计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(3):104−110 [doi: 10.15888/j.cnki.csa.006817]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于改进随机森林算法的停电敏感用户分类①

谢国荣[1], 郑　宏[1], 林伟圻[1], 徐　鸣[2], 郭　昆[3,4], 陈基杰[3,4]

[1](国网信通亿力科技有限责任公司, 福州 350001)
[2](国网福建省电力有限公司 电力科学研究院客户服务中心, 福州 350003)
[3](福州大学 数学与计算机科学学院, 福州 350116)
[4](福建省网络计算与智能信息处理重点实验室, 福州 350116)
通讯作者: 郭　昆, E-mail: gukn123@163.com

摘　要: 目前, 我国电网企业对于识别停电投诉风险, 开展用户停电敏感程度分析的研究工作还处在起步阶段. 为了有效地分析停电用户的敏感程度, 提出了一种基于改进随机森林算法的停电敏感用户分类算法. 首先, 对原始数据进行清洗、特征选择等预处理; 接着, 采用 SMOTE 算法增加少数敏感用户样本数据量, 解决数据分布不均匀问题; 然后, 以 Fisher 比作为特征的重要性度量, 按比例随机采样选取具有代表性的特征构成子特征空间; 最后, 利用随机森林算法识别停电敏感用户. 通过在真实停电数据上的实验, 验证了提出的方法不仅具有较好的准确性和时间性能, 而且可以有效处理高维、冗余特征的数据.

关键词: 停电敏感度分类; 随机森林; 不平衡数据; SMOTE 算法; Fisher 准则

引用格式: 谢国荣,郑宏,林伟圻,徐鸣,郭昆,陈基杰.基于改进随机森林算法的停电敏感用户分类.计算机系统应用,2019,28(3):104–110. http://www.c-s-a.org.cn/1003-3254/6817.html

Power Outage Sensitive Customers Classification Based on Improved Random Forest Algorithm

XIE Guo-Rong[1], ZHENG Hong[1], LIN Wei-Qi[1], XU Ming[2], GUO Kun[3,4], CHEN Ji-Jie[3,4]

[1](State Grid Info-Telecom Great Power Science and Technology Co. Ltd, Fuzhou 350001, China)
[2](Customer Service Center, Electric Power Research Institute, State Grid Fujian Electric Power Co. Ltd., Fuzhou 350003, China)
[3](College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350116, China)
[4](Fujian Provincial Key Laboratory of Network Computing and Intelligent Information Processing, Fuzhou 350116, China)

Abstract: At present, the research on the risk identification of power outage complaints and the customer sensitivity analysis in power grid companies is at its early stage. In order to effectively analyze the sensitivity of power outage customers, a sensitive customer classification algorithm based on the improved random forest algorithm is proposed. First, the data is preprocessed by methods of data cleaning, feature selection, and so on. Second, the SMOTE algorithm is used to increase the number of sensitive customers to solve the problem of data imbalance. Third, the representative feature space is selected by proportional random sampling. The Fisher ratio is used as the characteristic importance measure. Then, the random forest algorithm is used to recognize the customers that are sensitive to power outage. Finally, the experiments on real power outage data show that the proposed method not only has better accuracy and time performance but also can effectively deal with high-dimensional data with redundant features.

① 基金项目: 国家自然科学基金 (61300104, 61300103, 61672158); 福建省高校杰出青年科学基金 (JA12016); 福建省高等学校新世纪优秀人才支持计划 (JA13021); 福建省杰出青年科学基金 (2014J06017, 2015J06014); 福建省科技创新平台计划 (2009J1007, 2014H2005); 福建省自然科学基金 (2013J01230, 2014J01232); 福建省高校产学合作项目 (2014H6014, 2017H6008); 海西政务大数据应用协同创新中心

Foundation item: National Natural Science Foundation of China (61300104, 61300103, 61672158); Excellent Young Scientists Fund of Higher Education of Fujian Province (JA12016); New Century Talent Supporting Program of Higher Education of Fujian Province (JA13021); Excellent Young Scientists Fund of Fujian Province (2014J06017, 2015J06014); Science and Technology Innovation Platform of Fujian Province (2009J1007, 2014H2005); Natural Science Foundation of Fujian Province (2013J01230, 2014J01232); Industry-University Cooperation Project of Fujian Province (2014H6014, 2017H6008); Big Data Application Collaborative Innovation Center of Haixi District

