基于Stacking模型融合的专变用户电费回收风险识别方法

潘国兵[1],龚明波[2],贺　民[2],邬程欢[2],唐小淇[3],杨　吕[1],欧阳静[1]
(1. 浙江工业大学 分布式能源与微网研究所,浙江 杭州 310012;
2. 国网浙江省电力有限公司宁波供电公司,浙江 宁波 315000;
3. 浙江华云信息科技有限公司,浙江 杭州 310012)

摘要:针对当前电力公司面临的专变用户电费回收风险,提出一种基于Stacking模型融合的专变用户电费回收风险识别方法。对专变用户数据进行特征处理、特征构造与特征筛选,从样本分布和特征属性上优化模型的泛化性能;利用Stacking模型融合多个基学习器,构建专变用户电费回收风险识别模型。实验结果表明,相较于其他常用的分类算法,所提方法具有更优的精确率、召回率、*P*-*R*调和均值、AUC值以及模型泛化性能,对专变风险用户的识别率也更高。

关键词:专变用户;电费回收;风险识别;Stacking模型融合;LGBM
中图分类号:TM 73　　**文献标志码**:A　　DOI:10.16081/j.epae.202010022

0　引言

随着我国电力体制的不断改革、市场机制的不断推进,电力市场的营销工作也在同步优化。电力营销中最重要的工作是电费回收,电费回收工作的顺利开展直接关系到电力公司未来的经济效益,因此电力公司对用户的电费回收风险识别工作尤为重要。同时随着社会经济的发展,电力用户的用电需求开始趋于多样化,专变用户的数量也逐渐增多。专变用户是指变压器由用户投资、电力公司代管的用户,其用电量一般很大,产生的电费金额很高[1]。准确识别专变用户的电费回收风险,并及时采取有效的应对措施,对于电力公司的健康发展具有重要意义。

人工智能技术的迅速发展给电力行业带来了巨大变革[2],机器学习技术也已在电费回收风险识别领域得到了广泛使用。目前国内外常用的用户电费回收风险识别算法有逻辑回归LR(Logistic Regression)[3]、随机森林RF(Random Forest)[4]、支持向量机SVM(Support Vector Machine)[5]等。文献[6]提出一种基于LR模型的电力客户欠费预测算法,引入人工智能算法改变原先事后欠费管理的被动局面。LR算法虽然具有模型简单、精度高等优点,但是在处理电力用户这种具有大量特征的数据集时可能存在欠拟合的情况,并且对于特征的多重共线性较为敏感。文献[7]针对用户分布不均衡、模型参数确定困难等问题,基于改进RF算法进行电力用户欠费风险预警。对于电力用户数据这样的典型不平衡数据集而言,RF虽然可以平衡误差,但是算法的解释性不好。尽管RF中各棵决策树的决策性能不同,但是决策权重却是相同的,这在一定程度上会削弱模型的精度。文献[8]采用长短期记忆网络算法进行电费回收风险预警,进一步提高了预警的准确率,但模型的计算量、复杂度也大幅增加。文献[9]基于集对分析和马尔可夫链方法对电力客户进行信用风险评估,进一步完善了电费回收风险识别方法。

上述识别用户电费回收风险的算法虽然都提供了新颖的解决思路,但是也各有缺点,且上述算法都是采取单一算法,容易因为样本的随机性而导致模型的识别性能不佳。模型融合在用户信用识别、风险预测、电力设备故障诊断等电力系统数据挖掘领域效果极佳[10-13],Stacking模型融合的思想就是充分结合各种算法的优势,取长补短利用模型融合实现整体模型的性能提升。

鉴于现有用户电费回收风险识别方法存在的问题,本文提出一种基于Stacking模型融合的专变用户电费回收风险识别方法。首先针对专变用户样本倾斜与特征繁杂的问题进行特征处理,实现训练数据样本分布均衡与特征规范化;然后利用Stacking模型融合方法进行专变用户电费回收风险识别;最后利用某省电力公司的专变用户数据集进行有效性验证,结果表明基于Stacking模型融合的方法进行专变用户电费回收风险识别,在模型泛化性能与鲁棒性方面均优于其他方法。

1　专变用户特征处理

1.1　数据清洗与不平衡数据集过采样处理

专变用户数据集中通常存在大量的异常值与缺失值,本文对异常值进行剔除,对缺失值进行填充。对专变用户数据集中的异常值采用阈值法进行处

收稿日期:2020-01-16;**修回日期**:2020-08-30
基金项目:国家重点研发计划项目(2017YFA0700300);浙江省重点研发计划项目(2018C01081)
Project supported by the National Key R&D Program of China(2017YFA0700300) and the Key R&D Plan Project of Zhejiang Province(2018C01081)

理，即判定超过预设阈值的特征值为异常特征值，并从数据集中剔除。以专变用户第i月用电量异常值剔除为例，首先采取箱型图分析的方法确定第i月用电量的阈值，定义第i月用电量x这个特征的正常取值范围为$x\in\left[x_{\mathrm{L}}-1.5I_{\mathrm{QR}},x_{\mathrm{U}}+1.5I_{\mathrm{QR}}\right]$，其中$x_{\mathrm{L}}$为第$i$月用电量下四分位数，即有1/4的第$i$月用电量取值比它小，$x_{\mathrm{U}}$为上四分位数，即有1/4的第$i$月用电量取值比它大，$I_{\mathrm{QR}}=x_{\mathrm{U}}-x_{\mathrm{L}}$为四分位数间距，若第$i$月某用户用电量不在正常取值范围内，则判定为超过阈值的异常特征值，并删除该条专变用户数据。同时对专变用户数据集进行完整性检验，若其中存在缺失特征值，则采用同类均值插补的方法进行补齐。

