



专题：智能通信技术的应用

移动通信网络投诉热点问题智能预警方法

朱琳, 赵娟, 王伊婷, 冯俊兰, 邓超

(中国移动通信有限公司研究院, 北京 100053)

摘要: 在移动通信网络中, 一旦发生投诉热点问题, 通常会影响大量用户的上网、通话体验, 进而引发大量用户投诉, 即网络投诉热点问题, 该问题因其影响范围大、用户多, 一旦发生往往影响恶劣, 需实时监控并提前预警和处理。提出一种基于用户级信令数据的投诉热点问题预警方法。首先基于对业务逻辑的理解, 选择了 S1 接口数据中与用户体验相关的 30 个关键字段; 然后, 提取 one-hot 特征、统计衍生特征和差分特征 3 类特征来详细刻画用户感知状况; 针对数据中噪声大及正负样本不均衡等问题, 采用泛化能力较强且针对样本不均衡问题有所改善的 LightGBM 分类器实时识别受影响的用户。这一方法可以实时输出潜在受影响的用户与区域, 先于用户投诉提前处理问题或进行客户关怀, 有效降低影响, 提升客户体验。试验结果与某省现网部署应用均验证了算法的有效性。

关键词: 投诉热点问题智能预警; 信令数据; 特征提取; LightGBM 分类器

中图分类号: TN915

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2019099

Smart prediction of the complaint hotspot problem in mobile network

ZHU Lin, ZHAO Juan, WANG Yiting, FENG Junlan, DENG chao

China Mobile Research Institute, Beijing 100053, China

Abstract: In telecom communication network, a hot customer complaint often affects hundreds even thousands of users' service and leads to significant economic losses and bulk complaints. An approach was proposed to predict a customer complaint based on real-time user signaling data. Through analyzing the network business layer logic, 30 key segments related to the user experience in the S1 interface data were selected. Further, one-hot features, statistical derived features, and differential features were extracted to classify user perceptions in detail. Considering the problems of noise data and unbalanced training samples, LightGBM was chosen to train the prediction model. Experiments are conducted to prove the effectiveness and efficiency of this proposal. As of today, this approach has been deployed in our daily business to locate the hot complaint problem scope as well as to report affected users and area.

Key words: prediction of the complaint hotspot problem, signaling data, feature extraction, LightGBM classifier



1 引言

通信网络结构复杂,网元数量庞大,给日常运维管理带来巨大的挑战。在通信网络中一旦发生区域性网络问题,通常会影响大量用户的上网、通话体验,进而引发大量用户投诉,当某个区域或某类业务投诉超过了一定阈值时,客服中心将建立投诉热点问题,该问题因其影响范围大、用户多,一旦发生往往影响恶劣,因此需要尽可能提高处理效率,降低影响范围。实际环境中网络运营部门往往是在接收到投诉热点问题后,才利用专家经验分析网元指标、小区指标及单个投诉用户的信令数据来排查网络问题,然而这种基于投诉信息再处理问题的方式均为后处理方式,效率低,无法在第一时间监控投诉热点问题。因此,亟需发展投诉热点问题自动监控及高效预警方法,先于用户投诉进行处理,减少影响范围。

为了解决投诉热点问题定位和预警问题,最近一些学者和专家基于用户投诉信息或设备 KPI 开展投诉问题分析和故障定位的研究工作,然而现有的投诉热点问题预警存在 3 个难点:用户投诉描述可能不准确,直接利用投诉信息进行分析,会影响问题定位的准确性;客服中心通常在用户体验受影响一段时间后才收到投诉,即客户投诉往往会晚于实际问题发生的时间,因此,基于投诉信息的故障定位方法难以在第一时间发现网络问题;通常投诉热点问题是偶然出现的,因此,难以获取大量正样本数据,会存在明显的数据不平衡问题。为了有效解决上述问题,本文提出一种基于用户信令大数据的投诉热点问题预警方法。首先从 S1 接口中选择与用户体验相关的关键字段;根据网络业务流程的特点,提取了 3 种类型的特征;考虑到样本不均衡的问题,引入权重调整的 LightGBM 分类器来识别受影响的用户。最后,试验证明了本文提出方法的准确性和有效性。

