基于机器学习的网络投诉预测分析

万仁辉，王洁，戴鹏程，张旭阳，辛潮

（中国移动通信集团设计院有限公司，北京 100080）

摘　要　减少网络相关的投诉一直是运营商的重点工作之一。目前，网络投诉用户预警方案多以网优工程师经验为主导，准确率和效率都较低。本文通过对历史网络投诉用户数据进行全面深入的分析，基于XGboost算法建立投诉用户特征模型，实现了对网络投诉用户的预测。该方法预测准确率较高，与其它网优系统对接后能够定位用户质差原因，使网络部门能够提前进行网络优化，提升用户满意度。

关键词　机器学习；网络投诉；投诉预测

中图分类号　TP39　　**文献标识码**　A　　**文章编号**　1008-5599（2020）08-0045-06

DOI:10.13992/j.cnki.tetas.2020.08.010

用户投诉处理是运营商服务工作中尤为重要的一个环节，是网络日常维护中不可缺少的部分。通过投诉能够及时了解网络和业务中的不足。以客户感知为导向，可以持续提升网络服务能力，不断优化经营策略，提升工作效率。因此，提高电信服务质量和减少客户投诉将是电信运营商工作重心之一，也是能否领跑其他运营商的关键因素。

为了减少用户的投诉量，一方面可以从运营商自身出发，提高电信产品服务质量；另一方面可以从预测用户投诉行为入手，通过数据挖掘模型事先对高概率投诉用户进行预警，分析其潜在投诉的原因，提前进行预防。电信运营商拥有各类网络质量数据，结合历史投诉用户数据，如何去预警潜在投诉用户是运营商可探索的一项重要工作。

1　网络投诉预测现状

（1）质差用户挖掘法：根据用户感知指标传统的质差门限，筛选质差用户，根据筛选的质差用户与实际投诉的用户进行对比分析，不断调整质差门限，确定最佳质差门限。该方法周期长，涉及到多个感知指标，无法评估指标之间的关联性，此方法潜在投诉用户准确性判定较差，且效率较低。

（2）不满意行为判定法：根据不满意用户得分关联用户异常事件，寻找用户不满意行为与异常事件的关联性，形成不满意用户评价体系，再通过体系挖掘不满意用户作为潜在的投诉用户。该方法难点在于如何寻找不满意行为与异常事件的关联性，实际应用情况来看，该方法过于主观，无法形成合理的不满意用户评价体系，与投诉用户的关联度也较低。

总体来看，现有的投诉用户预警方案多以网优工程师经验为主导，缺乏客观的科学依据，且挖掘出来的潜在投诉用户与实际投诉用户匹配程度较低。机器学习能够较为准确的获取历史投诉用户的特征，因此，本文通过引入机器学习算法提升投诉用户预警效率及准确性，

收稿日期：2020-03-30

提前进行网络优化，最终降低用户网络质量相关的投诉量。

2 基于机器学习的网络投诉预测方法

本文通过对上网类历史网络投诉用户进行深入分析，研究影响用户投诉的各类因素，并通过机器学习建立潜在投诉用户预测模型。对接信令数据、MR 数据和经分数据，生成用户级指标，基于模型对全网潜在投诉用户进行预测，并定位用户质差原因。

2.1 网络投诉用户分析

网络投诉是用户对于网络质量不满的客观反映，因此，可对历史的网络投诉用户进行分析，研究用户的业务感知、投诉原因和网络指标等，初步确定训练指标集，作为机器学习的输入特征。总体来看，投诉行为受客观体验和主观因素影响。

投诉用户客观体验主要包括信号不好、下载速率慢、网页无法打开和视频卡顿等，可归类为信号差和业务体验差两方面，分别用 MR 数据和 XDR 数据进行反映。基于现有网络指标体系对投诉用户和全网用户对比分析，网络投诉平均质差指标个数为 2.5 个左右，而全网用户平均质差数为 1.4 个，网络投诉用户相比于全网用户网络质差较为明显。

投诉用户主观因素主要集中在性格倾向性和行为倾向性两个方面，例如用户对于网络质量的忍受度和是否经常上网等。用户主观因素不确定性较大，可部分用经分数据进行反映。对用户经分属性进行分析可知，投诉用户的男性占比、APRU 值、DOU、星级级别、语音通话时长和通话次数明显高于全网用户。

