



IPTV 用户体验异常的自动化检测

谭晓敏, 方艾, 金铎, 李长江

(中国电信股份有限公司广州研究院, 广东 广州 510630)

摘 要: IPTV 系统架构复杂, 涉及大量终端、网元和连接等。为此, 运营商建立了较为完善的监控体系, 特别是采集了海量的 EPG 体验数据, 形成多维度的监控指标, 旨在监控用户体验水平。然而, 监控指标繁多, 导致运维人员监察各项指标费时费力, 无法及时发现异常, 也难以确定异常原因。为解决上述人工运维的痛点, 采用一种智能异常检测方法, 并根据实际应用进行改进, 高效地实现对海量数据的实时分析。实践表明, 该方法计算成本较低, 适应现网异常变化, 快速准确地检测异常, 从而减少人力成本, 提高运维效率, 推进智能运维。

关键词: IPTV; EPG; 自动异常检测; 智能运维

中图分类号: TP393

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2019051

Automated anomaly detection of IPTV user experience

TAN Xiaomin, FANG Ai, JIN Duo, LI Changjiang

Guangzhou Research Institute of China Telecom Co., Ltd., Guangzhou 510630, China

Abstract: Architecture of IPTV system is complex, involving a large number of terminals, network elements and connections. Therefore, a relatively complete monitoring system, which collected massive EPG experience data and formed multi-dimensional monitoring indicators, had been established to monitor user experience. Due to a large number of indicators, manual monitoring was time consuming and laborious. It was hard to detect anomalies in time and it was impossible to determine the cause of abnormality. To solve the pain points of the operation, an intelligent algorithm was implied and improved to analyze massive experience data. The practice indicates that the algorithm with low calculation cost adapts to abnormal changes in the network and detect anomalies accurately and quickly, which reduces labor costs, improves operation efficiency and promotes intelligent operation.

Key words: IPTV, EPG, automated anomaly detection, intelligent operation

1 引言

现阶段, IPTV 已建立了全方位的用户体验监控体系, 监控内容包括视频播放质量、EPG 操控

体验等。用户体验监控体系涵盖多个维度, 如机顶盒、EPG 页面、EPG 服务器以及用户区域等; 形成了多个量化指标, 如 UV/PV、加载时间、操控响应缓慢用户数/次数等。现有的运维方式下,



工程师通过查看各指标数据趋势,从中发现可能存在的异常,并做出判断,例如观察每日 UV/PV 的波动以确认业务是否正常、观察操控响应缓慢次数的增长或下降变化来判断用户体验的好坏等。然而,依赖人工的运维效率低,特别是随着监控指标的不断增加,运维工程师需监控上百项指标,无法满足快速发现和准确定位异常的运维要求。因此,需要引入智能算法,实现自动异常检测。

IPTV 指标以小时粒度的时间序列形式呈现,指标异常的第一时间感知涉及时间序列的异常检测^[1]。时间序列具有严格顺序,传统的异常检测技术分为以下 3 类:基于统计的异常检测方法、基于聚类的异常检测方法和基于预测的异常检测方法^[2]。

基于统计的异常检测方法根据数据偏离统计分布或模型的程度来检测异常,分布的统计属性包括均值、中值、分位数等,如移动平均值算法 MA^[3]和高斯正态分布模型^[4]等。参考文献[4]中,当序列中某时刻的值超出 3 个标准差则视为异常,但由于模型中的残差非正态分布,结果产生许多假阳性异常点。

基于聚类的异常检测方法是观察值分类的过程,偏离聚类中心的数据视为异常,如 K-means^[5]算法等。参考文献[5]利用滑动窗口分割时间序列,提取分割子序列的特征构造三维向量,并采用三维空间的 K-means 聚类方法对时序数据进行聚类,最后将聚类对象与其簇中心的距离作为异常判断的依据。由于异常检测标准无法体现在聚类

