



基于决策树的 EPG 体验异常原因定位

谭晓敏, 方艾, 梁冰, 杨豪杰

(中国电信股份有限公司广州研究院, 广东 广州 510630)

摘要: 用户体验的提升一直是 IPTV 的运维重点。然而, 运营商 IPTV 组网相对复杂, 涉及机顶盒、接入网、EPG 页面、EPG 服务器、媒体服务器等多个环节和调度链路, 一旦发生异常, 运维人员需排查各环节的监控指标, 凭经验定位引起异常的原因。该运维模式专业性强、效率低下, 无法快速止损。为解决上述运维痛点, 基于大数据技术, 采用决策树对 EPG 体验数据建模, 根据决策树模型的决策规则, 快速定位异常原因。现网中, 决策树模型能够准确快速定位 EPG 体验的异常原因, 为辅助运维人员指明排查方向, 提高运维效率。

关键词: IPTV; 决策树; 大数据; 智能运维

中图分类号: TP393

文献识别码: A

doi: 10.11959/j.issn.1000-0801.2019087

Locating causes of abnormality of EPG experience based on decision tree

TAN Xiaomin, FANG Ai, LIANG Bing, YANG Haojie

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

Abstract: Improving user experience has always been the focus of IPTV operation. However, IPTV network is relatively complex, including multiple scheduling links such as set-top box, access network, EPG page, EPG server and media server. Once the abnormality occurs, monitoring indicators of each link will be checked to locate the cause of abnormality by human experience. The operation mode relies on professionals, which is inefficient and unable to stop losses quickly. Thus, to solve the problem above, based on big data technology, decision tree was implied to locate the cause of abnormality according to the classification rule of the tree model. Decision tree model has been used in actual networks of IPTV and helps to find out the cause of abnormality, which improves the efficiency of operation.

Key words: IPTV, decision tree, big data, AIops

1 引言

IPTV 作为一种新兴互动网络电视技术, 为用户提供各种视频内容和数字服务。近几年 IPTV 的迅猛发展, 使之成为电信运营商日益重要的一项

业务。随着 IPTV 用户数的快速增长, 如何及时发现并修复故障是保障业务质量的基本要求。IPTV 组网复杂, 在目前的运维工作中, 故障排障难度较大, 如何及时地发现 IPTV 异常并高效快速地定位异常原因已成为当务之急。

收稿日期: 2019-02-13; 修回日期: 2019-04-23

IPTV 故障种类繁多, 包括 EPG 响应故障、视频质量故障^[1]、网络系统故障^[2]等。传统的故障定位方式基于用户投诉, 运维人员根据用户的描述查看各项指标是否正常, 再根据经验判断故障节点。实际上, 仅仅根据投诉用户的简单描述并不能准确定位故障, 针对这一缺陷, 出现了基于机顶盒和 QoS 参数故障定位的运维方式。“机顶盒法”^[3]将 IPTV 业务指标分解成可测的质量指标后, 通过机顶盒内置监测工具来评估业务体验质量; “QoS 参数法”^[4]在适当的位置部署探针, 关联 MPEG-TS 层和 RTP 层 QoS 参数定位平台侧和网络层的故障。

近几年, 随着移动互联网规模的扩大和 IPTV 用户数量的快速增长, 现有的运维技术已经难以适应需求, 运维问题的复杂性和良好的用户体验之间的矛盾越发尖锐, 因此需要一种高效定位故障的新技术。IPTV 的上百万用户量是 AIOps (智能运维) 良好的前提条件, 海量多样的数据处理也逐渐应用到 IPTV 故障定位领域。自适应 SVM 分类器^[5]用于预测潜在 IPTV 投诉用户, 进而对投诉用户关联的故障设备进行提前维护, 有效地指导一线运维工作。改进的 AdaBoost 算法^[2,6]适用于不平衡数据集, 能够较好地预测 IPTV 故障。聚类分析^[7]用于聚类问题设备、定位故障。

通过机器学习挖掘故障原因, 实际上是分类模型输出特征重要性的过程, 能够解决目前运维效率低、不准确、不及时等缺陷。常用的分类模型有决策树、XGBoost^[8]、神经网络等。决策树可处理非数值型数据, 容易进行可视化, 结果易理解; XGBoost 是以决策树为基础的集成学习, 适合于中小数据集, 高效但难于可视化; 神经网络具有很强的学习能力, 数据量较大时模型更有效, 但数据预处理工作比较苛刻, 处理时间较长。考虑排障时间和可视化的需求, 本文基于现网 EPG 操作日志, 采用决策树建立分类模型、自动定位异常原因, 在实践中多次发现 EPG 体验异常并准