通过对网络投诉用户的分析，可对用户的信令指标、MR 指标、经分属性进行关联，作为训练样本的特征集，见表 1。

2.2 网络投诉模型建立

机器学习分为有监督学习和无监督学习，本文采用历史投诉用户作为训练数据的标签，是一种有监督学习。投诉用户预测属于离散型预测，因此采用了分类算法。本文采用了性能较好、适合大数据并行计算的 XGboost 算法。在训练阶段，通过对训练数据进行清洗、特征选择、算法调参和模型验证，形成最终参数调优后的模型。

表1 训练样本特征

维度	指标	来源	维度	指标	来源
网络指标	大分组下行速率	信令系统	网络指标	平均 RSRP	MR
	即时通信大文件下行速率	信令系统		弱覆盖采样点占比	MR
	即时通信大文件上行速率	信令系统		平均 SINR	MR
	TCP 一二次握手时延	信令系统		低 SINR 采样点占比	MR
	TCP 二三次握手时延	信令系统	用户信息	性别	经分系统
	TCP 建立时延	信令系统		年龄	经分系统
	HTTP 响应时延	信令系统		用户入网时长	经分系统
	网页响应时延	信令系统		ARPU	经分系统
	视频响应时延	信令系统		DOU	经分系统
	即时通信响应时延	信令系统		星级级别	经分系统
	TCP 乱序率	信令系统		终端厂家	经分系统
	TCP 重传率	信令系统		终端型号	经分系统
	HTTP 响应成功率	信令系统		语音通话时长	经分系统
	网页响应成功率	信令系统		语音通话次数	经分系统
	视频响应成功率	信令系统		套餐饱和度	经分系统

2.2.1 训练数据生成

训练数据由投诉用户指标集和非投诉用户指标集构成。

投诉用户指标集基于用户投诉时间点向前汇聚一段时间内的用户信令数据和 MR 数据，生成用户级信令指标和 MR 指标。同时，获取用户投诉时间的经分属性，并与信令数据和 MR 数据进行关联，形成最终的投诉用户级指标。非投诉用户指标集与投诉用户类似，随机选取同时间段内的非投诉用户，形成非投诉用户指标集。由于投诉用户远少于非投诉用户，可通过对投诉用户进行过采样，使得训练阶段的投诉和非投诉用户样本相对平衡，再进行模型训练时能够提高模型准确性。