过程中,聚类算法偏向于多个时间序列的相似性判断,并不适用于单个时间序列中异常值的检测。

基于预测的异常检测方法通过预测模型预测未来值,若实际值超出预测值的合理范围,则视为异常,如自动回归 AR^[6]、差分自回归移动平均模型 ARIMA^[7-9]、支持向量回归 SVR 等。参考文献[8]采用 ARIMA-SVR 模型预测未来水文数据,准确检测水文时间序列中的异常值。参考文献[9]采用 ARIMA 模型预测网络流量,再采用 SVR 模型校正预测结果,快速检测异常流量。另外,机器学习成为一种新兴的预测方法,参考文献[10]采用多任务最小二乘支持向量机训练历史数据进行时间序列的预测。

2 自动化异常检测

2.1 方案选型

对于时间序列中的异常点检测,通常考虑的方案有基于统计和基于预测的异常检测,如常用的 MA 算法、ARIMA 模型等,表 1 是两种算法的对比。

MA 算法不局限于时间序列的特点,易实现,部署成本较低,生产中异常检测对实时性的要求较高,因此采用 MA 算法作为异常检测算法。

2.2 检测算法

移动平均算法^[11-12]是一种轻量级的时间序列异常检测算法,通过计算滑动窗口内数据(一定长度的时间序列)的平均值,确认序列的移动趋势,设一时间序列 $y_1, y_2, y_3, \dots, y_n$, 滑动窗口 $w=N$, 按照序列顺序依次推移求出平均值,即可得到移动平均值:

表 1 异常检测方案对比

	MA	ARIMA
方法特点	基于统计量,适用于不同类型异常的时间序列	基于预测模型,适用于平稳或差分后平稳的时间序列
技术成本	简单、易实现	引入新数据需重新建立模型
计算成本	滑动窗口遍历序列	平稳性判断、模型识别、参数估计、模型检验
准确性	较强的顽健性;依赖于窗口大小,窗口大小未覆盖周期长度时效果较差	依赖于模型预测能力,序列不平稳时预测效果较差

$$V_t = \frac{y_t + y_{t-1} + \dots + y_{t-N+1}}{N}, t > N \quad (1)$$

其中, V_t 为第 t 时期的移动平均值。

图 1 是某指标的时间序列, 可见指标数据呈现明显的周期波动, 高峰与低谷差异明显。为了展示细节, 对加载时间原始值做了对数处理, 可见移动平均算法对指标的周期规律适应性较强, 其移动平均值曲线反映原始数据的上升变化。

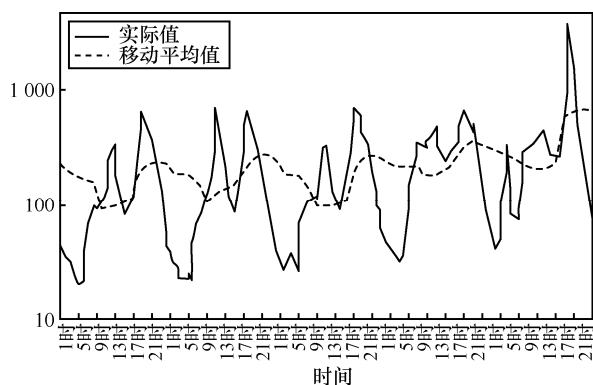


图 1 移动平均值曲线

2.3 方案实现

一般情况下, 指标随着时间的变化呈现一定的规律, 若指标出现不规则波动, 即可找到对应的异常时间点。因此, 计算实际值残差的标准差, 确认实际值的离散程度, 可进一步判断实际值是否异常。异常检测方案实现流程如图 2 所示。

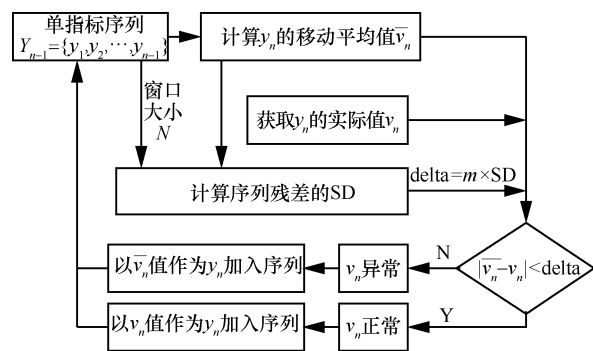


图 2 异常检测方案实现流程

首先, 以一定时间内的正常数据为基础生成时间序列 Y_{n-1} , 设定滑动窗口大小 N 。滑动窗口遍历时间序列, 计算移动平均值 $\bar{v}_n$ 和序列残差的