2 相关工作

最近,一些学者和专家研究了基于 KPI 和测量报告^[1-3]或用户投诉信息^[1,4-7]的投诉处理和网络故障定位方法。Chao 等^[1]研究了一种基于核密度估计的网络投诉热点定位自动更新方法,新的投诉热点是根据过去几周内的坐标和客户投诉数量确定的。然后,引入核函数和密度函数来对每个位置的投诉量进行求和。最后,根据 KPI 和信号质量调整热点权重。Tuğba 等^[2]提出了一种基于决策树的面向客户的智能诊断和解决方案模型方法。Zhang 等^[3]介绍了一种基于测量报告的网络主动优化方法。Feyzullah 等^[4]设计了一个基于 LDA 的主题模型来分析客户投诉和投诉主题分布情况。考虑到用户投诉的不准确性,Li 等^[5]基于故障投诉信息创建了农村配电网故障定位算法。Yigit 等^[6]研究了一种主题建模方法来检测客户投诉中的网络问题。Yang 等^[7]结合决策树和手机客户投诉数据、电信设备故障数据来预测设备故障。这些研究主要基于 KPI 与测量报告等小区级性能指标或用户投诉信息两类数据进行问题定位。基于 KPI 与测量报告的方法能够实时监测受影响的区域,但是不能及时定位受影响的潜在用户,无法先于投诉进行用户主动关怀。基于投诉的问题定位方法无法及时定位受影响的区域、排查故障。并且该类方法属于投诉后处理的方式,存在较大时延。此外,用户投诉可能存在描述不准确的问题,会导致投诉问题定位不准确。

针对上述问题,本文提出一种基于信令大数据的投诉热点问题预警方法。实时监测信令数据能够即时判断是否存在发生投诉热点问题的风险、找到投诉热点问题的潜在受影响区域,并输出潜在受影响的用户组,其结果具有较高的准确性和可解释性。该方法能嵌入投诉预处理部分,帮忙运维人员及时进行主动运维及客户关怀。

3 基于信令大数据的投诉热点问题预警方法

一旦网络中发生潜在导致投诉热点问题发生的网络故障或问题,用户的体验往往会受到一定程度的影响,该影响通常会实时反映在用户的信令数据和业务数据流中。以 LTE 网络为例,用户与网络间的信令交互和业务流数据都会通过 S1 接口数据体现,其中, S1-MME 接口提供寻呼、切换、UE 上下文管理、E-RAB 管理以及 MME 和 UE 之间消息的透明传输的功能,主要体现用户和网络间通信信道链路建立情况的体验感知。S1-U 是用户平面接口,用于承载 eNode B 和 SGW 之间终端接收的用户数据流量,主要体现用户在业务传输过程中的体验感知。本文聚焦于 LTE 网络投诉热点问题预警,选用用户 S1-MME 和 S1-U 接口数据进行建模分析实现投诉热点问题的预警。

3.1 预处理

根据网络业务逻辑与用户业务流程的专业知识,本文筛选出 30 个与用户体验相关的关键字段,包括 S1MME 数据 6 个关键字段和 S1U 数据 24 个关键字段,如信令流程编码、流程状态、下行流量、上行 IP 分组数等。其中,共包含 9 个枚举型字段,如流程编码为 1 表示 attach, 2 表示 service request; L4 协议为 1 表示 TCP, 2 表示 UDP。上、下行流量,上、下行 IP 分组数等 21 个字段为数值型字段,如 u32UITraffic 表示从内层 IP 分组头开始计算的数据分组大小总和。然后对全量 S1MME 与 S1U 数据进行数据清洗,通过一致性检查 (consistency check) 去除无效数据、错误数据与重复数据,并对缺失值进行填补。

为能更详细地表征用户的业务流程,将控制面存在业务相关性的字段对进行拼接构成新的字段,如流程编码与流程状态拼接成一个字段后, 10 则表示用户附着成功。最终,共得到 33 个关键字段。

3.2 特征提取

3.2.1 one-hot 特征提取

枚举型字段的值是人为规定的,没有大小、顺序关系,为了防止枚举值的设置对模型训练结果的影响,引入 one-hot 方法将枚举型字段转化为 binary feature (二值特征)。假设某枚举型字段共包含 n 个值,那么 one-hot 处理后可转化为 n 个特征,值对应的特征为 1,其余特征为 0,如图 1 所示。

过程类型	PT_1	PT_2	PT_3	...	PT_n
1	1	0	0	...	0
3	0	0	1	...	0
2	0	1	0	...	0
1	1	0	0	...	0
n	0	0	0	...	1
⋮	...	...	...	...	...