专变用户数据集大多存在数据倾斜方面的问题，风险用户所占比例远低于正常用户，属于典型的不平衡数据集。由于样本分布的不平衡会在很大程度上影响分类算法的精度，因此本文采用合成少数类过采样技术SMOTE（Synthetic Minority Oversampling TechniquE）对专变用户数据集进行过采样处理[14]，使训练集中风险用户和正常用户的比例提升至1∶1，这样可以有效缓解样本分布不平衡带来的问题。电力公司进行专变用户电费回收风险识别的主要目的是识别风险用户，采用SMOTE增加风险用户的数量，将会使样本分布均衡，降低模型的过拟合风险，提高模型在测试集的泛化性能，增加模型对专变风险用户的识别率。

1.2　连续特征规范化

如果专变用户各个连续型特征之间的取值范围相差甚远，则值域偏大的特征具有较高的“权重”，值域偏小的特征造成的影响较低。因此对数据清洗后的连续型特征还需进行最小–最大规范化处理，以保证值域的一致性，如式（1）所示。

$$x'=\frac{x-x_{\min}}{x_{\max}-x_{\min}} \tag{1}$$

其中，x'为规范化后的值；x为原始值；$x_{\min}$为最小值；$x_{\max}$为最大值。

1.3　离散特征One-hot编码

专变用户特征中存在多种离散的标称型特征，没有明确的大小和序列关系，不能使用简单的数值型变量进行替代，并且多数模型不能直接处理这种离散特征。本文采用表1所示One-hot编码对离散特征进行处理，可以保证特征值权重的取值正常，并且起到扩充属性特征的作用。

表1　One-hot编码

Table 1　One-hot encoding

属性名	整数值	编码			
		Poor	Ok	Good	Great
Poor	1	1	0	0	0
Ok	2	0	1	0	0
Good	3	0	0	1	0
Great	4	0	0	0	1

2　专变用户特征构建

2.1　专变用户衍生特征

专变用户的用电量趋势可以在一定程度上反映该用户是否隐含电费回收风险，基于时间序列利用每月用户的用电量构造新的专变用户衍生特征。由于通常正常用户的用电量比较平稳，而风险用户的用电量波动较大且用电量呈现明显的下降趋势，因此本文将用户用电量进行最小二乘法拟合，将得到的斜率作为衡量用电量趋势的依据。基于时间序列使用滑动窗口的方法构造衍生特征，如图1所示。图中，时间轴上展示了该专变用户每月的用电量，时间轴下表示使用滑动窗口构造用电量趋势特征。利用最小二乘法拟合生成各滑动窗口内用电量的斜率衍生特征，其中第i月的用电量趋势是考虑前4个月和第i月当月得到的用电量斜率，如图1所示在滑动窗口W_1内利用最小二乘法拟合生成的斜率衍生特征代表第5月的用电量趋势，考虑了第1月至第5月的用电量斜率，计算公式如式（2）所示。

$$k_i=\frac{\sum_{n=i-4}^{i}(Q_n-\overline{Q})(n-\bar{n})}{\sum_{n=i-4}^{i}(n-\bar{n})^2}\quad i=5,6,\cdots,12 \tag{2}$$

其中，k_i为第i月用电量趋势值；Q_n为第n月的用电量；$\overline{Q}=\frac{1}{5}\sum_{n=i-4}^{i}Q_n$；$\bar{n}=\frac{1}{5}\sum_{n=i-4}^{i}n$。

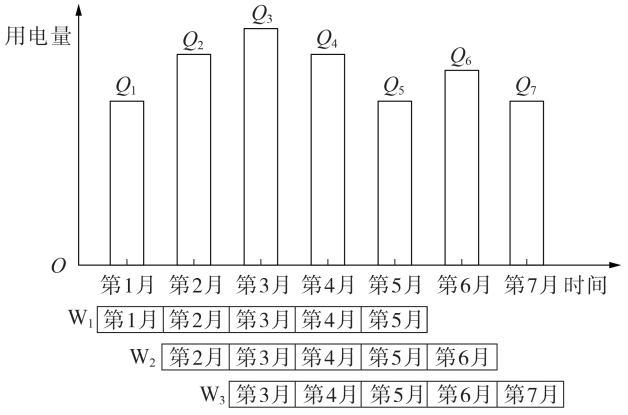

图1　滑动窗口构造衍生特征示意图

Fig.1　Schematic diagram of derived features construction with sliding window

在实际电力公司的电费回收工作中发现，专变用户（如工业用户）的用电量趋势可以反映企业经营情况的好坏，在一定层面反映该用户是否有窃电行为，且可侧面反映该用户电费回收风险[15]。因此若用户连续几个月的用电量趋势值不断下降（不包含季节、工作日等引起的周期性下降），则认为该用户隐含电费回收风险。计算5个月内用电量趋势值递