标准差, 同时设定 m 倍标准差作为指标的正常波动范围 δ 。

然后, 判断当前时刻 y_n 的实际值 v_n 是否落在 $\bar{v}_n \pm \delta$ 区间内, 若是, 则 y_n 为正常点, 否则为异常点。

最后, 引进新数据, 并剔除最早一刻的数据。若 y_n 为正常点, 将实际值 v_n 作为 y_n 时刻的数据加入序列, 若 y_n 为异常点, 将移动平均值 $\bar{v}_n$ 作为 y_n 时刻的数据加入序列, 进行下一时刻 y_{n+1} 的检测。采用移动平均值替代异常值, 能减少异常值对后续时间序列均值的影响, 提高异常检测的准确率。

3 研究与应用

3.1 数据来源

某省 EPG 交互数据采集分析系统基于该省全网用户的体验数据, 以小时为粒度分析存储了 UV/PV、操控响应缓慢用户数/次数、机顶盒加载时间、页面加载时间以及 EPG 服务器加载时间等指标。

选择操控响应缓慢用户数 (指标一) 和机顶盒加载时间 (指标二) 数据进行方案验证。影响用户体验的因素繁多复杂, 除了各节点故障以外, 机顶盒升级、页面改版等也可能引起用户体验的变化。因此, 针对实际异常检测过程中遇到的问题, 应根据实际业务特点不断改进和优化异常检测方案。指标一和指标二的实际数据 (连续 30 天 $\times$ 24 h) 如图 3 所示, 两项指标的数据均具有 24 h 的周期间隔, 存在明显的异常点, 其中指标二还存在上升趋势。

3.2 检测结果

针对图 3 (a) 中缓慢用户数 UV 暴涨的情况, 运维人员排查结论是某台 EPG 服务器故障引起的; 图 3 (b) 中加载时间上升, 是机顶盒升级引起的。异常检测结果如图 4 所示, 黑色圆点表示指标的实际值, 灰色曲线是移动平均值曲线, “ $\times$ ” 表示检测到的异常值。检测结果表明, 基于移动平均算法的异常检测方案能准确地计算序列的波

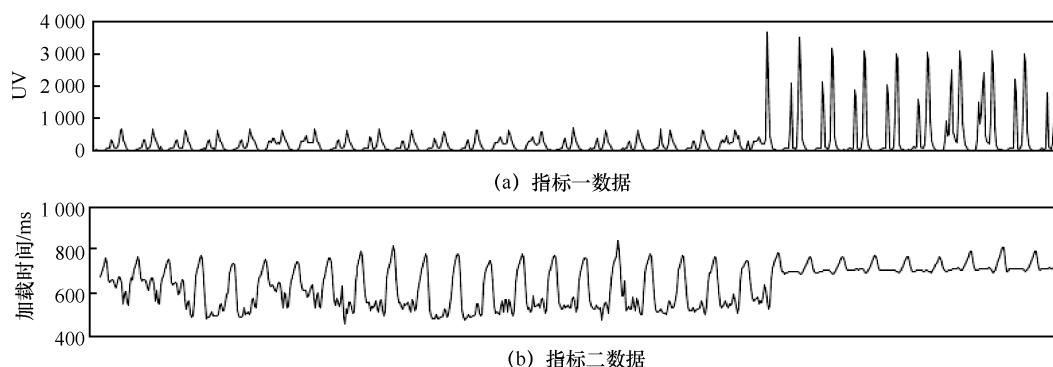


图3 指标一和指标二的时间序列

动情况,成功检测出大部分异常点,尤其是周期规律发生变化的指标二时间序列,机顶盒升级后均视为指标异常。

基于移动平均算法的异常检测方案虽能达到最基本的异常检测要求,但在实际业务背景下,如周期性差异、机顶盒升级等,检测结果存在大量假阳性点(图4椭圆内)。针对不同的业务需求,需对检测方案做出改进与优化。