2.2.2 数据清洗

针对数据预处理生成的用户级数据，对数据进行清洗，提高样本数据质量。

（1）去除所有行以及所有列大部分为空的数据。

（2）去除缺失率高的特征。

（3）去除数值型且标准差较小的特征。

（4）对缺失值进行填充，采用 0、均值等方式填充。

（5）对于字符型的特征值，采用独热编码进行处理。

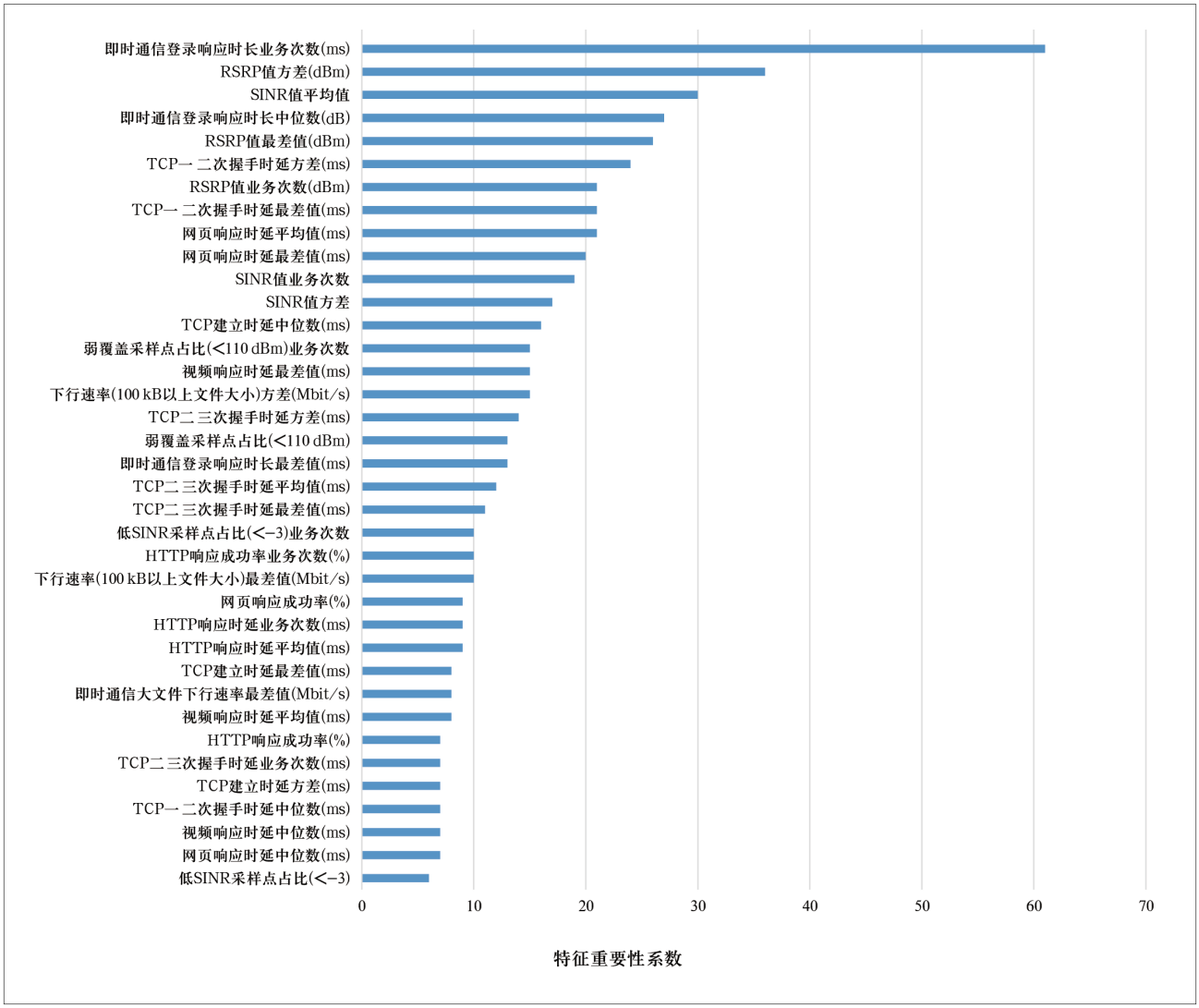

图1 XGboost算法输出的特征重要性系数

对数据进行分析后发现，部分用户无网络质量指标，此类用户级数据将被删除。此外，TCP重传率和乱序率缺失较高，两项特征数据也将会被删除。对于中文字符的经分数据，如星级级别等信息，将通过独热编码进行处理。

2.2.3 特征选择

特征选择过程中，首先需要分析训练数据每列特征的分布，研究每两列特征之间和特征与类别之间的相关性。根据特征的相关系数矩阵，删除高度相关的特征。采用预先训练的模型，XGboost算法可计算出特征重要性系数，评估多种特征重要性系数阈值下的特征子集的模型性能，如图1所示。通过这种方式可识别出影响用户投诉的关键指标，提高建模的速度和精准度，调参后可重复进行特征优化。

2.2.4 算法调参

XGboost算法相关的参数主要分为通用参数、模型参数和学习目标参数3类，其中模型参数是算法优化的重点。下面以XGboost模型参数中的分类器和最大深度为例，阐述调参过程。

对分类器数目进行调参时，以LogLoss作为评估准则，分析不同分类器数目下的LogLoss值。如图2所示，分类器数目在200个左右获得模型调优的最大值，即LogLoss最小。XGboost算法的基分类器是树模型，树的最大深度越大，模型越复杂，越能学到更具体更局部的样本，但是容易过拟合。以实际投诉训练情况来看，树的最大深度在7附近获得模型调优的最大值。

2.2.5 算法验证

机器学习分类模型的评估度量准则包括模型准确率、精确率（查准率）和召回率（查全率）等。机器学习调参的目的是尽可能地让模型准确，使得准确率、精确率和召回率都越高越好。由于投诉用户与非投诉用户