减的月数，如式(3)所示。

$$L(i)=\begin{cases}1 & k_i<k_{i-1}\\ 0 & k_i\geqslant k_{i-1}\end{cases} \tag{3}$$

其中，$L(i)$为判断第i月与第i–1月相比用电量趋势值是否递减的函数；k_{i-1}为第i–1月用电量趋势值。

则这5个月的用电量下降趋势指标为：

$$S=\sum_{n=i-3}^{i}L(n) \tag{4}$$

对于用电量下降趋势指标大于预设阈值3的特征置1，其余置0。通过上述过程完成用户用电量趋势这一衍生特征的构造。

2.2 专变用户特征筛选

专变用户特征筛选的意义在于选取更有助于辨别专变风险用户的特征，特征子集的选择对于专变用户电费回收风险识别的准确率影响较大。对专变用户的特征筛选包括去除冗余特征和不相关特征，特征筛选可以显著提升模型的精度与性能。

首先计算专变用户各特征(原始特征和基于时间序列构造的衍生特征)与标签(该用户是否为风险用户)之间的相关性，筛选出其中的弱相关特征，不作为输入模型的属性特征；然后对各特征之间的相关性进行计算，若相关系数大于预先设定的阈值，则随机剔除其中任意一个冗余特征。

对于标称型特征，如专变用户用电类别、电压等级、城乡标志、违约标识、用电量趋势指标等分别使用卡方检验与标签进行相关性分析，删除与标签相关性低的标称型特征，卡方相关性值χ^2的计算公式如下：

$$\chi^2=\sum_{i=1}^{m}\sum_{j=1}^{t}\frac{(s_{ij}-e_{ij})^2}{e_{ij}} \tag{5}$$

其中，s_{ij}为联合事件(x_i,y_j)的观测频度，e_{ij}为(x_i,y_j)的期望频度，其中特征x有m个不同的值x_1、x_2、…、x_m，特征y有t个不同的值y_1、y_2、…、y_t。

对于数值型特征，如用户违约次数、月用电量、电费金额、缴费变更次数等采用皮尔逊相关系数进行相关性分析，皮尔逊相关性值r的计算公式为：

$$r=\frac{\sum_{i=1}^{m}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{m}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{m}(y_i-\bar{y})^2}} \tag{6}$$

其中，特征x有m个不同的值x_1、x_2、…、x_m；特征y有m个不同的值y_1、y_2、…、y_m；$\bar{x}$与$\bar{y}$分别为特征x和y的均值。对于皮尔逊相关系数，一般认为：$|r|\leqslant 0.3$表示2个特征不存在线性关系；$0.3<|r|\leqslant 0.5$表示2个特征低线性相关；$0.5<|r|\leqslant 0.8$表示2个特征高线性相关；$|r|>0.8$表示2个特征极高线性相关，需要删除其中一个数值型特征。

3 专变用户电费回收风险识别模型

3.1 Stacking模型融合算法

模型融合的算法包括Stacking、Blending、Bagging、Boosting等。本文采用Stacking模型融合算法，主要分为2层：第一层将原始数据集划分成若干个大小相等的子训练集，输入第一层的K个基学习器中进行训练，每个基学习器都会输出各自的预测结果；第二层将第一层中生成的K个预测结果进行汇总，输入第二层的元学习器中进行训练，再由该元学习器输出最终的预测结果，具体见附录中图A1[16]。通过Stacking模型融合算法，可以使用元学习器对所有基学习器产生的预测结果进行泛化，以此来提高模型的整体预测精度。

Stacking模型融合算法对于第一层的基学习器有2个要求：模型预测结果多样性与分类性能相似性。模型预测结果多样性是指第一层的各个基学习器之间的预测结果应该存在一定的差异，产生的预测结果应有所不同，否则进行模型融合也不能提升太多精度。模型的分类性能相似性是指各个基学习器的分类性能应该基本处于同一水平，否则低性能的基学习器会对模型的整体性能造成较大影响[17]。

3.2 基于Stacking模型融合算法的专变用户电费回收风险识别

本文收集的数据包括专变用户基本属性(用户编号、行业分类、城乡标志等)、各月用电行为数据(各月的用电量、电费金额等)与缴费行为数据(缴费变更次数、各月缴费日期等)。首先对原始专变用户数据进行特征处理；然后构建专变用户衍生特征即用电量趋势指标，分别使用卡方检验与皮尔逊相关系数对标称型特征与数值型特征进行特征筛选；最终确定输入模型的特征包含用户编号、行业分类、用电类别、电压等级、城乡标志、违约标识、用电量趋势指标、用户违约次数、月用电量等23个特征，如表2所示，划分训练集与测试集进行风险识别模型建模。

表2 特征简介

Table 2 Features introduction

特征	取值范围	计算方式	相关系数计算方式
用户编号	—	—	—
行业分类	0~19	One-hot	卡方
用电类别	0~3	One-hot	卡方
电压等级	0/1	One-hot	卡方
城乡标志	0/1	One-hot	卡方
违约标识	0/1	One-hot	卡方
5—12月用电量趋势指标	0/1	One-hot	卡方
违约次数	0~12	规范化	皮尔逊
缴费变更指标	0/1	One-hot	卡方
7—12月用电量/(kW·h)	40~55 000	规范化	皮尔逊
标签	0/1	—	—