小时的子序列为基础,作移动平均曲线,分别检测各小时的异常。

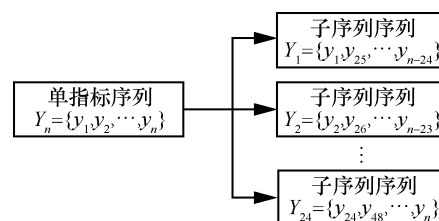


图5 子序列划分示意图

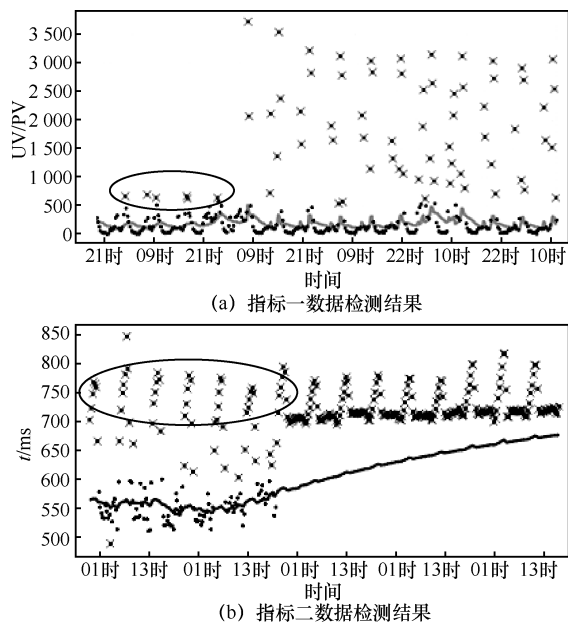


图4 异常检测结果

3.3 周期性差异分析

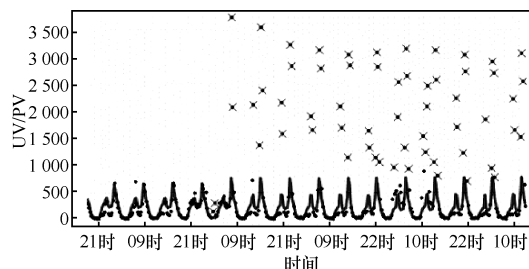
周期性差异导致高峰时段值错被检测为异常值。为了消除指标的周期性差异,将单指标序列按小时划分为24个子序列,如图5所示。以各个

划分子序列的异常检测结果如图6所示,真实异常点的检测效果与原始检测(图4)效果相当,但却成功消除了晚高峰的假阳性异常点。移动平均值走势更接近于实际值,由此可见,按小时划分的子序列更能反映各小时的波动趋势。结果表明,进行划分子序列改进的异常检测方案,有效地消除了实际指标周期性的影响,减少了异常误判。

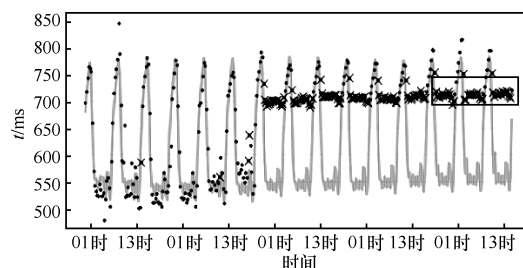
3.4 趋势突变分析

由于机顶盒升级,指标二上升了一个台阶,升级后的数据值均判断为异常,如图6(b)所示。由此可见,正常的业务改版优化引起的趋势变化,是不能简单地采用移动平均值替代“异常值”作为新的数据引入时间序列当中的。

为解决因机顶盒升级、页面改版等引起的趋势变化,加入人工干预,实现半监督异常检测。一旦确认指标数据“异常”是正常业务造成的,则人为地保留实际值,不再采用移动平均值替代实际值加入新序列,待滑动窗口能完全覆盖升级后



(a) 指标一数据检测结果



(b) 指标二数据检测结果

图6 划分序列的异常检测结果

的正常趋势后,再恢复原有的异常检测方式。图7是经过人工干预的异常检测结果。从图7可知,移动平均曲线呈上升趋势,一段时间后,数据值不再被检测为异常。结果表明,加入人工干预可更新序列趋势,解决因正常突变引起的长期异常误判问题。