收稿时间: 2018-08-20; 修改时间: 2018-09-18, 2018-10-18; 采用时间: 2018-10-23; csa 在线出版时间: 2019-02-22

Key words: power outage sensitivity classification; random forest; imbalanced data; SMOTE algorithm; Fisher criterion

1 引言

随着社会经济的不断发展, 各部门对电力的稳定性要求越来越高, 而电网公司通过各种供电可靠性措施, 不断加快抢修速度, 已使得停电次数和时间大大减少[1]. 尽管电网公司的服务水平越来越高, 用户对于用电的需求也在不断的提升, 一些停电敏感的用户对于供电更是具有严格的要求. 由于停电给用户带来的负面影响大小不同, 造成用户存在不同程度的停电敏感度[2]. 停电敏感用户是指对停电事件关注度较高的用户. 通过分析用户的行为特征, 借助数据挖掘、机器学习等技术对用电客户进行停电敏感度分类预测, 不仅可以有效的提高供电服务, 还有助于减少 95 598 的客户投诉量[3,4].

目前, 很多电网企业已经开启了客户关系管理 (CRM), 结合数据挖掘技术对用户的停电敏感度进行标识, 并根据不同用户提供差异化的增值服务, 提高用户的满意度[5]. Kaminski 等人[6]开发了一套基于决策树的停电用户敏感分类框架, 利用该框架可以计算出每个用户的停电敏感概率, 从而达到敏感用户分类的目的. 刘平等人[7]根据电力用户的满意度调研数据以及专家分析, 结合停电时段信息, 建立用户在不同时间段的停电敏感等级指数, 这种方法只针对不同类型的用户进行划分, 没有深入到个体用户层面. 严宇平等人[8]通过分析停电用户的属性特征, 利用逻辑回归和 SVM 算法等机器学习算法, 建立停电敏感程度预测模型, 模型可以准确的预测用户停电敏感度, 但模型训练时间较慢. 郑芒英等人[9]通过建立随机森林模型对用电用户停电敏感度进行分析, 可以区分出用户的敏感程度, 但是未能有针对性对目标用户清单进行筛选且未对用户停电特征进行评估, 模型的稳定性不高. 耿俊成等人[10]提出了基于 K-support 稀疏逻辑回归的停电敏感度预测模型, 通过优势分析法对特征属性的显著性进行分析, 提升了模型的准确性, 但该模型未考虑数据分布不均匀的问题.

本文针对传统的机器学习方法在不平衡数据集的处理性能较差, 及易存在过拟合等问题, 提出一种基于改进随机森林算法的停电敏感用户分类算法 POSCC (Power Outage Sensitive Customer Classification). 论文的主要创新点有: (1) 引入 SMOTE (Synthetic Minority Oversampling TEchnique) 算法提高少数类停电敏感用户的数据比例, 解决数据分布不均匀问题; (2) 改进随机森林算法特征选择阶段, 将 Fisher 比作为特征重要性的衡量指标, 依据比例和顺序选择重要的特征构成子树的特征集, 降低高维数据冗余特征的影响; (3) 通过与标准随机森林算法和经典 SVM 算法对比实验表明, 本文提出的算法在节省运行时间的同时具有较高的精度.

2 停电敏感用户分类算法

2.1 基本设计思想

本文提出的基于改进随机森林算法[11,12]的停电敏感用户分类算法 POSCC 主要包括数据预处理、数据分布不均匀处理、特征选取、分类模型训练构建四步. 预处理部分主要包括数据填充、异常值处理、数据标准化等操作. 分类算法是通过 Bootstrap 重采样的方式来构造每棵树的训练集, 以此保证基分类器的多样性. 整个训练集数据包括多数类样本和少数类样本数据, 使用 SMOTE 算法生成与少数类样本相似的子集, 再与少数类样本合并形成新的训练集, 通过这种方式可以有效的处理数据分布不均匀问题. 为了降低高维数据冗余属性的影响, 本文在特征选取部分做了改进, 首先对每个特征的重要程度进行计算[13], 然后根据权值的顺序对子特征进行分区划分, 接着对每个区按比例随机抽取特征, 最后构造出特征子空间. 根据生成好的训练集和特征子空间, 生成多棵决策树, 完成停电敏感用户分类算法的训练. 对于待测样本, 根据每棵树的分类结果, 采用投票的机制决定每个样本的分类结果. 图 1 给出 POSCC 算法的训练和预测流程.