图 1 one-hot 特征

3.2.2 统计特征提取

用户的每一次操作都会产生一条信令并生成一个 XDR 记录,然而因为在通信网络流程中设计了大量信令重试的机制来保证用户体验,单独分析每一条记录意义并不大,例如对正常用户,因终端或特殊情况可能存在某一信令流程 (如 service request) 失败情况,但重试 1~2 次后往往就会成功,并不会影响用户体验,而受热点问题影响的用户往往会出现频繁失败甚至超时的情况,产生多条信令,该情况下才会影响用户体验。考虑到通信网络具有上述特殊性,主要提取用户信令数据某一段时间内的统计特征作为模型的输入,以便更好体现和刻画用户的体验特征,而非将每一条 XDR 记录都作为一个样本。

对于枚举型特征,计算用户在一段时间内每个业务的总数以及信令数作为统计特征,如统计 5 min 内用户 service request 失败次数,这



只需要将 one-hot 向量相加。对于数值型字段，计算一段时间内的最大值、最小值、平均值、标准差、中位数、和值来详细描述用户的业务感知情况，如统计用户 TCP 建链响应时延，相较于不受影响的用户，受影响的用户时延可能会较长，这六维统计特征能够有效区分受影响与不受影响的用户。

3.2.3 差分特征提取

通过上述特征提取过程得到了用户特征的一些时间序列。为了提取用户数据在时间维度的变化趋势，引入一阶差分计算挖掘趋势特征。一阶差分计算式如下所示：

$$\Delta y_t = y(t+1) - y(t), t = 0, 1, 2, \dots, n \quad (1)$$

其中， y 表示用户的连续时序数据，共包含 $n+1$ 个值。考虑到一阶差分后序列中仍然蕴含长期递增的趋势，于是对一阶差分序列再做一次差分运算，得到二阶差分。计算式如下所示：

$$\begin{aligned} \Delta(\Delta y_t) &= \Delta y_{t+1} - \Delta y_t \\ &= (y(t+2) - y(t+1)) - (y(t+1) - y(t)) \end{aligned} \quad (2)$$

对数据进行两次差分后能够较充分地提取原序列中蕴含的长期趋势。通过上述特征提取过程，获得最终用于模型分类的所有特征数据。

3.3 分类器

由于网络问题与信令数据的特殊性，在选择分类模型时面临着两个问题：一是为了帮助专业人员做后续相关分析，需要分类模型具有可解释性，实现对特征重要性的排序；二是原始数据含有一定量噪声，分类模型需要有较强的泛化能力。综合考虑，选择 GBDT (gradient boosting decision tree) [8-10] 作为分类器模型。

GBDT 用作回归时会在每一轮拟合一棵新的决策树，在 GBDT 将要拟合第 t 棵决策树时，它使用前 $t-1$ 棵决策树在所有样本上的回归值的残差作为待拟合值。而当 GBDT 用于分类时，模型会在每一轮同时拟合 K 棵决策树，对应于 K 个分

类。GBDT 模型在第 t 轮拟合时使用样本在前 $t-1$ 轮时所产生的回归值 f_m^{t-1} 的 Softmax 值表示样本属于该分类的概率，如式 (3) 所示：

$$P_m^t(x_i) = \frac{e^{f_m^{t-1}(x_i)}}{\sum_{p=1}^K f_p^{t-1}(x_i)} \quad (3)$$

令样本 x_i 的真实标签表示为一个形如 one-hot 的向量 $[q_{m,i} | m=1, 2, \dots, K]$ ，仅当样本 x_i 属于第 k 类时， $q_{k,i}$ 为 1，那么 GBDT 模型在第 t 轮第 m 棵决策树处拟合的残差为 $\sum_{i=1}^N q_{m,i} - P_m^t(x_i)$ 。

