基于箱线图算法的数据建模在投诉异动实时监控中的应用研究

王玉静，徐宇，徐超

（中移在线服务有限公司山东分公司，山东 济南 250022）

摘要：针对投诉异动点难以及时觉察、导致投诉处理实时性无法有效提升的痛点，本文提出基于箱线图算法建设投诉异动实时监控模型，实现数据分析结果可视化呈现，突破了传统的T+N数据报表分析模式。模型上线后，投诉异动反馈时延较上线前缩短24小时，实现了投诉产生事中控制，降低了投诉处理压力，提升了生产效能。

关键词：投诉异动；箱线图；实时监控；数据可视化

1 引言

随着移动业务的飞速发展，日益严峻的投诉处理压力对投诉问题解决的时效性、妥善性要求越来越高。在此形势下，投诉异动点难以及时发现的问题愈发凸显，严重制约了投诉处理效率的提升。

鉴于此，本文提出基于箱线图算法搭建投诉异动实时监控模型（图1），将数学模型应用于实时投诉数据流，精准定位投诉异动问题，实现投诉异动实时监控、服务策略及时介入、突发投诉事中干预及时化解，突破了传统的数据报表T+N分析模式。

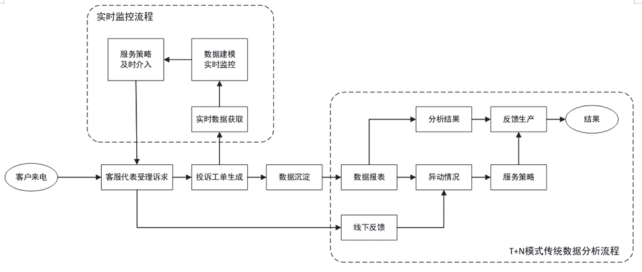

图1 投诉异动实时监控流程与T+N模式传统数据分析流程对比图

2 投诉异动分析模型搭建

2·1 数据预处理

以某运营商客服中心2021年1月的真实投诉数据作为训练集，以2月份的真实投诉数据作为测试集。首先，在训练集中去除数据字段维度不全、内容重复、工单已废弃等无效数据、脏数据；然后，筛选热线人工正常服务时间段数据；接着，将经两步数据预处理后的有效样本数据按小时维度进行分组；最后，抽取数据表中“业务类型名称”字段，对数据维度做进一步拆分细化[1]，得到800余项投诉问题类型作为投诉异动抓取目标问题，并按列表存储。

2·2 箱线图算法原理

训练集数据样本及生产环境数据均为离散的时序数据，具有随机性强、不服从特定分布规律的特征。目前行业内外应用于时序数据异常值检测的成熟算法中，基于正态分布的3σ法则或z分数方法[2]均存在假定数

收稿日期：2021-03-26

作者简介：王玉静（1986—），女，硕士研究生，中级工程师，研究方向为大数据分析建模。E-mail:15966305033@139·com

据服从正态分布的前提条件，但生产环境数据并不服从于正态分布，且上述方法判断异常值的标准是以计算全量样本集数据的均值和标准差为基础，而均值和标准差的抗干扰性极小，所以异常值本身会对算法的性能产生较大影响。以3σ法则为例，σ代表标准差，μ代表均值，认为数值的正常区间为（μ-3σ，μ+3σ），数值分布在此区间的概率为99·7%，即通过此算法得出的异常值个数不会多于样本集总数的0·3%。综合考虑生产环境数据特征与实际生产要求，投诉异常值产生的比例是不可预知的，故这一算法显然不是最优解决方案。

箱线图算法则很好地解决了这一问题。箱线图[3]是一种计算一组数据分散情况的数学算法，也用于反映数据分布特征。其优势一方面在于箱线图算法仅依靠实际数据，而不需要事先假定数据样本集服从特定的分布形式，没有对数据做任何限制性要求，所以能真实、直观地反映数据变化趋势。另一方面，箱线图判断异常值的标准是以四分位数和四分位距为基础，四分位数具有一定抗干扰性，多达25%的数据可以变得任意远而不会在很大程度上干扰四分位数，所以异常值不会对这个标准产生影响，这样使异常数据识别结果相对更加客观。箱线图算法对异常数据的判定原理如图2所示。

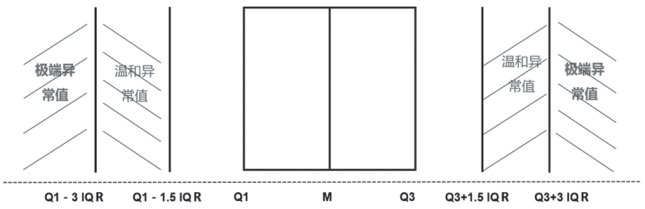

图2 箱线图算法异常数据判定原理

Q1为一组数据的第一四分位点，M为中位数，Q3为第三四分位点，四分位数间距IQ R为Q3与Q1两数之差。当检测值小于Q1-3IQ R或大于Q3+3IQ R时，认为该值属于极端异常值；当检测值小于Q1－1·5IQ R或大于Q3+1·5IQ R，且未达到极端异常值界限时，认为该值属于温和异常值。

2·3 数学模型训练

箱线图算法异常值检测模型计算步骤如下：

（1）数据预处理。首先遍历42万条原始样本数据，按小时维度进行分组，若字段为空则以0填充，剔除非夜间服务时段数据（因为此时间段内产生投诉数量较少或不产生，不具备参考性）；然后将分组数据中同一投诉类型进行合并处理，确定有效的小组样本数据为420个〔n=14（单天正常服务时间段）×30（监测时间周期）〕，将处理后的样本数据按升序排序。