确定位导致异常的原因, 较好地辅助运维人员及时止损, 提升运维效率。

2 异常原因自动化定位

某省各地市用户的 EPG 体验水平有所差异, 如图 1 所示。正常情况下, 各地市的质差用户数在一定范围内呈周期性波动。若质差用户数暴增, 说明当前时间段用户的 EPG 体验变差, 存在异常, 如图 1 中箭头所指的时刻。

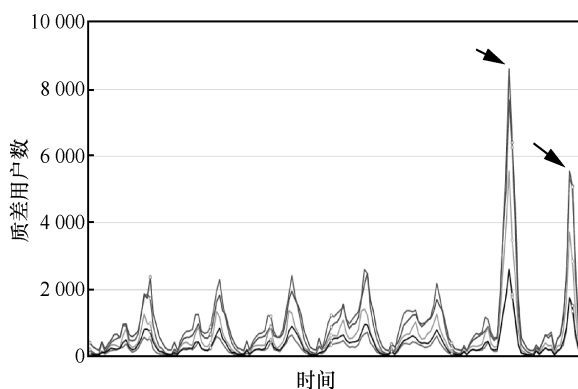


图1 各城市质差用户数

运维人员通过监测各市质差用户数指标主动发现异常。然而, 现网中有数百台服务器、上百个 EPG 页面、几十款机顶盒, 难以通过人工快速准确地定位导致异常的具体因素, 无法在排障的黄金时间内提出合适的应对方案, 及时止损。针对 IPTV 智能运维中的难点, 发现异常后, 取异常时间段的用户体验数据, 把用户分为质差用户和正常用户两类, 结合决策树模型自动挖掘现网异常用户的信息, 定位异常原因。

2.1 决策树模型

决策树^[9]是数据挖掘中一种简单有效的分类算法, 通过对训练样本进行归纳学习, 从无规则的实例中推理出以决策树表示的分类规则。决策树由一个根节点、多个内部节点和叶节点构成, 其中根节点是整个决策树的开始, 内部节点包含特征属性测试条件, 用以区分具有不同特征的记录, 每个叶节点代表一个类别。根节点到每个叶



节点的路径对应了一个结果判定, 建立模型的基本流程如下。

步骤 1 创建新节点, 对于当前节点的数据集 D , 若其样本个数小于阈值, 停止递归, 返回决策树。

步骤 2 根据划分度量选择最优划分特征属性, 把数据集划分子集, 生成分支节点。

步骤 3 对新的子节点递归调用步骤 1 和步骤 2, 生成决策树。

步骤 2 中如何选择最优划分特征属性是决策树的核心, 根据划分度量将样本划分成较小的子集, 划分度量则是指划分前和划分后样本的类分布。任意节点的类分布都可以表示为 (p_0, p_1) , 其中 $p_1 = 1 - p_0$ 。通常根据划分后子节点不纯性的程度选择最佳的划分度量, 不纯的程度越低, 类分布就越倾斜。不纯性度量包括信息增益、增益率和 Gini 系数。

信息增益是较为常用的度量指标。一般而言, 信息增益越大, 特征属性 a 划分获得的“纯度”越大。ID3 算法以信息增益为准则, 根据信息增益最大 (划分前后信息熵的差) 的特征分裂出各子节点, 采用递归的方法建立决策树的分支, 直到只含一种类别时停止。ID3 算法学习能力强, 但偏袒数目较多的特征属性, 且只能处理离散值属性。假定当前样本集合 D 中第 k 类样本所占比例为 $p_k (k=1, 2, \dots, |y|)$, 信息增益定义如下:

$$\text{Ent}(D) = - \sum_{k=1}^{|y|} p_k \lg p_k \quad (1)$$

$$\text{Gain}(D, a) = \text{Ent}(D) - \sum_{v=1}^V \frac{|D^v|}{|D|} \text{Ent}(D^v) \quad (2)$$

为了减少偏袒带来的不利影响, C4.5 算法采用增益率替代信息增益作为度量指标, 从候选特征属性中找到信息增益高于平均水平的特征属性, 再从中选择增益率最高的特征属性作为最优划分特征^[10]。C4.5 算法能处理连续值属性。增益率定义如下:

$$\text{IV}(a) = - \sum_{v=1}^V \frac{|D^v|}{|D|} \lg \frac{|D^v|}{|D|} \quad (3)$$

$$\text{Gain}_{\text{ratio}}(D, a) = \frac{\text{Gain}(D, a)}{\text{IV}(a)} \quad (4)$$