其中对制造业、建筑业等20个行业分类使用0~19赋值；对居民生活、农业生产用电等4个用电类使用0~3赋值；对电压等级、城乡标志等使用0／1赋值。

为了有效提升Stacking模型融合算法的整体泛化性能，需要选取性能强的学习算法作为第一层的基学习器。轻量梯度提升学习器LGBM（Light Gradient Boosting Machine）算法作为微软机器学习团队对极限梯度提升学习器XGBoost（eXtreme Gradient Boosting）算法进行改进而提出的一种高性能、高精度算法，有助于提升模型的整体预测效果。相较于XGBoost算法，LGBM算法的优势主要包括带深度限制的决策树Leaf-wise叶子生长策略、Histogram算法、直方图做差优化、直接支持类别特征、多线程优化[18]。

XGBoost在面对大量数据时，训练速度慢，耗费内存大，其采用的Level-wise叶子生长策略如附录中图A2所示，既费时又低效。LGBM采用的带深度限制的决策树Leaf-wise叶子生长策略，如图2所示。在分裂次数相同的情况下，对于专变用户电费回收风险识别这种数据量大、特征维数高的数据挖掘问题，与Level-wise相比，Leaf-wise生成的决策树精度更高，生成速度更快。通过对决策树的深度限制，既提高了模型泛化性能，也避免了模型过拟合的发生。

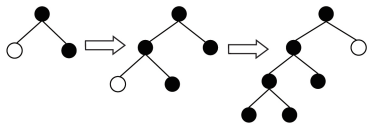

图2　LGBM的Leaf-wise叶子生长策略

Fig.2　Leaf-wise leaf growth strategy of LGBM

基于Stacking模型融合的专变用户电费回收风险识别方法，针对第一层的各个基学习器，首先将训练集进行4折划分，利用其中的3个子集训练模型，剩余的1个子集验证模型，然后重复进行该过程可能的4种选择。在完成4次训练后，验证集将会覆盖所有的训练集，将这4个验证集的预测结果（x_1—x_4）拼接，作为第二层元学习器的训练输入S^k_{train}（k表示第k个学习器的结果）。在每次训练完基学习器后，使用该学习器对测试集进行预测，得到测试集的预测结果。最终将测试集的4次预测结果（c_1—c_4）取平均得到S^k_{test}，作为第二层元学习器的测试输入。采用相同的方法完成对第一层其他基学习器的训练，本文Stacking模型融合算法第一层的具体原理见附录中图A3。

在完成对第一层各个基学习器的训练后，将得到的S^k_{train}和S^k_{test}分别按照基学习器的次序进行拼接，得到数据集S_{TRAIN}和S_{TEST}。将第一层得到的数据集S_{TRAIN}输入第二层元学习器进行训练，再使用训练好的Stacking模型融合算法对数据集S_{TEST}进行预测，得到Stacking模型融合的预测结果。

本文基于Stacking模型融合的专变用户电费回收风险识别模型第一层的K个基学习器不仅选取了LGBM和RF这种泛化性能较强的集成算法，同时也选取了LR与SVM这种鲁棒性较好的学习算法。LGBM和RF算法分别是Bagging和Boosting集成算法的典型代表，而SVM算法是小样本学习法的典型代表，这些差异性高、性能相近的基学习器有助于提升模型的整体性能。由于第一层的输出结果需要输入第二层模型进行再次泛化，因此本文选择模型简单、泛化性能好的LR算法作为第二层的元学习器。基于Stacking模型融合算法的专变用户电费回收风险识别方法的整体示意图见附录中图A4。

4　算例分析

本文所提出的基于Stacking模型融合的专变用户电费回收风险识别方法使用PyCharm与Anaconda编程实现，硬件环境为Intel（R）Core（TM）i7-8550U CPU@1.80 GHz　1.99 GHz、内存为8 GB，软件环境为Windows 10操作系统。采用某省电力公司的专变用户数据集进行验证，将本文所提基于Stacking模型融合的风险识别算法结果与LR、RF、SVM和LGBM算法的结果进行对比。

4.1　模型性能度量

专变用户电费回收风险识别是一个典型的二分类问题，对于二分类问题常用的评价指标有精确率（P）、召回率（R）、$P-R$调和均值（F_1）。对于专变电费回收风险用户和专变电费回收正常用户的分类问题，混淆矩阵如表3所示。表中，N_{TP}为将专变风险用户预测为风险用户数；N_{TN}为将专变正常用户预测为正常用户数；N_{FP}为将专变正常用户预测为风险用户数；N_{FN}为将专变风险用户预测为正常用户数。

表3　混淆矩阵

Table 3　Confusion matrix

真实类别	用户数	
	预测为风险用户	预测为正常用户
风险用户	N_{TP}	N_{FN}
正常用户	N_{FP}	N_{TN}

精确率、召回率、$P-R$调和均值具体计算公式为：

$$P=\frac{N_{\text{TP}}}{N_{\text{TP}}+N_{\text{FP}}} \tag{7}$$

$$R=\frac{N_{\text{TP}}}{N_{\text{TP}}+N_{\text{FN}}} \tag{8}$$

$$F_1=\frac{2PR}{P+R} \tag{9}$$

精确率和召回率是一对不可兼顾的评价指标，

若要综合考虑2个指标，可以使用P–R调和均值来对模型进行综合评价。本文使用精确率、召回率和P–R调和均值对模型进行综合考量。

专变用户样本中存在类别分布不平衡的问题，导致测试集的类别分布可能会随着时间的推移而有所不同，这使得测试集的精确率、召回率和P–R调和均值也会不断变化，因此仅使用上述指标对模型进行评价是不够全面的。受试者工作特征(ROC)曲线是研究模型泛化能力的有力工具，当模型的测试集样本分布发生变化时，ROC曲线依然可以保持较好的稳定性。ROC曲线的纵轴是真正率TPR(True Positive Rate)，公式为：

$$r_{\mathrm{TPR}}=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FN}}} \tag{10}$$