（2）列举观察数据。列出样本数据中最大值MAX、最小值MIN与中位数M，仅与异常数据、中间过程数据进行直观对比，而不参与计算过程。训练集样本中以“网络无法连接”问题为例，最大值MAX=336，最小值MIN=0，中位数M=113·5。

（3）计算分位数值。计算训练集样本数据第一四分位数Q1、第三四分位数Q3，选择四分位数计算方式为训练集样本数据总量n乘以第p分位。若np是小数，选择位置加1的数据Q[int(np)+1]；若np为整数，则按1/2[Qnp + Q(np+1)]计算。按此公式计算训练集样本数据中“网络无法连接”问题数据，第一四分位数Q1=26·5，第三四分位数Q3=145·5。

（4）计算样本数据四分位数间距值。训练集样本数据中“网络无法连接”问题的四分位数间距为IQR=Q3-Q1=145·5-26·5=119。

（5）划定异常值区间范围。按箱线图算法原理规则，训练集样本数据的正常数据值区间范围为(Q1-1·5IQR,Q3+1·5IQR)，温和异常值范围为(Q1-3IQR,Q1-1·5IQR)与(Q3+1·5IQR,Q3+3IQR)，极端异常值范围为小于(Q1-3IQR)或大于(Q3+3IQR)，因此，“网络无法连接”问题的正常值区间范围理论值为(-152,324)。实际上投诉问题数量不会为负数值，所以正常值区间为(0,324)，温和异常值范围为(324,502·5)，极端异常值范围为大于502·5，判定出该类问题存在两个异常值325与336。

2·4 数据模型结果验证

针对数据模型结果进行测试，将测试集数据带入模型进行验证，抓取 800 余项投诉问题维度异常值数据，并将结果用数据表进行呈现。例如，在 2 月 24 日 10:00-11:00 时段抓取突发投诉问题“机顶盒损坏”，同时将抓取结果与实际生产情况进行比对，确认该时段内部分地市突发魔百和故障，导致投诉量短时间突增，表明数据模型结果与应用场景吻合。

3 数据可视化分析应用

投诉异动实时数据监控模型实际应用场景是客服中心话务接续的生产环境作业流程，对实时数据按时间窗口设置进行流计算，针对产生的投诉问题进行异常点实时抓取，并应用数据模型计算将异动情况前端可视化结果展示，后端实时告警、主动反馈。

首先，建立数据归集定时任务 [4]，实时获取增量投诉工单数据，与历史数据关联合并，并通过时间滑动窗口与数据模型规则保持计算样本数据组 n=420 恒定。运用箱线图算法对动态的样本数据进行运算，通过计算结果判定当前时间段 800 余项投诉问题数据是否存在异常。

前端通过数据可视化分析，将冗杂的实时异动计算结果高效提炼出可应用于生产管理的有效信息，按时间段实时更新整体投诉异动情况与详细信息，清晰地将数据模型分析效果推送至业务需求部门。例如，通过数学模型计算 2021 年 3 月 26 日 10：00-11：00 时间段内投诉问题点异动情况，结果显示“营销宣传与实际不符”为异常增加的投诉业务类型，该时间段内投诉量升至 88 例。数据可视化分析效果如图 3 所示。

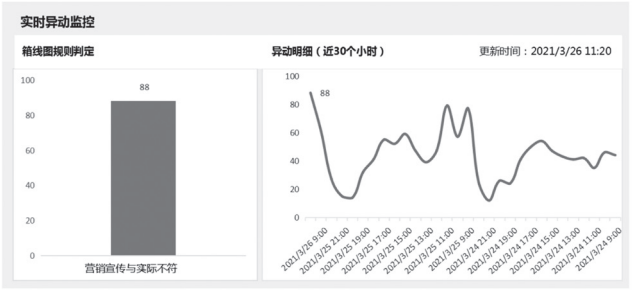

图3 投诉异动实时监控数据可视化

为规避投诉业务处理人员无法及时查看可视化平台的问题，在通过数据分析可视化输出结果的同时，将告警信息以邮件与短信方式同步主动推送至投诉处理部门接口人。

4 结论

投诉异动模型上线并在生产环境应用后，实现了当前时间段投诉异动信息的实时抓取、数据可视化展示与投诉异动信息的实时推送，突破了传统的 T+N 报表分析模式。模型上线后，投诉异动反馈时延较上线前至少缩短 24 小时，大大提升了服务策略介入的及时性，实现了投诉产生事中控制，降低了投诉处理压力，提升了生产效能。

参考文献：

[1] 唐盛涛 · 基于数据挖掘的运营商客户投诉分析方法研究 [J]· 互联网天地 ,2016(3):53-55.
[2] 陈阳 , 凌俊民 , 蒙圣光 · 投诉数据智能挖掘分类管理系统 [J]· 数字技术与应用 ,2011(6):146-149.
[3]GAU R A E I,B R USEY J,ALLEN M,etal·Edge mining the Internet of things[J]·IEEE Sensors Journal,2013,13(10):3816-3825.
[4] 李玮 , 黄秀彬 · 基于层次分析的客服中心实时数据流自动监测方法 [J]· 自动化与仪器仪表 ,2020(1):193-196.