2.2 数据预处理

停电用户数据是直接从各个停电业务系统中获取, 存在数据缺失和数据不一致的情况. 并且, 通过整合而来的数据经常会出现某些字段值缺失或异常等问题. 如果不对这些脏数据进行处理而直接在其上进行敏感用户划分, 则会严重影响预测的精度. 因此, 需要在建模前对数据进行预处理, 预处理的过程主要包括数据填充、异常值处理、数据规范化等.

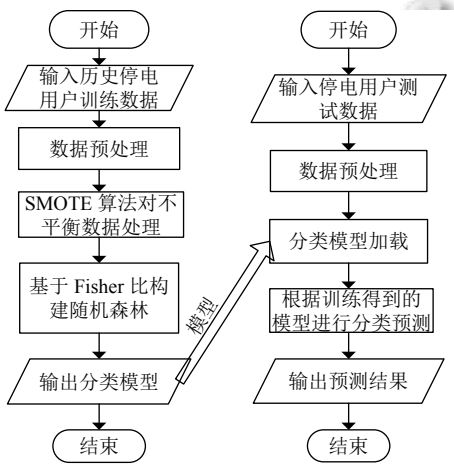
2.2.1 数据填充

针对停电用户数据的实际背景, 本文采用的数据

填充方案为: 对于类别型字段, 如行业分类、用电类别、电压等级、客户类别、行政区域、缴费类型等字段采用默认值填充, 分别填充预指定的缺失类别; 对于数值型字段, 采用平均值的方式处理. 针对某个时刻缺失的记录, 根据相邻多个时刻正常的记录进行求和取平均, 将平均值作为该缺失值记录的填充[14]. 计算方式如式 (1) 所示:

$$x_t = \frac{1}{m}\sum_{i=t-m}^{t-1} x_i \tag{1}$$

其中, x_t 表示第 t 时刻对应的值, m 为前相邻时刻正常数据的记录数. 这种方式综合考虑了最近多个时刻的信息, 弱化了其他因素的影响, 可以更合理地对缺失值进行填充, 进而保留了停电数据的连续性.



(a) 训练过程 (b) 预测过程

图 1 停电敏感用户分类算法训练和预测过程

2.2.2 异常值处理

在对数据进行处理的过程中会发现, 某些记录一个或多个字段的值与其他记录的值相差很大, 或者根本没有意义, 那就认为这些记录是异常数据. 比如, 用电量、停电次数、诉求次数等这些字段的值过大或者过小, 甚至出现为负的情况, 则说明这些字段存在异常值. 处理异常值的方法与处理缺失值的方法类似, 当发现异常数据时, 采用相邻多个时刻正常数据的平均值来替换该异常值, 从而降低噪声对停电敏感用户划分的影响. 平均值的计算参见式 (1).

2.2.3 数据标准化

停电用户数据中不同数值字段的取值范围可能会存在较大差异. 例如: 合同容量、本月电量等普遍是以百位数, 千位数的数值居多, 而停电次数、投诉次数等字段却以个位数或者十位数的数值居多. 因此需要进行标准化处理, 使不同值域的特征字段数据处在同一个量级范围内, 以便更好的进行建模分析.

鉴于停电用户数据的值域差异较大, 且各个特征字段的最大值和最小值都可求, 所以采用 Min-Max 标准化[15]来对数值字段数据进行数据标准化处理. 在标准化过程中, 若遇到某些特征字段值的最大值和最小值一样时, 则直接让该字段的值都为 0.5, 不进行线性变换. Min-Max 标准化如式 (2) 所示:

$$\underset{i,j\in[1,n]}{y_i} = \frac{x_i - \min(x_j)}{\max(x_j) - \min(x_j)} \tag{2}$$

其中, n 为数据的记录数, $\max(x_j)$ 特征字段的最大值, $\min(x_j)$ 为特征字段的最小值, x_i 为特征字段的值, y_i 表示为标准化之后的值.