横轴是假正率FPR(False Positive Rate)，公式为：

$$r_{\mathrm{FPR}}=\frac{N_{\mathrm{FP}}}{N_{\mathrm{FP}}+N_{\mathrm{TN}}} \tag{11}$$

其中，r_{TPR}为TPR；r_{FPR}为FPR。

ROC曲线如图3所示。图中每个点对应于模型一个分类阈值时的TPR和FPR。点(0,1)可以理解为模型将所有专变用户样本正确分类，这是专变用户电费回收风险识别模型最理想的情况；点(1,0)可以理解为模型将所有专变用户样本错误分类，这是模型最差的一种情况；点(0,0)可以理解为模型将所有专变用户样本都预测为专变正常用户；点(1,1)可以理解为模型将所有专变用户都预测为专变风险用户；对角虚线可以理解为随机猜测模型。AUC(Area Under Curve)为ROC曲线下的面积值，其可以直接反映专变用户电费回收风险识别模型的性能。ROC曲线越接近点(0,1)，代表模型的性能越好，显然此时的AUC值越大。最终可综合ROC曲线、P–R调和均值和AUC值，对模型性能进行全面度量。

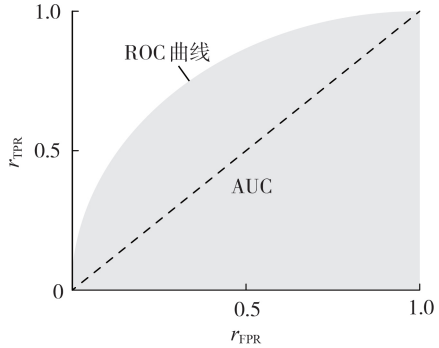

图3 ROC曲线图

Fig.3 ROC curve

4.2 专变用户数据集验证

本文采用的专变用户数据集来自某省电力公司2018年1月至2019年2月的专变用户信息，包括用户基本属性、历史用电行为信息与历史缴费记录信息。基于2018年12个月的用户信息，预测该用户2019年1月和2月是否为专变电费回收风险用户。对于专变电费回收风险用户的定义是当月产生的违约金大于0的用户。

首先对原始专变用户数据进行用户特征处理与用户特征构建，经过特征筛选后得到取值情况如表2所示，包含用户编号、行业分类等23个特征。对经过处理后的数据分别使用LR、RF、LGBM、SVM和Stacking模型融合算法构建专变用户电费回收风险识别模型。通过实验得到5种算法的ROC曲线如图4所示，Stacking模型融合算法的ROC曲线在LR、RF、LGBM和SVM算法的曲线之上，且更接近于点(0,1)，显然Stacking模型融合的AUC值也最大，故Stacking模型融合的专变用户电费回收风险识别算法的泛化性能更好。

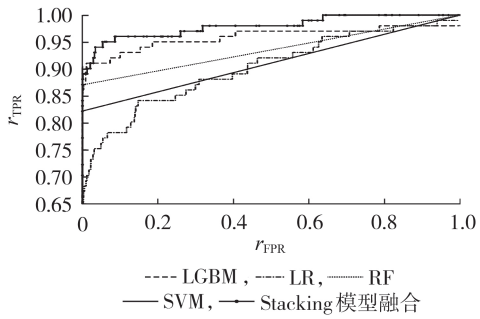

图4 5种算法ROC曲线对比图

Fig.4 ROC curve comparison among five algorithms

通过测试集进行实验得到5种算法的精确率、召回率、P–R调和均值以及AUC值如表4所示。分析5种算法的精确率和召回率可知：LR模型对专变风险用户的识别率不高，偏向于将风险用户预测为正常用户，这样会遗漏风险用户进而造成电力公司较大的经济损失；SVM模型的召回率达到了0.78，这表明模型对专变风险用户的识别率尚可，但是精确率最低，为0.62，这表明模型偏向于将专变正常用户预测为风险用户；RF模型对专变风险用户的识别率一般，模型的精确率尚可；LGBM模型对专变风险用户的识别率很好，模型的召回率很高，但是精确率不高，这表明模型偏向于将专变正常用户预测为专变风险用户；Stacking模型的召回率和精确率都很高，

表4 5种算法效果对比

Table 4 Effect comparison among five algorithms

算法	P	R	F_1	AUC值
LR	0.74	0.56	0.64	0.90
SVM	0.62	0.78	0.69	0.89
RF	0.86	0.75	0.80	0.94
LGBM	0.65	0.87	0.74	0.96
Stacking模型融合	0.86	0.88	0.87	0.98