CART 是二分类回归树, 采用 Gini 系数最小化准则选择特征^[11], 特征较多时, 根据特征的重要性忽略对分类没有贡献的特征属性。Gini 系数定义如下:

$$\text{Gini}(D) = \sum_{k=1}^{|y|} \sum_{k' \neq k} p_k p_{k'} \quad (5)$$

2.2 评价标准

决策树模型的性能评价指标可基于混淆矩阵, TP 指类别 1 (positive) 被预测为类别 1 的样本数, TN 指类别 0 (negative) 被预测为类别 0 的样本数, FP 指类别 0 被错误预测为类别 1 的样本数, FN 指类别 1 被错误预测为类别 0 的样本数, 见表 1。

表 1 样本混淆矩阵

样本类别	预测为正常 1	预测为异常 0
正常 1	true positive (TP)	false negative (FN)
异常 0	false positive (FP)	true negative (TN)

准确率 (ACC) 是较常用的评价指标, 是分类正确样本数与所有样本数的比值。通常来说, 准确率越高, 分类模型越好, 但在异常类别 0 和正常类别 1 样本不平衡的情况 ($0 << 1$) 下, 分类模型把大部分测试样本正确划分为 1, 准确率为 99%, 即使小部分类别 0 样本被分类错误, 也对准确率影响极小。

$$\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (6)$$

精确率 (precision) 是正确预测分类的样本数与实际样本数的比值。正常类别和异常类别的精确率计算式分别为式 (7) 和式 (8)。

$$\text{precision}_1 = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (7)$$

$$\text{precision}_0 = \frac{\text{TN}}{\text{TN} + \text{FN}} \quad (8)$$

召回率 (recall) 是正确预测分类的样本数与实际被预测为该类的样本数的比值。正常类别和异常类别的召回率计算式分别为式(9)和式(10)。

$$\text{recall}_1 = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (9)$$

$$\text{recall}_0 = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (10)$$

F1 分数则是精确率和召回率的综合指标。

$$\text{F1} = \frac{2\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (11)$$

3 模型建立与分析

3.1 数据来源

采用决策树来区分缓慢用户和正常用户, 输入数据取自 EPG 交互数据采集分析系统中异常小时段的采集日志, 表 2 显示了示例样本, 字段包含用户账号、接入的 EPG 服务器、用户机顶盒型号、访问的页面以及访问页面的加载时间。

表 2 日志样本

用户账号	EPG 服务器	机顶盒型号	访问页面	加载时间
用户 1	IP1	ZXT*	首页 tab1	7 012 ms
用户 2	IP1	FIBER*	首页 tab2	801 ms
用户 3	IP2	EC6108V*	首页 tab2	423 ms

3.1.1 特征向量提取

为了对用户进行分类 (异常用户和正常用户), 提取影响用户加载时长的特征——机顶盒、访问页面和接入的 EPG 服务器。由于全网机顶盒、EPG 页面和 EPG 服务器的个数多, 可采用独热编码对上百个特征进行编码形成特征向量, 记特征 $F = (f_1, f_2, \dots, f_n)$, n 为特征个数, 则经过独热编码的特征向量长度为 n 。独热编码可防止人为引入名词特征之间不存在的相对关系。

对表 1 中的日志样本进行独热编码, 特征共 8 个, 分别为 $f_1=IP1$ 、 $f_2=IP2$ 、 $f_3=ZXT^*$ 、 $f_4=FIBER^*$ 、

$f_5=EC6108V^*$ 、 $f_6=$ 首页 tab1、 $f_7=$ 首页 tab2、 $f_8=$ 首页 tab3, 用户 1 生成的特征向量为 $V=[1,0,1,0,0,1,0,0]$ 。

通过“加载时间”将用户分为异常用户和正常用户两类, 若加载时间大于 5 000 ms, 则定义用户为异常用户, 编码为 0; 反之, 定义用户为正常用户, 编码为 1。

3.1.2 训练样本

异常时间段的用户日志具有特殊性, 反映当前时间段的故障情况, 因此采用异常时间段的用户日志作为样本, 随机抽取 70% 作为训练样本, 30% 作为测试样本。实验采用 10 例现网异常数据, 每例样本量达 10 万级。