2.3 不平衡数据处理

在停电敏感用户划分的研究中, 由于真正属于停电敏感用户的数据相对较少, 因此就会存在数据分布不均匀的现象. 如果直接用这些少数类的数据进行分类建模, 则很容易让模型学习到的信息过于特别而不够泛化, 从而让模型产生过拟合的现象. 因此, 采用 SMOTE 算法[16]来解决数据分布不均匀问题, 通过对少数停电敏感用户类样本数据进行抽样, 并将抽样的数据合成新样本添加到数据集中, 以此来提高少数类样本的比例. 使用 SMOTE 算法解决数据分布不均匀的流程如下:

(1) 对于停电敏感用户类中的每一个样本 x_i, 利用 k 近邻算法[17]得到样本 x_i 的 k 个近邻.

(2) 然后从这 k 个近邻中随机选择一个样本 $x_{i(nn)}$, 再生成一个 0 到 1 之间的随机数 $R_{0,1}$, 根据式 (3) 合成一个新的样本.

$$x_{new} = x_i + R_{0,1}\cdot(x_{i(nn)} - x_i) \tag{3}$$

(3) 将步骤 2 重复进行 N 次, 从而形成 N 个新的样本, N 即是根据采样比例确定的采样倍率.

SMOTE 算法是通过随机采样来生成新样本, 并非直接从实例复制而来, 这样可以缓解过拟合的问题, 同时不会损失有价值的信息. 所以, 采用 SMOTE 算法能够有效的处理停电敏感用户少数类数据分布不均匀问题.

2.4 基于 Fisher 特征比的特征选择

在对停电敏感用户分类模型的训练过程中, 每一步都要求提升数据的纯度, 以便达到更好的分类效果.

由于停电用户数据中的特征较多, 有些特征对于算法的贡献度不高, 甚至会对算法的训练过程会产生负面影响. 而且在高维特征空间中, 往往存在部分冗余特征, 因此需要对数据进行特征选择, 使得每次选出的特征更具有代表性. 特征选择是指从高维特征集合中, 根据某种评估标准选择那些输出性能最优的特征构成特征子集. 然后, 直接对这些特征子集进行建模, 从而降低模型的计算代价, 提高模型算法的预测精度.

本文改进随机森林算法的特征选择阶段首先用 Fisher 比计算每个特征的重要性权值, 根据权值进行从大到小排序; 然后以权值的均值为界, 将特征空间划分为两个特征子空间; 最后在划分好的特征子空间中, 按比例随机选择特征, 构造新的特征子空间. 具体步骤如下:

(1) 设停电数据集共有 n 个样本, 分属于 C 个类别: 对于第 w 类的集合, 样本个数为 n_w, 第 w 类中第 k 维特征的均值为 μ_{wk}, 第 w 类中第 k 维特征的方差为 σ_{wk}^2, 全部样本的第 k 维特征的均值为 μ_k, 则类内方差如式 (4) 所示:

$$S_W^k = \frac{1}{n}\sum_{w=1}^{C} n_w \sigma_{wk}^2 \tag{4}$$

类间方差如式 (5) 所示:

$$S_B^k = \frac{1}{n}\sum_{w=1}^{C} n_w (\mu_{wk} - \mu_k)^2 \tag{5}$$

则 Fisher 比计算方式如式 (6) 所示:

$$F_d = \frac{S_W^k}{S_B^k} \tag{6}$$

(2) 通过对每个特征计算 Fisher 比 F_d 之后, 按 F_d 从大到小对特征进行排列. 然后计算所有特征 F_d 的均值, 作为分界线, 将特征划分为重要特征区和次要特征区, 划分方式如式 (7) 和式 (8) 所示.