尽管 GBDT 在许多数据集上分类表现优秀，但它仍存在两个问题：一个问题是模型面对海量数据和特征维度很高的数据集时，算法的效率和扩展性很难让人满意。其原因是模型在拟合每一棵决策树的过程中需要进行特征选择和节点分裂，这就需要模型对特征值进行预排序，遍历所有可能的划分点，然后计算信息增益，从而选择出最优的分裂点。模型在每次拟合时要对整个训练集进行遍历，这个计算过程既耗费内存也非常耗时。另一个问题是模型对于不平衡的训练样本分类表现不够优秀。针对这两个问题，引入了一种高效的 GBDT 实现——LightGBM^[11]。该模型在 GBDT 的基础上主要做了两方面的优化：一方面，利用 GOSS 算法对模型训练时的样本点的采样方法进行优化；另一方面，利用 EFB 在选择分裂点时进行特征维度的压缩，这两种优化方法可以让模型在选择分裂点时减少计算量，提高算法的性能。

GOSS 优化是指模型在对某个特征 j 处进行节点分裂时，不再根据信息增益最大的原则对样本遍历找到最佳阈值 d 。模型先将样本根据梯度值排序，然后选择梯度值最大的 $a\%$ 个样本，并且在剩余 $(1-a)\%$ 个样本中随机抽取 b 个样本组成新的训练集去拟合分类器，同时给梯度较小的 b 个样本乘一个较大的系数。分裂节点时信息增益计算量可以被表示为：

$$\tilde{V}_j(d) = \frac{1}{n} \left(\frac{\left(\sum_{x_i \in A_l} g_i + \frac{1-a}{b} \sum_{x_i \in B_l} g_i \right)^2}{n_l^j(d)} + \frac{\left(\sum_{x_i \in A_r} g_i + \frac{1-a}{b} \sum_{x_i \in B_r} g_i \right)^2}{n_r^j(d)} \right) \quad (4)$$

其中, $A_l = \{x_i \in A: x_{ij} \leq d\}$, $A_r = \{x_r \in A: x_{ij} > d\}$, $B_l = \{x_i \in B: x_{ij} \leq d\}$, $B_r = \{x_r \in B: x_{ij} > d\}$ 。这样每次拟合决策树时模型使用非全量数据训练, 既可以从训练数据维度减少计算量, 又增加了每一个决策树的多样性, 提高了泛化能力。

对于稀疏高维数据, LightGBM 可以通过 EFB 算法来绑定互斥特征, 达到减少特征的效果。这意味着直方图建立时的复杂度将会从 $O(\text{data} \times \text{feature})$ 减少到 $O(\text{data} \times \text{bundle})$, 从而加速 LightGBM 的训练过程。

对于第二个问题, 在训练 LightGBM 模型时对正样本权重做了一些调整, 即提高正样本的权重为:

$$\text{weight} = \frac{\text{Negative_Sample}}{\text{Positive_Sample}} \quad (5)$$

总体流程如图 2 所示。

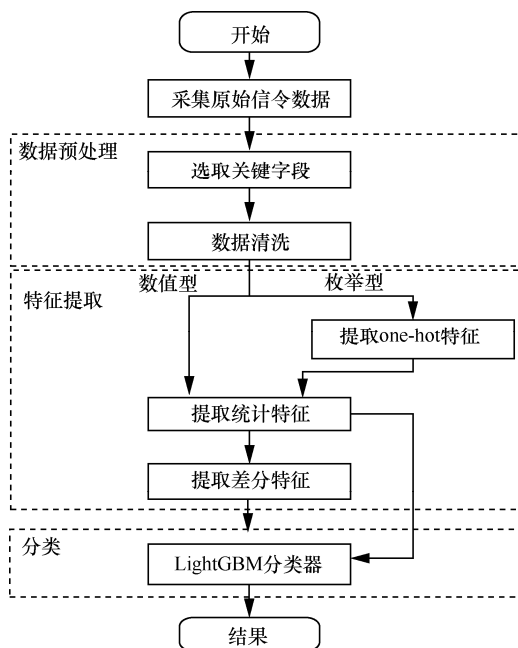


图2 总体流程

4 试验效果分析

为了验证所提出方法的有效性, 本文以大气波导现象导致用户批量投诉的场景为例进行效果验证, 大气波导现象导致信号跨越保护时隙并严重干扰远程基站的上行链路, 这种现象导致 VoLTE 语音通话质量恶化、无法上网或上网速度慢, 从而影响用户体验并导致批量投诉。