异常时间段中, 即使异常用户数有所上升, 但也无法达到正常用户的数量级。在机器学习算法中, 正反样本的不平衡会导致分类结果更倾向于样本数多的类别。为解决样本不平衡的问题^[12-13], 可采用欠采样方法从多数类别中随机抽取样本以减少多数类别样本的数量, 如 ENN (edited nearest neighbor) 算法; 采用过采样方法增加少量类别的数量, 以达到样本数量的平衡, 如 SMOTE (synthetic minority oversampling technique) 算法。另外, 还可以通过设置损失函数权重, 加大少数类别的惩罚因子, 使少数类别损失大于多数类别, 增强分类的准确性。

样本平衡后的决策树模型 ROC 曲线如图 2 所示, ENN 算法随机减少多数类别样本数, 样本信息的缺少或缺失对决策树分类效果的影响是负面的。SMOTE 算法效果最好, 其 ROC 曲线下的面积最大, 但 SMOTE 处理样本时间为 302 s。权衡运维对时间的要求, 设置损失函数是较折中的平衡样本处理方式, 处理时间不超过 5 s。

3.2 决策树的生成

由于样本量大、特征多, 为防止过拟合, 需设定以下 3 个参数: 决策树的最大深度 max_depth、叶节点包含的最小样本数 min_samples_leaf、最大叶节点 max_leaf_nodes。

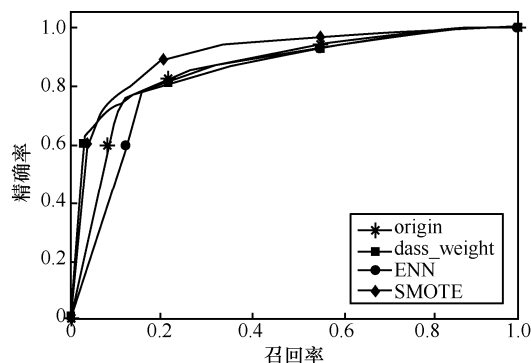


图2 样本平衡决策树模型 ROC 曲线

根据经验和实验设定决策树的最大深度 $\text{max_depth} = 10$, 叶节点最少样本数 $\text{min_samples_leaf} = 5$, 最大叶节点 $\text{max_leaf_nodes} = 30$ 。测试用例 mm527 总样本量 157 731, 训练样本量 110 411, 采用 Gini 分类标准和 CART 算法对训练样本构建决策树模型, 进行样本分类。

样例 mm527 决策树模型可视化图形的一部分如图 3 所示, 根据独热编码的规则, 特征 X_{17} 取值 1, 因此, 对于判断条件 $X_{17} \leq 0.5$, 若用户信息中包含特征 X_{17} , 则通向错误路径的内部节点; 否则通向正确路径的内部节点。决策树模型错误路径下的各叶节点均为异常, 例如包含特征 X_{17} 和 X_{88} 的用户被分类为慢用户。

决策树模型得到特征的重要程度 $X_{17} > X_{72} > X_{41}$, 权重值依次为 0.517、0.139、0.121。

权重值越大表明特征越重要, 即特征 X_{17} 是划分正常用户和异常用户的重要特征, 由此可见特征 X_{17} 是导致用户异常的主要原因。实际上, X_{17} 是现网某台服务器, 服务器运维人员确认当时该服务器出现了故障。

决策树模型多用于样本的分类, 侧重于预测新样本的分类结果。本文利用决策树模型的内部规则, 侧重分类属性, 重点关注异常类别的划分特征, 定位导致异常的原因。利用决策树模型定位异常是在异常时间段异常用户数增大的前提下实现的, 在异常用户数较少的情况下, 决策树模型并不能准确反映出各特征与类别的划分关系。一个好的分类模型 ROC 曲线应尽可能地靠近左上角, PRC 曲线应尽可能地靠近右上角。发生异常和没有发生异常条件下的决策树模型的 ROC 曲线和 PRC 曲线如图 4 所示, 正常时间段样本训练得到的决策树模型 PRC 曲线向原点凹陷, 召回率越大, 精确率衰减得越快, 分类模型没有具体意义。