这表明模型没有偏向于将风险用户预测为正常用户或者偏向于将正常用户预测为风险用户。对比*P*-*R*调和均值、ROC曲线与AUC值可知，Stacking模型融合算法的优势较大。Stacking模型融合算法的精确率达到0.86，与单个模型最高值RF模型的结果相同；召回率达到0.88，比单个模型最高值LGBM模型的结果提高了0.01；*P*-*R*调和均值达到0.87，比单个模型最高值RF模型的结果提高了0.07；AUC值达到0.98，比单个模型最高值LGBM模型的结果提高了0.02。综上，利用Stacking模型融合算法进行专变用户电费回收风险识别可以更好地识别风险用户。

5　结论

专变用户作为电力公司的用电大户，如果拖欠电费，将会对电力公司的健康发展造成巨大影响。为了解决当前电力公司对专变用户电费回收风险难以识别的问题，首先对专变用户的特征进行处理，改善训练样本的分布情况，然后进行特征工程与特征筛选，进一步提升模型的识别率，最后针对单一模型在识别专变用户电费回收风险时泛化性能不佳的问题，提出一种基于Stacking模型融合的专变用户电费回收风险识别方法。

专变用户数据集的实验验证结果表明，本文方法在精确率、召回率、*P*-*R*调和均值以及AUC值方面都更优，实现了对不同算法的取长补短，获得了泛化性能更优的融合模型。

附录见本刊网络版（http://www.epae.cn）。

参考文献：

[1] 危娟. 如何做好专变用户电费回收管理工作[J]. 数字通信世界，2018(12)：219.

[2] 鞠平，周孝信，陈维江，等. “智能电网+”研究综述[J]. 电力自动化设备，2018，38(5)：2-11.
JU Ping，ZHOU Xiaoxin，CHEN Weijiang，et al. “Smart Grid Plus” research overview[J]. Electric Power Automation Equipment，2018，38(5)：2-11.

[3] 涂莹，林士勇，欧阳柳，等. 基于市场细分的逻辑回归模型在电费回收风险预测中的应用研究[J]. 电力需求侧管理，2016，18(4)：46-49.
TU Ying，LIN Shiyong，OUYANG Liu，et al. Research on the risk prediction of electricity fee recovery using logistic regression based on the market segmentation theory[J]. Power Demand Side Management，2016，18(4)：46-49.

[4] 陈羽中，郭松荣，陈宏，等. 基于并行分类算法的电力客户欠费预警[J]. 计算机应用，2016，36(6)：1757-1761.
CHEN Yuzhong，GUO Songrong，CHEN Hong，et al. Electricity customers arrears alert based on parallel classification algorithm[J]. Journal of Computer Applications，2016，36(6)：1757-1761.

[5] 仲春林，郑飞，邵俊，等. 基于改进支持向量机的电费回收风险预警[C]//2018电力行业信息化年会. 银川，中国：人民邮电出版社，2018：201-205.

[6] 周晖，王毅，王玮，等. 基于Logistic回归模型的电力客户欠费违约概率的预测[J]. 电网技术，2007，31(17)：85-88.
ZHOU Hui，WANG Yi，WANG Wei，et al. Predication of default probability of clients’ electricity charges arrears based on Logistic regression model[J]. Power System Technology，2007，31(17)：85-88.

[7] 李晓蕾，魏玲，王忠强，等. 基于改进随机森林的电力用户欠费风险分析预警[J]. 电测与仪表，2019，56(9)：56-62.
LI Xiaolei，WEI Ling，WANG Zhongqiang，et al. Arrears risk analysis and early warning of electricity customers based on optimized random forest[J]. Electrical Measurement & Instrumentation，2019，56(9)：56-62.

[8] 谢林枫，钱立军，季聪，等. 基于长短期记忆网络算法的电费回收风险预警[J]. 电力工程技术，2018(5)：98-103.
XIE Linfeng，QIAN Lijun，JI Cong，et al. Application of long short-term memory network algorithm in tariff recovery risk early warning for large power customers[J]. Jiangsu Electrical Engineering，2018(5)：98-103.

[9] 宋连峻，徐志勇. 基于集对分析和马尔可夫链的电力客户信用风险评估[J]. 电力自动化设备，2009，29(12)：37-40.
SONG Lianjun，XU Zhiyong. Assessment of power customer credit risk based on set pair analysis and Markov chain model[J]. Electric Power Automation Equipment，2009，29(12)：37-40.

[10] 辛永，刘燕秋，黄文思，等. 基于多模型融合的客户投诉风险预测方法[J]. 电力大数据，2018(11)：31-37.
XIN Yong，LIU Yanqiu，HUANG Wensi，et al. Customer complaint risk prediction method based on multi-model fusion[J]. Power Systems and Big Data，2018(11)：31-37.

[11] 周子程. 基于Stacking模型融合的电信客户信用度模型研究与设计[D]. 广州：华南理工大学，2018.
ZHOU Zicheng. The research and design of telecom customer credit model based on Stacking model fusion[D]. Guangzhou：South China University of Technology，2018.

[12] 李刚，于长海，范辉，等. 基于多级决策融合模型的电力变压器故障深度诊断方法[J]. 电力自动化设备，2017，37(11)：138-144.
LI Gang，YU Changhai，FAN Hui，et al. Deep fault diagnosis of power transformer based on multilevel decision fusion model[J]. Electric Power Automation Equipment，2017，37(11)：138-144.