为保证模型的顽健性, 即模型更详细全面地刻画受影响用户数据的特征, 选择大气波导时间段不受影响的用户、其他投诉用户、正常用户 3 类用户数据构造负样本集。收集了 702 598 条 S1-MME 接口数据和 2 073 038 条 S1-U 接口数据, 其中包含受大气波导影响的 1 250 565 条数据 (335 314 条 S1-MME 接口数据和 915 251 条 S1-U 接口数据)。根据网络服务逻辑和用户业务流程的专业知识选择 30 个与用户体验相关的关键字段, 如开始时间、结束时间、业务类型、业务状态、请求原因、应用类型代码、应用类型整体、L4 协议、上传流量、下载流量、上传 IP 分组数、下载 IP 分组数等。然后通过一致性检查删除无效数据、错误数据和重复数据。考虑到数字型字段存在缺失值现象, 利用平均值来进行填充。再将存在业务关联的字段对拼接构成新字段, 如业务类型-业务状态、业务类型-请求原因、过程类型-失败原因。

在特征提取阶段, 首先将枚举字段通过 one-hot 转换为 binary feature。然后, 为了提取统计特征, 按时间窗对数据进行切片处理。时间窗宽度的选择基于以下考虑: 大气波导现象通常会在几分钟内影响用户的服务, 如果时间窗过长, 那么统计特征会过于平滑, 使得它们与普通用户的特征一致; 如果时间窗过短, 则统计特征可能无法有效区分投诉热点问题和偶发的小规模网络故障。因此, 窗口宽度选择为 5 min。切



片处理后,提取每个切片的统计特征,包括每个二值特征的和值,每个数字字段的最大值、最小值、平均值、标准差、中位数。图3显示了TCP构造链接确认时间和下行IP分组数的统计特征。可以看出,受影响的用户和普通用户的统计特征具有一定差异。

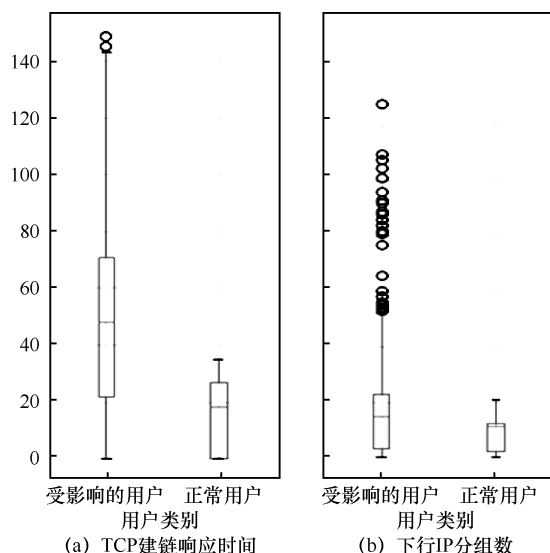


图3 受影响用户与不受影响用户TCP构造链接确认时间和下行IP分组数的统计特征

为全面表征用户感知状况,将S1-MME数据和S1-U数据按时间进行拼接。再提取差分特征,

图4显示了一阶差分和二阶差分后的数据趋势。可以看出,差分特征能够有效提取出相邻时间段的变化趋势。

通过上述步骤,得到 $252\,325 \times 1\,878$ 维特征矩阵,其中,正负样本比例约为1:9,按7:3划分为训练集与测试集。考虑到样本的不均衡性与实际业务需求,准确率不能客观衡量分类算法的优劣,因此选用F1分值进行结果评估。使用十折交叉验证法训练LightGBM分类器,最优结果与平均结果见表1。平均ROC曲线如图5所示,其中“...”代表随机分类情况。多次试验发现,提高负样本权重能够有效提升F1分值。