实验结果表明, 沿着决策树模型“异常用户”叶节点自下而上追溯异常根源是可行的, 通过决策树模型可准确快速定位异常原因, 减少运维成本。

3.3 结果分析

分类评价指标中, 精确率和召回率是矛盾的, 召回率高而精确率低, 见表 3。现网异常原因定位

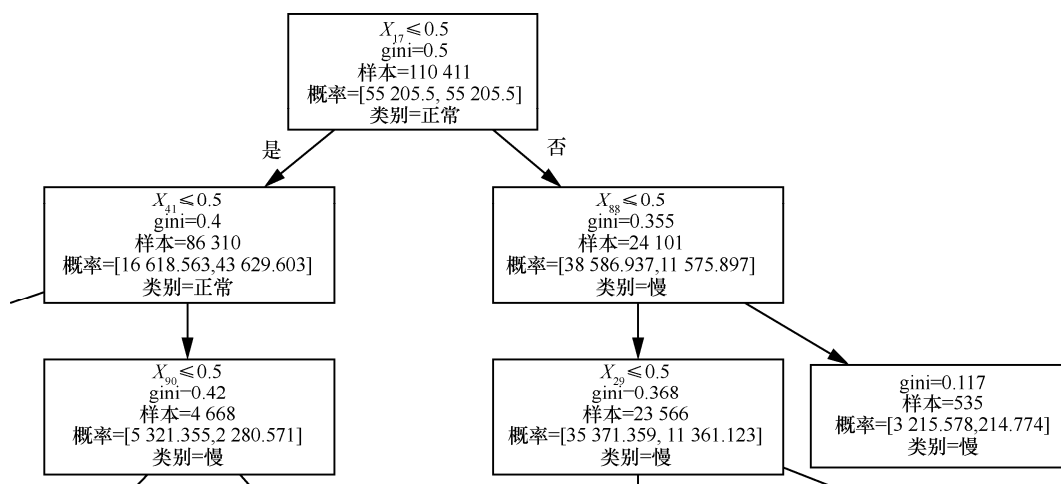


图3 mm527 样例决策树模型

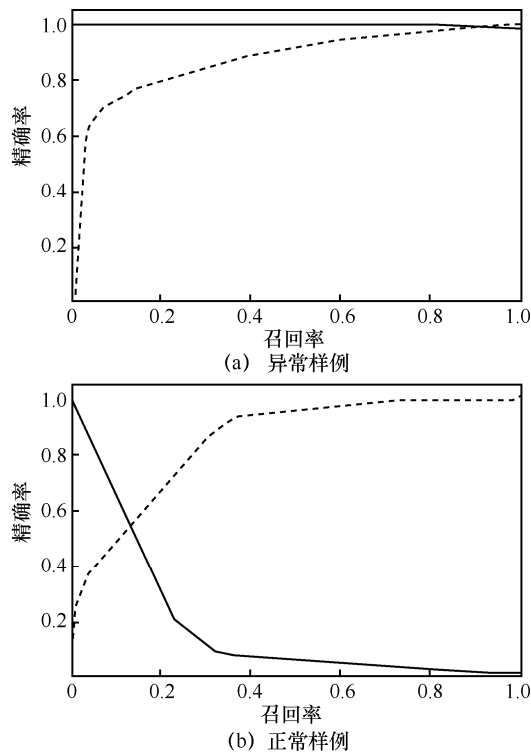


图 4 决策树 ROC/PRC 曲线

的实际情景下，本文更关注召回率，异常用户尽可能地分类正确，保证异常用户分类特征的准确性。表 3 中的 10 个样例是现网真实案例，“异常”类别召回率的均值为 0.82。10 个样例中，有 3 例异常通过决策树模型定位到某 EPG 服务器出了问题，报告相关 EPG 维护人员，检查发现问题服务器确实存在故障；有 7 例定位到 EPG 首页页面出了问题，报告相关页面维护人员，检查发现是问题页面获取资源超时导致用户长时间等待。

根据每一样例的决策树模型，寻找异常主因。由于异常突发在不同的时间点和不同的区域，决策树模型只针对当前时间段的异常，不要求决策树模型作为一个通用的模型适用于每个异常情景，因此，决策树参数也可以根据不同的情景做出调整，以提高召回率。

4 结束语

本文采用损失函数减少 EPG 用户正负样本不平衡对分类的影响，建立 EPG 用户分类决策树模

表 3 决策树模型评价指标

样例	类别	精确率	召回率	F1 值
mm527	异常	0.05	0.91	0.1
	正常	1.00	0.73	0.84
zq521	异常	0.09	0.84	0.16
	正常	0.99	0.71	0.83
yj514	异常	0.11	0.88	0.19
	正常	0.99	0.76	0.86
cz507	异常	0.07	0.81	0.12
	正常	0.99	0.76	0.86
hy507	异常	0.05	0.84	0.10
	正常	1.00	0.77	0.87
jm507	异常	0.15	0.88	0.26
	正常	1.00	0.93	0.96
jy507	异常	0.01	0.70	0.01
	正常	1.00	0.69	0.81
mm507	异常	0.22	0.86	0.35
	正常	1.00	0.94	0.97
mz507	异常	0.04	0.72	0.08
	正常	1.00	0.80	0.89
st507	异常	0.04	0.81	0.07
	正常	1.00	0.75	0.85
yj507	异常	0.13	0.83	0.22
	正常	1.00	0.93	0.96