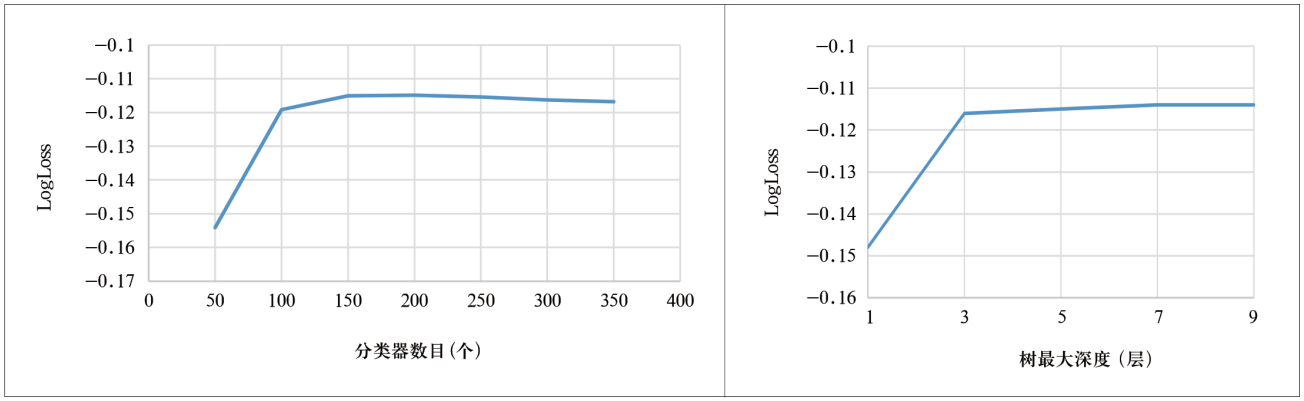

图2 XGboost算法分类器及深度调优

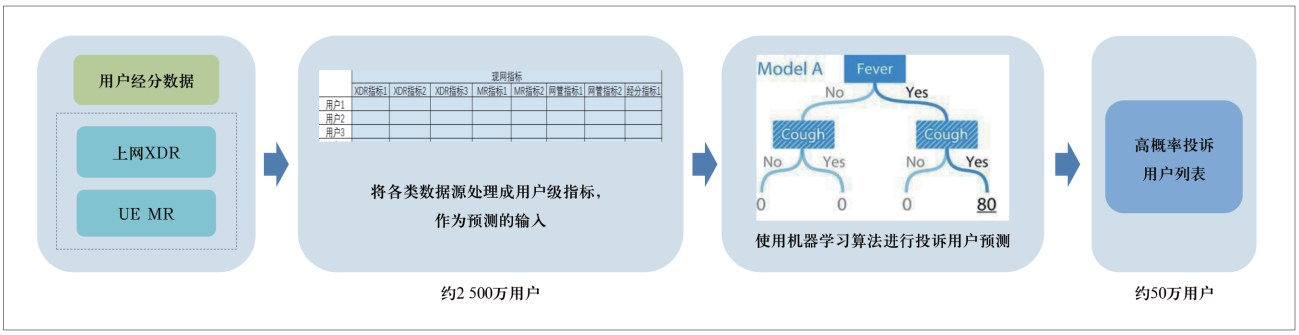

图3 投诉预测现网用户数据处理流程

分布非常不均衡，因此，为保证对潜在投诉用户的识别，需要更加偏向召回率的优化。

对 5 026 条投诉用户数据和 52 395 条非投诉用户数进行训练，经过数据清洗、特征选择和参数调优后，训练准确率达到 93%，精确率达到 56%，召回率达到 42%。随着投诉用户数据的累积，训练结果的精确率和召回率也有一定的提升。

2.3 应用效果

对投诉用户模型调优后，在网络进行了模型部署。采集某省全量用户 S1−U 接口 XDR 话单、MRO 和经分数据，其中 XDR 每天约 450 亿条、MRO 约 200 亿条。基于 MapReduce 对全量 XDR 和 MRO 进行处理，生成用户质量指标，并与经分数据关联形成最终约 2 500 万条用户级数据。最后，利用机器学习模型进行预测，输出约 50 万用户高概率投诉用户，如图 3 所示。