通过特征重要性计算,得到了全部特征的重要性排序,用户面和控制面最重要的5个特征见表2。

为了进一步评估提出方法的有效性,设置了几组对比实验,结果见表3。可以看出,LightGBM模型的召回率和F1分值均高于其他模型,说明该模型能发现更多的正样本,能够有效避免漏报问题。考虑到现网中的实际业务需求,LightGBM分类器更适用于网络批量投诉问题的预警。

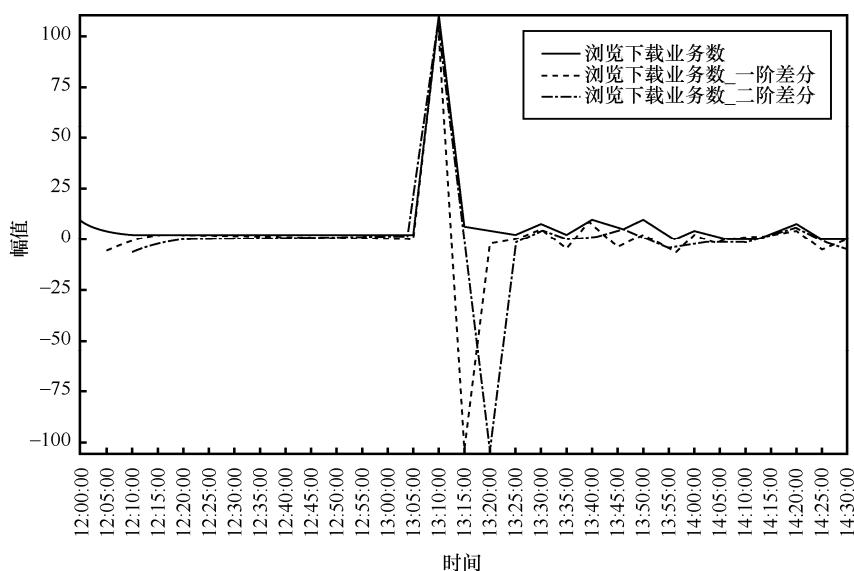


图4 5 min内浏览下载数的一阶差分特征与二阶差分特征

表 1 10 折交叉验证最优结果与平均结果

	精度	召回率	F1 分值
最优结果	83.96%	85.93%	84.93%
平均结果	83.35%	85.78%	84.54%

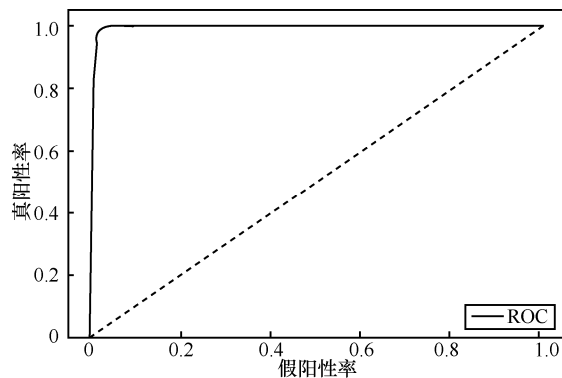


图 5 LightGBM 分类结果的 ROC 曲线

表 2 特征重要性结果

控制面特征	用户面特征
信令数	Spendtime 最大值
Paging 成功次数	windowSize 标准差
E-RAB 释放数	TCPSynNum 最小值
信令数_二阶差分	上行流量最小值
认证失败次数	Spendtime 总时长

表 3 对比实验结果

算法	精度	F1 分值
LightGBM	83.35%	84.54%
决策树	75.71%	71.06%
随机森林	83.25%	78.12%
GBDT	85.35%	80.90%
DNN	75.59%	78.81%
CNN	79.25%	76.83%

目前,提出的方法已经在中国某省现网部署实施,通过将批量投诉预警模型引入投诉预处理环节,实现投诉问题预处理前置,预测时间可压缩至 5 min 内,准确率高达 85%。