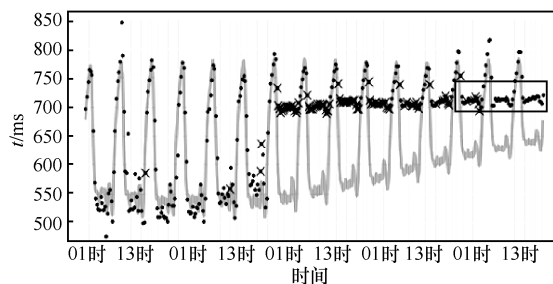


图7 保留实际值的异常检测结果

3.5 参数分析

基于移动平均算法的异常检测方案的核心是定义合适的滑动窗口大小和正常值区间。滑动窗口大小决定了均值走势,滑动窗口较小,子序列受局部波动影响较大,难以反映指标数据长期的波动情况;滑动窗口较大,包含的数据信息量越大,得到的平滑效果越好,但移动平均值对波动则不敏感。正常值区间给出了未来时刻观测值是否正常的取值区间,区间较大,会遗漏部分异常点;区间较小,假阳性点增多。

为了选取合适大小的滑动窗口和合理的正常值区间,设置不同的滑动窗口值和标准差倍数,采用特异性和假负率评估指标,进一步评估检测方案的有效性。特异性指的是异常点被检测为异常点的概率。假负率是正常点被错误检测为异常点的概率。

图8展示了在不同窗口值和不同正常值区间下的异常检测效果。各滑动窗口大小和正常值区间得到的特异性平均值为0.806,说明改进后的异常检测方案能有效检测出80%的异常点。由图8可知,正常值区间设定越窄(倍数 m 越小),此时即使滑动窗口较小,特异性均能维持在0.800以上。但是,滑动窗口越小,假负率越高,说明很多正常值被误检为异常,产生许多假阳性点。因此,滑动窗口和正常值区间的选择需要折中,在保证特异性水平的基础上尽可能降低假负率。滑动窗口大小设置为 $N=30$ 上下较为合适,正常值区间范围选择3倍标准差较为合适。

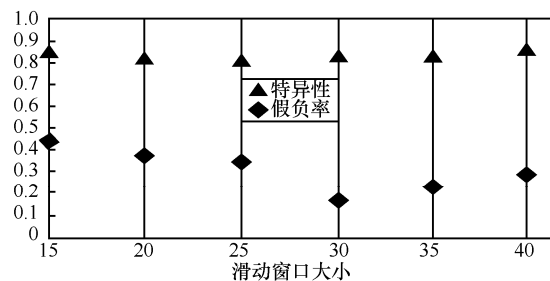
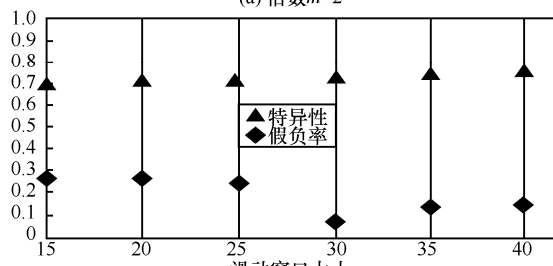
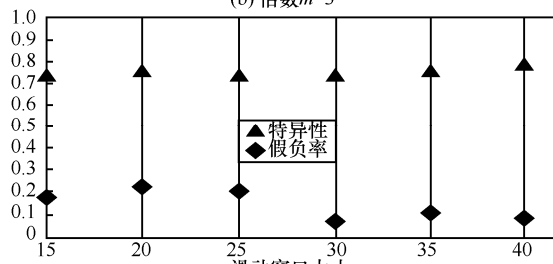
(a) 倍数 $m=2$ (b) 倍数 $m=3$ (c) 倍数 $m=4$

图8 不同参数下异常检测结果



4 结束语

本文将基于移动平均算法的异常检测方案应用到 IPTV 用户体验的异常检测中, 针对实际生产中的主要问题, 如误报、升级导致趋势突变等, 进行了划分子序列、实际值保留的相关改进, 并且通过调参选择合适的参数, 实现对多个指标的实时监控和自动异常检测。自该方案部署以来, 减少了运维人员人工监察的工作量, 多次准确检测出异常指标, 及时地协助运维人员解决问题, 提高了运维效率。

但本文中的异常检测方案处于研发初期, 依然存在异常点漏报、错报等问题, 下一步工作可考虑采用预测模型、监督学习等方案, 减少误报, 减少人工干预, 提高异常检测的准确性。