[13] 刘云鹏，付浩川，许自强，等. 基于AdaBoost-RBF算法与DSmT的变压器故障诊断技术[J]. 电力自动化设备，2019，39(6)：166-172.
LIU Yunpeng，FU Haochuan，XU Ziqiang，et al. Transformer fault diagnosis technology based on AdaBoost-RBF algorithm and DSmT[J]. Electric Power Automation Equipment，2019，39(6)：166-172.

[14] 石洪波，陈雨文，陈鑫. SMOTE过采样及其改进算法研究综述[J]. 智能系统学报，2019，14(6)：1073-1083.
SHI Hongbo，CHEN Yuwen，CHEN Xin. Summary of research on SMOTE oversampling and its improved algorithms[J]. CAAI Transactions on Intelligent Systems，2019，14(6)：1073-1083.

[15] 张良均. Python数据分析与挖掘实战[M]. 北京：机械工业出版社，2015：144-163.

[16] 马健. 多模型融合学习方法与应用[D]. 南京：南京大学，2016.
MA Jian. Multi-model integrated learning methods and applications[D]. Nanjing：Nanjing University，2016.

[17] 史佳琪，张建华. 基于多模型融合Stacking集成学习方式的负荷预测方法[J]. 中国电机工程学报，2019，39(14)：4032-4041.
SHI Jiaqi，ZHANG Jianhua. Load forecasting based on multi-model by Stacking ensemble learning[J]. Proceedings of the CSEE，2019，39(14)：4032-4041.

[18] 王向鹏. 基于不平衡三分类LGBM模型的贷后风险预警研究

[D]. 兰州:兰州大学,2019.
WANG Xiangpeng. Research on post-loan risk warning based on unbalanced three-classification LGBM model[D]. Lanzhou: Lanzhou University,2019.

作者简介:

潘国兵(1978—),男,湖北襄樊人,副教授,博士,主要研究方向为光伏发电及微电网控制、电力系统数据挖掘(**E-mail**:gbpan@zjut.edu.cn);


潘国兵

龚明波(1975—),男,浙江宁波人,高级工程师,主要研究方向为电能计量(**E-mail**:nb_gmb@163.com);

贺 民(1970—),男,浙江宁波人,高级工程师,主要研究方向为电能计量(**E-mail**:nbdyhm70@163.com)。

(**编辑** 王锦秀)

Identification method of electricity charge recovery risk of specialized transformer user based on Stacking model fusion

PAN Guobing[1],GONG Mingbo[2],HE Min[2],WU Chenghuan[2],TANG Xiaoqi[3],YANG Lü[1],OUYANG Jing[1]
(1. Institute of Distributed Energy and Microgrid,Zhejiang University of Technology,Hangzhou 310012,China;
2. Ningbo Power Supply Company of State Grid Zhejiang Electric Power Co.,Ltd.,Ningbo 315000,China;
3. Zhejiang Huayun Information Technology Co.,Ltd.,Hangzhou 310012,China)

Abstract: Aiming at the current electricity charge recovery risk of specialized transformer users faced by electricity companies, an identification method of electricity charge recovery risk of specialized transformer users is proposed based on Stacking model fusion. Feature processing, feature construction and feature selection are carried out for the data of specialized transformer users, and the model generalization performance is optimized from the sample distribution and feature attributes. Stacking model is used to integrate multiple base learners to construct an identification model of electricity charge recovery risk for specialized transformer users. The experimental results show that, compared with other commonly used classification algorithms, the proposed method has better precision, recall rate, *P-R* harmonic mean value, AUC(Area Under Curve) value, and model generalization performance, and has higher identification rate of specialized transformer risk users.

Key words: specialized transformer user; electricity charge recovery; risk identification; Stacking model fusion; LGBM

(上接第 134 页 continued from page 134)

Multi-objective planning of microgrid with electric vehicle charging load based on information gap decision theory

PENG Qiao[1],WANG Xiuli[1],SHAO Chengcheng[1],SHI Shuo[1],QI Shixiong[1],WANG Zhidong[2],YAN Sheng[3]
(1. School of Electrical Engineering,Xi'an Jiaotong University,Xi'an 710049,China;
2. State Grid Economic and Technological Research Institute Co.,Ltd.,Beijing 102209,China;
3. State Grid Corporation of China,Beijing 102209,China)

Abstract: In order to build the microgrid considering both the charging of electric vehicles at the operation level and the long-term increasing of load at the planning level, the economic costs of microgrid planning are analyzed. To deal with the contradiction between load fluctuation and economic benefit, a multi-objective planning model is proposed. Meanwhile to deal with the uncertainty of long-term increasing of load, the information gap decision theory is used for simulation. Then through fuzzy membership function weighted fuzzy programming, the maximum satisfaction programming model of microgrid with charging control of electric vehicle considering uncertainty is established to coordinate the planning of microgrid. Finally, the simulative results show that the proposed multi-objective programming model can balance the economy and load fluctuation, which can help the planning decision-maker to cope with the uncertainty of load at the optimal cost. Furthermore, the impact of interaction between microgrid and large grid on the access of electric vehicles is further studied.