5 结束语

本文提出了一种基于机器学习和信令大数据的投诉热点问题预警算法。首先通过预处理原始

S1 接口数据,选择与用户体验相关的关键字段,其次通过提取的 one-hot 特征、分片统计特征与差分特征等信息刻画受影响用户与未受影响用户之间的差异,最后应用 LightGBM 分类器用于实时识别投诉热点问题影响的用户及区域。通过对大气波导现象用户热点投诉问题预警的实验,证明本文提出的方法具有更好的性能,更适合用于投诉热点问题预警。在中国某省现网的部署应用证明了该方法的有效性,能够帮助操作和维护人员第一时间定位问题、输出潜在的受影响区域和用户组,实现先于用户投诉的主动运维和客户关怀,进而有效避免批量投诉。

参考文献:

- [1] CHAO H C, TANG K A, LIU Y H, et al. Using Kernel density estimation to target customer complaint handling service[C]//Asia-Pacific Network Operations and Management Symposium, Sep 27-29, 2017, Seoul, Korea. [S.l.:s.n.], 2017: 215-218.
- [2] TUĞBA K, ÖNEŞ O, ALBEY E, et al. Smart diagnosis and solution systems for network based customer complaints[C]//Signal Processing & Communication Application Conference, May 16-19, 2016, Zonguldak, Turkey. Piscataway: IEEE Press, 2016: 1753-1756.
- [3] ZHANG K, YANG P. WCDMA ratio network optimization approach based on measurement report[C]//IET International Conference on Communication Technology & Application, Oct 14-16, 2011, Beijing, China. Piscataway: IEEE Press, 2011: 379-383.
- [4] FEYZULLAH K, ZEYDAN E, YIGIT I O, et al. A customer complaint analysis tool for mobile network operators[C]//ACM International Conference on Advances in Social Networks Analysis and Mining, August 28-31, 2018, Barcelona, Spain. New York: ACM Press, 2018: 609-612.
- [5] LI Y, LI J, LI W C, et al. Rural distribution network fault location algorithm based on fault complaint information[C]//Third International Conference on Measuring Technology & Mechatronics Automation, Jan 6-7, 2011, Shanghai, China. Piscataway: IEEE Press, 2011: 199-202.
- [6] YIGIT I O, ZEYDAN E, ATES A F. An application for detecting network related problems from call center text data[C]//International Black Sea Conference on Communications



- and Networking, June 5-8, 2017, Istanbul, Turkey. Piscataway: IEEE Press, 2017.
- [7] YANG Q, JI G Q, ZHOU W. The correlation analysis and prediction between mobile phone users complaints and telecom equipment failures under big data environments[C]//2017 2nd International Conference on Advanced Robotics and Mechatronics, August 27-31, 2017, Hefei, China. Piscataway: IEEE Press, 2017: 201-206.
- [8] FRIEDMAN J H. Greedy function approximation: a gradient boosting machine[J]. The Annals of Statistics, 2001, 29(5): 1189-1232.
- [9] FRIEDMAN J H. Stochastic gradient boosting[J]. Computational Statistics & Data Analysis, 2002, 38(4): 367-378.
- [10] 陈晓芳, 张天韵, 张儒中, 等. 基于 GBDT 等数据挖掘算法的场景用户识别方法[J]. 电信科学, 2018, 34(S2): 167-175.
- CHEN X F, ZHANG T Y, ZHANG R S, et al. Scene user identification method based on data mining algorithm such as GBDT[J]. Telecommunications Science, 2018, 34(S2): 167-175.
- [11] MENG Q, KE G, WANG T, et al. A communication-efficient parallel algorithm for decision tree[J]. arXiv: 1611.01276, 2016.

[作者简介]



朱琳 (1983-), 女, 博士, 中国移动通信有限公司研究院人工智能算法高级技术研究员, 主要研究方向为网络智能化运营。



赵娟 (1993-), 女, 现就职于中国移动通信有限公司研究院, 主要研究方向为机器学习。



王伊婷 (1993-), 女, 现就职于中国移动通信有限公司研究院, 主要研究方向为机器学习。



冯俊兰 (1974-), 女, 博士, 中国移动通信有限公司研究院人工智能和智慧运营研发中心总经理、首席科学家, 主要从事大数据挖掘和人际交互领域的研究工作。



邓超 (1978-), 男, 博士, 中国移动通信有限公司研究院人工智能与智慧运营研发中心总经理助理、高级工程师, AIIA 联盟专家委员会委员, 主要从事人工智能、机器学习、大规模并行数据挖掘相关研究工作。