$$F_{1\text{st}} = \{F_d | F_d > \overline{F}\} \tag{7}$$

$$F_{2\text{nd}} = \{F_d | F_d < \overline{F}\} \tag{8}$$

(3) 在每一次的特征选择中, 根据划分好的重要特征区和次要特征区, 按比例从每个区中随机抽样 $m_{1\text{st}}$ 和 $m_{2\text{nd}}$ 个特征构造成特征子空间. 比例的计算公式如式 (9) 和式 (10) 所示.

$$P_{1st} = \frac{|F_{1\text{st}}|}{|F_{1\text{st}}| + |F_{2\text{nd}}|} \tag{9}$$

$$P_{2\text{nd}} = \frac{|F_{2\text{nd}}|}{|F_{1\text{st}}| + |F_{2\text{nd}}|} \tag{10}$$

其中, $|F_{1\text{st}}|$和$|F_{2\text{nd}}|$分别代表重要特征区总数和次要特征区总数.

对特征进行分区, 在一定程度上对特征选择的随机性做了约束, 保证选取出来的特征更具有代表性. 通过这一改进, 可以有效的降低维度的增加和冗余属性带来的影响, 并且能够提高模型的性能.

2.5 基于随机森林的停电敏感用户分类

应用随机森林算法构建停电敏感用户分类模型, 并基于该模型预测测试数据中哪些用户是停电敏感用户. 停电用户最终分为停电敏感用户和停电非敏感用户, 因此本文研究的是二分类问题. 分类的具体步骤如下:

(1) 采用 Bootstrap 策略, 有放回地随机抽取 n_1 (n_1<n) 个停电用户;

(2) 分别对 n_1 个用户构建决策分类树, 从原始的 d 个特征中随机选取 d_{try} 个特征, 再根据 2.4 节中介绍的 Fisher 比计算方法, 按比例选取具有代表性的用户特征构造特征子空间;

(3) 基于构造的特征子空间, 最大限度的让每棵树进行分裂生长, 直到 n_1 棵树组成的随机森林模型全部训练完成;

(4) 针对输入的测试用户数据, 由 n_l 棵树生成用户是否属于敏感用户的 n_l 个判断, 采用投票机制, 取票数最高的类别作为用户的最终类别. 若该类别为敏感用户类别, 则判断该用户为敏感用户.

3 实验结果与分析

3.1 数据集

本文所采用的数据是由南方某省各个地区在 SG186、95 598 业务支持系统、用电信息采集系统等采集汇总而来. 具体为该省份从 2016 年 1 月至 12 月的停电客户信息表, 包含客户档案信息、客户用电信息、台区停电信息、台区负荷信息、客户停送电诉求信息等共 56 个特征字段, 以及一个表示用户是否为敏感用户的类别字段. 数据大小共 9563 715 条, 由于数据集过大, 故采用 Spark 做并行化处理, 提高算法处理效率. 本文算法所涉及的字段信息如表 1 所示.

3.2 评估指标

在对停电敏感用户进行划分时, 连续型的数据已经被映射为离散型数据. 因此, 可用混淆矩阵[18]来展示预测的结果. 混淆矩阵可以直观的展示各个类别的预

测情况, 是分类算法中一种常见的评价指标. 矩阵的列表示预测类的实例, 行表示实际类的实例. 通过混淆矩阵的准确率和召回率可以很好的衡量分类算法的精度. 图 2 为混淆矩阵两类的分类结构.

表 1 具体字段信息

类别	字段名
用户信息	行业类别、用电类型、电压等级、用户类别、行政区域、合同容量、是否重要客户、重点保障用户、特殊用户标识、合同数量、缴费类型、本月电量.
台区信息	停送电延迟次数、平均停送电延迟时长、工作日负荷高峰时段、节假日负荷高峰时段、计划停电次数、故障停电次数、临时停电次数、平均停电时间、0611 停电次数、1218 停电次数、1921 停电次数、2205 停电次数、最近一次停电时长.
诉求信息	0611 诉求次数、1218 诉求次数、1921 诉求次数、2205 诉求次数、用户停送电来电次数、节假日诉求量、工作日诉求量、意见诉求量、临时停电诉求量、故障停电诉求量、计划停电诉求量、停电平均诉求量、平均停电敏感时长.

