示xi与xj之间的直线梯度，g—表示前面的平均梯度。m是考虑到前面的点的数量，实际中设定m=5。会发现第一个估计点起着决定性的作用，只需要将xn+1复制k次，添加到序列的尾部。

总之，SR算法只包含几个超参数，滑动窗口大小w，估计点数k，，异常检测阈值t。根据经验设置它们，效果比较好。

2.SR-CNN

原始的SR方法利用显著图上的单一阈值来检测异常点，但是这个规则太简单了，应该下选择更复杂的决策规则。

想法是在精心设计的合成数据上训练一个判别性模型作为异常检测器。

具体来说，就是在时间序列上随机选择几个点，计算注入值替换原来的值，得到显著图。注入值通过以下方式计算：

$$x = (\bar{x} + mean)(1 + var) \cdot r + x$$

x—是前面各点的局部平均值，mean和var是当前时间窗口的所有点的平均数和方差，r是N（0，1）随机采样的，这种就是异常点，其他点是正常点。

选择CNN作为判别模型架构。在实践中，收集带有合成异常的时间序列作为训练数据。

五.实验

1.数据集

用三个数据集评估模型。KPI和雅虎是公共数据集，通常用于评估时间序列异常检测的性能；微软是在生产中收集的内部数据集，涵盖不同时间间隔、广泛的时间序列模式。
异常点被标记为正样本，正常点被标记为负样本。

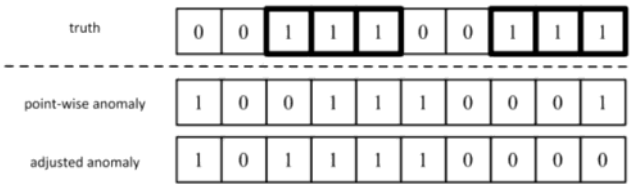
KPI是由AIOPS数据竞赛发布的，该数据集由多条带有异常标签的KPI曲线组成，来自不同的互联网公司，大多数间隔1分钟，一些间隔5分钟。

雅虎的部分时间序列曲线是合成的，另一部分是真实流量。

2.评价指标

从准确度、效率、通用性三方面评价模型。
使用精度、召回率和F1-score表示准确性。
其实在实际应用中，人类操作者不关心精确到点的指标，如果延迟时间不是太长，算法对一个连续异常段的任何一点触发警报都可以接受。因此，将整个连续的异常段标记为阳性样本，只有一个有效的检测计算在内。

评估策略如下图所示。



测试序列有10个连续点，2个异常段。
第二行是预测结果，如果允许延迟一个点，k=1，那么第一个片段就预测正确了，第二个片段就预测错误了。
第三行就是调整后的结果。第一个段的第一个点视为预测出来了，但是第二段的1就相当于没预测出来。
基于调整后的结果，开始计算准确率、召回率、F1-score。
在实际实验中，如果时间序列的间隔是1分钟，那么k=7；如果时间序列的间隔是一小时，k=3；如果时间序列间隔一天，那么k=1；应该遵循实际应用的要求。

效率：在系统中，必须每秒完成数十万次的计算。该模型的延迟足够小，以便它不会阻塞整个计算流水线。在实验中，评估了三个数据集的总执行时间，比较不同异常检测方法的效率。

通用性：一个工业异常检测模型应该有能力处理不同类型的时间序列。为了评估通用性，我们将雅虎数据集中的时间序列手动分为3大类，比较不同类别的F1-score。

3.SR/SR-CNN实验

将SR和SR/CNN与最先进的无监督时间序列异常检测方法进行比较。
基准模型包括FFT、Twitter-AD、Luminol、DONUT、SPOT和DSPOT。
FFT、Twitter-AD、Luminol不需要额外的数据启动，直接输入时间序列就行。
DONUT、SPOT和DSPOT需要数据训练模型，所以把时间序列分两半，前面训练，后面测试。
DONUT可以利用额外的标记数据来提高异常检测的性能。
但是因为目标是完全无监督下比较，公平起见，就不标注了。

实验是在流水线当中进行的，一个时间序列的点依次被提取到评估管道中。每个回合中，只判断最近的点是否是异常点，后面的点不可见。

实验发现，SR的性能明显优于最先进的无监督模型。
此外，SR-CNN在三个数据集上都取得了进一步的改进，说明CNN判别器取代单一阈值的优势。
F1-score远胜其他方法。
CPU总执行时间看，SR还是最有效的方法。

4.SR+DNN

前面的实验已经表明，SR模型在无监督的异常检测中体现出令人信服的结果。
但是有了异常检测标签之后，可以有更令人满意的结果。
采用基于DNN的监督模型，DNN的结构由一个输入层、一个输出层、两个隐藏层组成，在第二个隐藏层之后增加一个dropout层，dropout率为0.5。

效果也挺好的。

六.相关工作

1.异常检测器：

以前的工作可以分为统计、监督和非监督的方法。

统计学文献中提出过几个模型：假设检验、小波分析、SVD和自动回归综合移动平均线（ARIMA）。快速傅里叶变换是另一种用于时间序列处理的传统方法。

……

传统统计模型的性能在实际应用中不令人满意。

采用监督模型，但是标签又不够。

先进的无监督方法。

2.显著性检测方法

受到了视觉显著性检测模型的启发。光谱残差模型（SR）。假设图像可以分为冗余部分和突出部分，人们的视觉对突出部分更加敏感。图像的对数振幅频谱减去平均对数振幅频谱就可以捕捉到图像的显著部分。

然后就是SR+CNN。

七.结论和未来工作

时间序列异常检测是确保在线服务质量的一个关键模块。一个高效、通用、准确的异常检测系统很重要。

总结：目前工业界需要时间序列异常检测的方法，但是目前遇到的挑战：标签工作量大、要求泛化能力和效率高，总结就是要求准确、高效、通用性强。但是统计模型和无监督模型效果都不好。提出了SR+CNN的方法。首先就是通过一系列数学操作（这里没太看懂）把时间序列转换，这时候异常会更为突出。而且这里提到一个很有意思的方法就是通过向后预测点，让最新输入的点可以成为中间点。然后手工把部分点替换成异常点，从而使用CNN。在评价指标里面也有一个比较有意思的就是，允许一定程度的延迟判断。这个方法性能很好。