全网用户和实际投诉用户的预测概率分布如图 4 所示，可以发现实际投诉用户与预测结果匹配度较高，70% 以上的实际投诉用户预测概率在 0.5 以上，而 80%

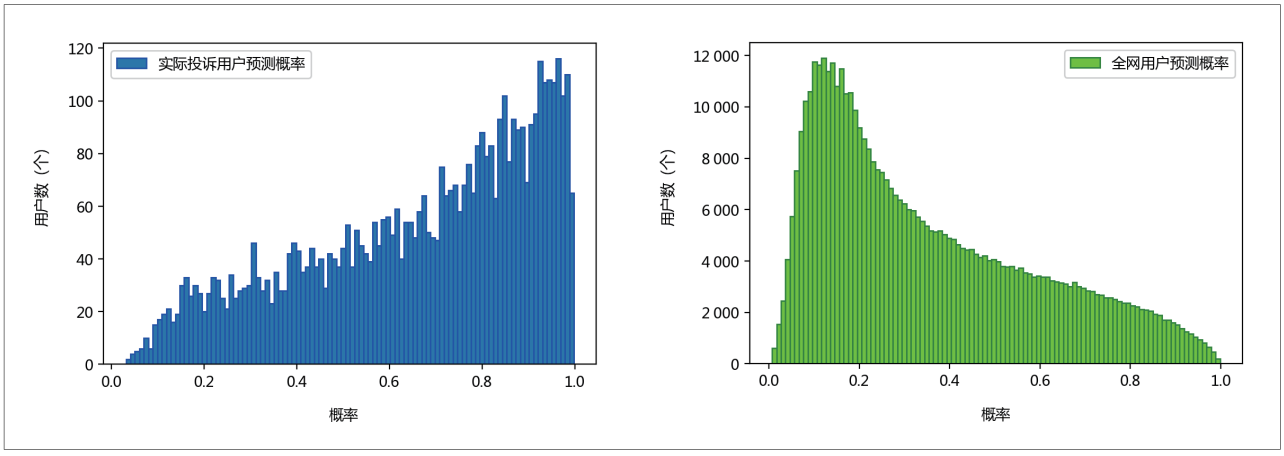

图4 全网用户和实际投诉用户的预测概率分布

图5 潜在投诉用户挖掘及预警平台

的全网用户投诉概率在0.5以下。

同时，与省内现有的质差用户分析平台进行对比分析，在和实际投诉用户匹配度方面，本平台预测命中的投诉用户是现有质差用户分析平台的3倍以上。

此外，通过搭建潜在投诉用户挖掘及预警平台，可对投诉用户预测结果进行呈现及分析，如图5所示。平台同时支持对用户级质差和小区级质差的钻取，与实际生产系统进行对接，定位质差原因，用于进一步的网络质量问题分析和优化。

3 结束语

基于机器学习的网络投诉预测通过对历史网络投诉用户的数据进行分析和学习，建立了投诉用户的网络特征模型。并且，通过不断累积投诉用户数据，可周期性更新和优化模型，提升预测准确率。在网络应用后，该方法预测效果较好，预测结果被网络部门和客服部门使用，有效地减小了网络投诉用户数量，提升了客户的网络满意度。随着5G网络的快速建设，5G用户也将迎来爆发性增长，此方法可同样运用于5G网络之中，为运营商提升5G客户满意度建立先发优势。

参考文献

[1] 李露，李一喆．基于AI的无线网络用户满意度分析[J]．邮电设计技术，2018(12)．
[2] 丁俊民，廖振松．基于大数据建模的投诉预测与应用[J]．信息通信，2015(9)．
[3] 钟鼎．基于神经网络的4G用户感知预警模型构建和应用[J]．电信技术，2016(11)．
[4] 李占山，刘兆赓．基于XGBoost的特征选择算法[J]．通信学报，2019(10)．

Prediction and analysis of network complaints based on machine learning

WAN Ren-hui, WANG Jie, DAI Peng-cheng, ZHANG Xu-yang, XIN Chao

(China Mobile Group Design Institute Co., Ltd., Beijing 100080, China)

Abstract Reducing network related complaints has always been one of the important work of telecom operators. At present, most of the early warning schemes for network complaint users are based on the experience of network optimization engineers, with low accuracy and efficiency. In this paper, through a comprehensive and in-depth analysis of the historical network complaint user data, a complaint user feature model is build based on XGboost algorithm, which can predict network complaint users. This method has a high prediction accuracy, and can locate the reasons for poor quality of users after docking with other network optimization systems, so that the network department can optimize the network in advance and improve user satisfaction.

Keywords machine learning; network related complaints; complaints predict