		预测值	
		正类	负类
实际值	正类	TP	FN
	负类	FP	TN

图 2 混淆矩阵

图 2 中, TP (True Positives) 为被正确地划分为正类的数量; FP (False Positives) 为被错误地划分为正类的数量; FN (False Negatives) 为被错误地划分为负类的数量; TN (True Negatives) 为被正确地划分为负类的数量.

通过混淆矩阵可以计算出每个类别的准确率、召回率和 F1 测度, 这些都是评估分类模型常用的重要指标.

(1) 准确率 (*Precision*):

准确率表示在预测结果中, 预测为正类且确实为正类的数据量占预测为正类数据量的比例:

$$Precision = \frac{TP}{TP+FP} \tag{11}$$

(2) 召回率 (*Recall*):

召回率是覆盖面的度量, 表示为预测为正类且确实为正类的数据量占所有正类数据量的比例:

$$Recall = \frac{TP}{TP+FN} \tag{12}$$

(3) *F*1 测度 (*F*1-measure):

*F*1 测度是 *Precision* 和 *Recall* 加权调和平均, 计算公式如式 (13) 所示:

$$F1 = \frac{2*Precision*Recall}{Precision+Recall} \tag{13}$$

3.3 实验方案

基于改进随机森林算法的停电敏感用户分类模型的代码是基于 Spark1.6.0 实现, 所使用的实验环境为 8 台硬件配置为 Core Duo 2.0 GHz CPU、16 GB 内存, 软件配置为 Ubuntu 12.04 的虚拟机组成的集群.

3.3.1 算法精度实验

为了比较基于改进随机森林算法的性能, 在实验中采用基于 Spark 的标准随机森林算法 (RF) 和 SVM 算法作为对比, 检验不同算法构建出来的停电敏感用户分类模型在测试数据上的精度. 选择 SVM 算法的原因是, SVM 算法泛化能力高, 通过选取合适的核函数可以处理高维特征的数据. 每次实验均通过 10-折交叉验证, 模型训练 20 次, 取均值作为模型运行一次的性能. 改进随机森林算法的参数 k 表示森林的规模, 本文实验设置为 300. 参数 f_{max} 表示数据的最大特征数, 由于总特征数不多, 取值为 none, 代表考虑所有特征数. 参数 d_{max} 表示树的最大深度, 根据实验数据集的特征数取值为 20. 标准随机森林算法的参数设置与本模型相同. SVM 算法中的参数 Gamma 为核函数设置, 一般设置为 1/m, m 为属性数. 参数 cost 为惩罚性因子, 一般取值为 1.0.

3.3.2 数据分布不均匀实验

由于停电敏感用户属于少数类, 约占数据集的 0.16%. 为了克服不平衡数据对算法精度的影响, 采用基于 Spark 的并行 SMOTE 算法对数据做相应的处理. SMOTE 算法是通过过采样方法重复选取少数类样本, 以提高少数类样本的数据比例. 在此基础上采用不同过采样比例进行多组分类实验, 寻找最佳的过采样比例, 克服不均衡数据的局限性, 以提高算法精度.

3.4 实验结果与分析

3.4.1 算法精度实验

图 3、图 4、图 5 为 3 种不同算法构建出模型的准确率、召回率和 *F*1 测度对比. 从图中可以看出, POSCC 算法在三个指标值上均高于其他的算法. 这主要是因为本文算法通过对特征选择步骤的改进, 在一定程度上降低了特征选择的随机性, 可以有效的处理多维向量相关的问题, 降低模型泛化误差, 并且可以避免过拟合问题, 提高了算法的分类精度. 由于 SVM 算法的核

函数是将高维数据映射到低维空间, 存在部分数据精度的缺失, 所以精度较低, 但又因为其泛化能力高, *Recall* 值较高于标准 RF 算法. 综上, POSCC 算法可以有效的对停电敏感用户进行分类.