Key words: microgrid; electric vehicles; multi-objective weighted fuzzy programming; information gap decision theory; uncertainty

附录：

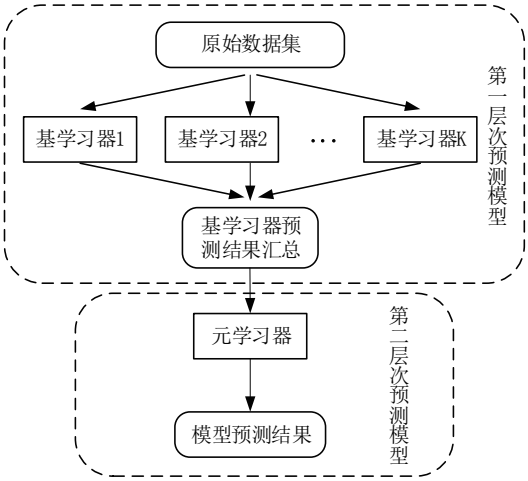

图 A1 Stacking 模型融合示意图

Fig.A1 Schematic diagram of Stacking model fusion

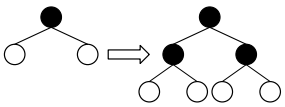

图 A2 XGBoost 的 Level-wise 叶子生长策略

Fig.A2 Level-wise leaf growth strategy of XGBoost

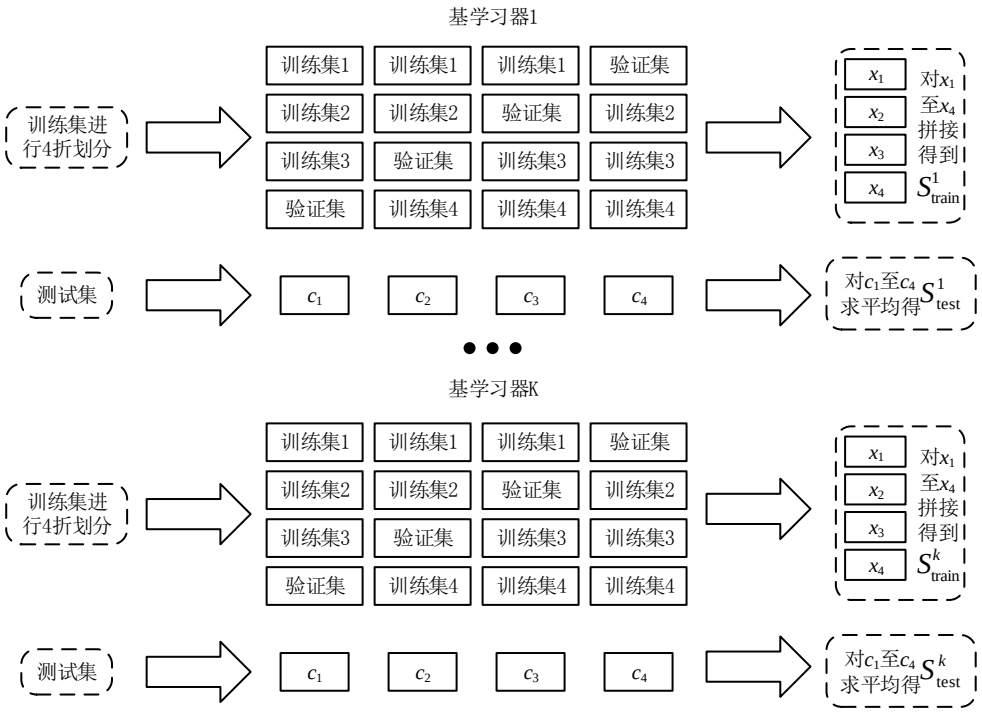

图 A3 Stacking 第一层原理解析图

Fig.A3 Principle diagram of Stacking at Level 1

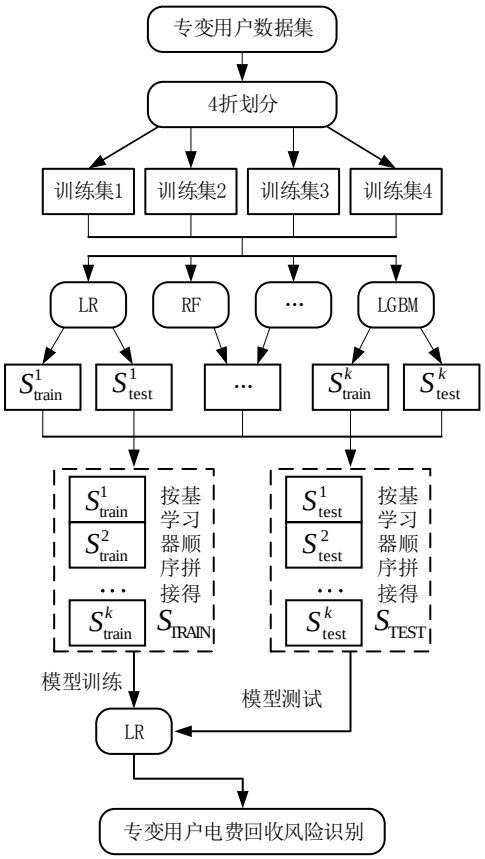

图 A4 基于 Stacking 模型融合的专变用户电费回收风险识别示意图

Fig.A4 Schematic diagram of risk identification of electricity charge recovery of specialized transformer users based on Stacking model fusion