型，通过分析决策树模型中“异常用户”类别的决策路径，确定导致异常的原因。实验表明，每一例异常的决策树模型只适用于异常本身，可以快速有效地从上百个影响因素中定位当时异常的原因，便于运维人员迅速发现问题，提高运维效率。

参考文献：

[1] 黄晓波, 黄志武. IPTV 视频质量故障定位及处理方法[J]. 信息通信, 2010(4): 67-68.
HUANG X B, HUANG Z W. IPTV video quality fault location and processing method[J]. Information and Communications, 2010(4): 67-68.

[2] 刘奇峰. 基于大数据的 IPTV 故障定位系统的设计与实现[D]. 南京: 南京邮电大学, 2017.
LIU Q F. Design and implementation of IPTV fault location



- system based on big data[D]. Nanjing: Nanjing University of Posts and Telecommunications, 2017.
- [3] 姚良, 肖晴, 陈文. IPTV 机顶盒内置质量监测工具的开发与应用[J]. 电信科学, 2008, 24(11): 80-83.
YAO L, XIAO Q, CHEN W. Development and application of built-in quality monitoring tool of IPTV set-top box[J]. Telecommunications Science, 2008, 24(11): 80-83.
- [4] 陈舒. 基于多协议层 QoS 参数关联的 IPTV 故障点定位方法[J]. 电视技术, 2013, 37(12): 55-58.
CHEN S. Method of IPTV fault isolation based on correlated analysis of multi-protocol-layer QoS parameter[J]. Video Engineering, 2013, 37(12): 55-58.
- [5] 吴志峰. 基于数据挖掘的 IPTV 故障诊断研究与实现[D]. 南京: 南京邮电大学, 2017.
WU Z F. The research and implement of IPTV fault diagnosis based on data mining[D]. Nanjing: Nanjing University of Posts and Telecommunications, 2017.
- [6] QIAN Y, WEI X, HUANG R, et al. Fault prediction in IPTV using improved AdaBoost algorithm[C]//International Conference on Wireless Communications & Signal Processing, Oct 13-15, 2016, Yangzhou, China. Piscataway: IEEE Press, 2016: 1-5.
- [7] 王泽, 屈海伟. 大数据分析在 IPTV 维护中的应用[J]. 电信技术, 2014, 1(10).
WANG Z, QU H W. Application of big data analysis in IPTV maintenance[J]. Telecommunications Technology, 2014, 1(10).
- [8] CHEN T, GUESTRIN C. XGBoost: a scalable tree boosting system[Z]. 2016.
- [9] 张琳, 陈燕, 李桃迎, 等. 决策树分类算法研究[J]. 计算机工程, 2011, 37(13): 66-67.
ZHANG L, CHEN Y, LI T Y, et al. Research on decision tree classification algorithms[J]. Computer Engineering, 2011, 37(13): 66-67.
- [10] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016.
ZHOU Z H. Machine learning[M]. Beijing: Tsinghua University Press, 2016.
- [11] STEINBERG D, COLLA P. CART: classification and regression trees[J]. The Top Ten Algorithms in Data Mining, 2009(9): 179.
- [12] MORE A. Survey of resampling techniques for improving classification performance in unbalanced datasets[J]. 2016, ar

Xiv: 1608.06048.

- [13] 熊冰妍, 王国胤, 邓维斌. 基于样本权重的不平衡数据欠抽样方法[J]. 计算机研究与发展, 2016, 53(11): 2613-2622.
XIONG B Y, WANG G Y, DENG W B. Under-sampling method based on sample weight for imbalanced data[J]. Journal of Computer Research and Development. 2016, 53(11): 2613-2622.

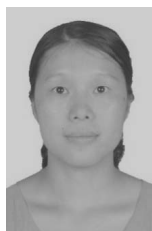
[作者简介]



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。



梁冰 (1978-), 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应用平台研发。



杨豪杰 (1982-), 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应用平台及智能网络软件开发。