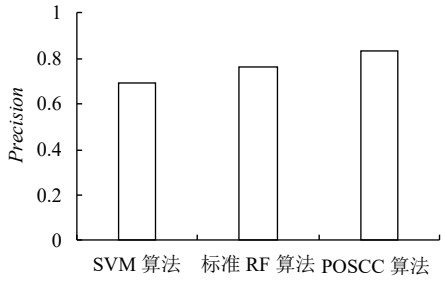


图 3 不同算法的准确率对比

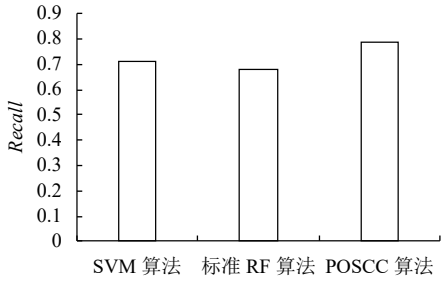


图 4 不同算法的召回率对比

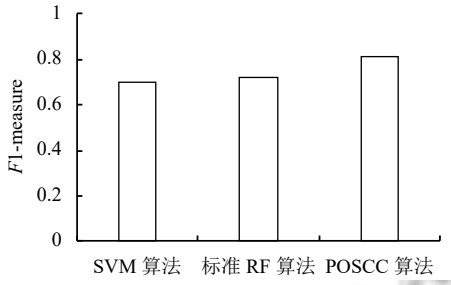


图 5 不同算法的 *F*1 测度对比

图 6 为三种算法构建出的模型运行时间对比. 从图中可以看出, POSCC 算法比标准随机森林算法和 SVM 算法就有较好的时间性能. SVM 算法需要对核矩阵进行分析, 需要大量的计算时间. 标准 RF 算法直接将用户的所有特征作为特征的输入, 因此也需要一定的运行时间. POSCC 算法通过将 Fisher 比的特征选择与随机森林算法相结合, 使得特征子空间更具有代表性, 降低了高维特征的数据中存在部分冗余特征的影响, 从而减少了模型在决策树节点分裂时对于冗余特征的重复计算, 有效的降低了算法分类判断的计算量. 因此 POSCC 算法运行时间较低.

3.4.2 数据分布不均匀实验

(1) 采用 SMOTE 算法与未采用 SMOTE 算法的实验结果

图 7 为采用 SMOTE 算法与未采用 SMOTE 算法的实验结果. 从图中可以看出, 采用 SMOTE 算法处理后模型的三种指标结果值比未采用 SMOTE 算法处理的结果高, 这是因为采用 SMOTE 算法后, 增加了停电敏感用户类别的样本数, 通过提升算法分类器的学习强度, 并降低随机森林中树之间的相关性, 最终让模型在停电敏感用户分类具有更好的分类效果.

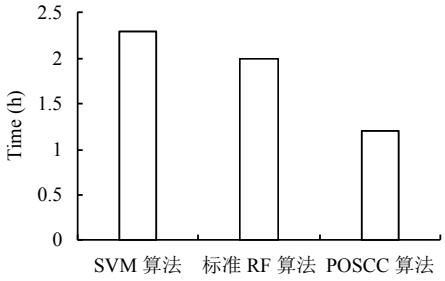


图 6 不同算法的运行时间对比

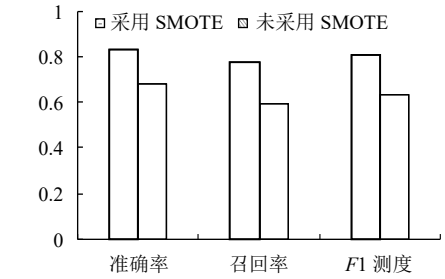


图 7 采用 SMOTE 算法与未采用 SMOTE 算法的实验结果

(2) 不同过采样比例实验结果

使用 SMOTE 算法进行过采样, 通过调整算法参数, 得到 5 组不同数据比例的实验结果, 如表 2 所示. 从表中看出, 随着采样比例的提高, 准确率和 *F*1 测度均有所下降, 当比例为 4:5 时就已经接近为 0. 实验表明, 为了达到较高的准确率, 应该保持较低的过采样比例, 而不能为了增加少数类样本数理一味的提高过采样比例. 所以, 本文采用 1:5 的比例作为 SMOTE 算法过采样比例的标准.