参考文献:

- [1] 万阳沙. 时序数据的异常检测——大数据运维应用[Z]. 2017. WAN Y S. Anomaly detection of time series data-big data operation and maintenance application[Z]. 2017
- [2] CHANDOLA V, CHEBOLI D, KUMAR V. Detecting anomalies in a time series database[R]. 2009.
- [3] CHOUDHARY P. Introduction to Anomaly Detection[Z]. 2017.
- [4] KIWARE S S. Detection of outliers in time series data[J]. Dissertations & Theses-Gradworks, 2010, 6425(6): 1572-1575.
- [5] 邵开霞, 陈淡泊, 周晓峰. 基于改进的 K -means 聚类方法的多站数据关联异常检测[J]. 微型电脑应用, 2016(11). SHAO K X, CHEN D B, ZHOU X F. Multistation data correlation anomaly detection based on improved K -means clustering method[J]. Microcomputer Applications, 2016(11).
- [6] FUJIMAKI R, YAIRI T, MACHIDA K. An anomaly detection method for spacecraft using relevance vector learning[C]//The 9th Pacific-Asia Conference Knowledge Discovery and Data Mining, PAKDD 2005, May 18-20, 2005, Hanoi, Vietnam. New York: ACM Press, 2005: 785-790.
- [7] MOAYEDI H Z, MASNADI-SHIRAZI M. Arima model for network traffic prediction and anomaly detection[C]//2008 International Symposium on Information Technology, Aug 26-28, 2008, Kuala Lumpur, Malaysia. Piscataway: IEEE Press, 2008: 1.
- [8] 孙建树, 姜渊胜, 陈裕俊. 基于 ARIMA-SVR 的水文时间序列异常值检测[J]. 计算机与数字工程, 2018, 46(2). SUN J S, LOU Y S, CHEN Y J. Outlier detection of hydrological time series based on ARIMA-SVR model[J]. Computer and

Digital Engineering, 2018, 46(2).

- [9] 刘雷杰, 陈雯, 陈聪. 基于 SDN 的网络流量异常检测技术[J]. 计算机科学与应用, 2018, 8(11): 1696-1705. LIU L J, CHEN W, CHEN C. Anomaly detection technology of network traffic in SDN[J]. Computer Science and Application, 2018, 8(11): 1696-1705.
- [10] 夏景明, 唐玲玲, 谈玲, 等. 基于 K -means 和 MTL-SVM 算法的生理参数监测系统[J]. 电信科学, 2017, 33(10): 49-55. XIA J M, TANG L L, TAN L, et al. Biometric monitoring system based on K -means & MTL-SVM algorithm[J]. Telecommunications Science, 2017, 33(10): 49-55.
- [11] 陈光, 白晓颖, 黄晓飞, 等. 使用移动平均线预测云平台服务性能趋势[J]. 计算机科学与探索, 2012, 6(6): 495-503. CHEN G, BAI X Y, HUANG X F, et al. Cloud performance trend prediction by moving averages[J]. Journal of Frontiers of Computer Science & Technology, 2012, 6(6): 495-503.
- [12] 余宇峰, 朱跃龙, 万定生, 等. 基于滑动窗口预测的水文时间序列异常检测[J]. 计算机应用, 2014, 34(8): 2217-2220. YU Y F, ZHU Y L, WAN D S, et al. Time series outlier detection based on sliding window prediction[J]. Journal of Computer Applications, 2014, 34(8): 2217-2220.

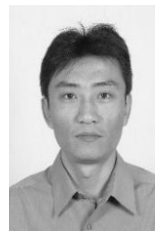
[作者简介]



谭晓敏 (1992—), 女, 中国电信股份有限公司广州研究院工程师, 主要研究方向为移动互联网大数据应用。



方艾 (1981—), 男, 中国电信股份有限公司广州研究院工程师, 主要研究方向为移动互联网与大数据应用、机器学习应用。



金铎 (1970—), 男, 中国电信股份有限公司广州研究院高级工程师, 主要研究方向为移动互联网应用、云计算、大数据分析。

李长江 (1989—), 男, 中国电信股份有限公司广州研究院工程师, 主要研究方向为 IPTV 终端技术。