表 2 不同过采样比例结果

指标	1:5	2:5	3:5	4:5	5:5
Precision	0.81	0.54	0.25	0.03	0.01
Recall	0.83	0.57	0.29	0.04	0.02
*F*1-measure	0.79	0.51	0.22	0.02	0.01

4 结束语

本文提出了一种基于改进随机森林算法的停电敏感用户分类算法 POSCC. 通过引入 SMOTE 算法对少数类样本进行处理, 提高停电敏感用户类数据比例, 降低模型算法的泛化误差. 再对停电用户数据特征的深入分析, 改进随机森林的特征选择方法, 根据 Fisher 比对特征进行分区, 按比例选取有代表性的特征, 在一定程度上降低算法的随机性, 提高了算法的性能. 实验表明, 相比较于其他算法, 本文提出的算法可以很好的对停电用户的敏感度进行分析, 具有较高的准确率和时间性能. 由于停电数据是实时更新的, 下一步的工作, 将考虑设计基于增量分类算法的停电敏感用户预测, 以进一步提升算法的准确性与实用性.

参考文献

1 刘自发, 张在宝, 杨滨, 等. 电网大停电社会综合损失评估. 电网技术, 2017, 41(9): 2928–2940.

2 Du Z W, Ha H X, Song Y, *et al*. New algorithm based on the sensitivity and the compensation methods for line-outage problem of power network. Power System Protection and Control, 2010, 38(16): 103–107.

3 许鑫, 王莉, 孙志杰, 等. 一种基于数据挖掘的频繁停电投诉预警模型. 信息记录材料, 2017, 18(2): 64–66.

4 程丽冰. 大数据时代的电力客户分群管理应用研究[硕士学位论文]. 广州: 华南理工大学, 2016.

5 Kumar G, Pindoriya NM. Outage management system for power distribution network. Proceedings of 2014 International Conference on Smart Electric Grid. Guntur, India. 2014. 1–8.

6 Kamiński B, Jakubczyk M, Szufel P. A framework for sensitivity analysis of decision trees. Central European Journal of Operations Research, 2018, 26(1): 135–159. [doi: 10.1007/s10100-017-0479-6]

7 刘平, 叶涛, 李立军, 等. 基于快速恢复供电的应急抢修研究. 电力安全技术, 2014, 16(4): 1–4. [doi: 10.3969/j.issn.1008-6226.2014.04.001]

8 严宇平, 吴广财. 基于数据挖掘技术的客户停电敏感度研究与应用. 新技术新工艺, 2015, (9): 89–93. [doi: 10.3969/j.issn.1003-5311.2015.09.029]

9 郑芒英. 用电客户停电敏感度分析[硕士学位论文]. 广州: 华南理工大学, 2014.

10 耿俊成, 张小斐, 孙玉宝, 等. 基于 K-support 稀疏逻辑回归的停电敏感度预测. 计算机与现代化, 2018, (4): 68–73. [doi: 10.3969/j.issn.1006-2475.2018.04.013]

11 Breiman L. Random forests. Machine Learning, 2001, 45(1): 5–32. [doi: 10.1023/A:1010933404324]

12 Liaw A, Wiener M. Classification and regression by random forest. R News, 2002, 2(3): 18–22.

13 马春来, 单洪, 马涛, 等. 一种基于随机森林的 LBS 用户社会关系判断方法. 计算机科学, 2016, 43(12): 218–222. [doi: 10.11896/j.issn.1002-137X.2016.12.040]

14 Han JW, Kamber M, Pei J. 数据挖掘: 概念与技术. 范明, 孟小峰, 译. 北京: 机械工业出版社, 2012.

15 Guerrero E, Wang H, Alvarez J, *et al*. A three-dimensional range-free localization algorithm based on mobile beacons for wireless sensor networks. Computer-Aided Design, Drafting and Manufacturing, 2010, 20(1): 83–92.

16 Chawla NV, Bowyer KW, Hall LO, *et al*. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 2002, 16(1): 321–357.

17 Cover T, Hart P. Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 1967, 13(1): 21–27. [doi: 10.1109/TIT.1967.1053964]

18 Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters, 2006, 27(8): 861–874. [doi: 10.1016/j.patrec.2005.10.010]