基于 LightGBM 算法的移动用户信用评分研究

国强强,朱振方
(山东交通学院 信息科学与电气工程学院,山东 济南 250357)

摘 要:随着科技进步、社会的发展,个人信用分值对于个人愈加重要,而传统的信用评分主要以个人消费能力等少数的维度来衡量,难以全面、客观、及时地反映个人的信用。旨在解决面向大样本、高维度数据的环境下的信用分预测问题,提出一种基于 LightGBM 算法的移动用户信用评分算法,完善信用评分体系。首先分析线性相关性来构建特征集合,然后通过 K-means 算法对特征集合进行聚类分析,最后通过 LightGBM 模型构建信用评分模型。通过在数字中国创新大赛所提供的真实数据上的实验表明,该方法能够充分挖掘数据特征并且精准地预测用户信用评分,较 GBDT、XGBoost 等算法具有较高的准确率和计算效率。通过对线性相关性分析基础上的数据特征集合进行聚类分析,并将其应用到基于 LightGBM 信用评分模型,能够更加准确地预测移动用户信用评分。

关键词:评分预测;LightGBM 算法;K-means 算法;特征数据;线性相关性;随机森林;信用评分
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2020)09-0210-06
doi:10.3969/j.issn.1673-629X.2020.09.038

Research on Mobile User Credit Score Based on LightGBM Algorithm

GUO Qiang-qiang, ZHU Zhen-fang
(Department of Information Science and Electrical Engineering, Shandong Jiaotong University, Jinan 250357, China)

Abstract: With the progress of science and technology and the development of society, personal credit score is becoming more and more important to individuals. However, the traditional credit score is mainly measured by a few dimensions such as personal consumption ability, which is difficult to reflect personal credit comprehensively, objectively and timely. In order to address the problem of credit score prediction in the environment of large sample and high-dimensional data, we propose a mobile user credit score algorithm based on LightGBM algorithm to improve the credit scoring system. The linear correlation is firstly analyzed to construct feature sets, and then the K-means algorithm is used to analyze the clustering of feature sets. Finally, the credit scoring model is built by LightGBM model. Experiments on real data provided by the digital China innovation competition shows that the proposed method can fully mine data features and accurately predict user credit score, which is more accurate and efficient than GBDT, XGBoost and other algorithms. By clustering the data feature set based on linear correlation analysis and applying it to LightGBM credit scoring model, mobile users' credit scores can be predicted more accurately.

Key words: score prediction; LightGBM algorithm; K-means algorithm; data features; linear dependence; random forest; credit scoring

0 引 言

随着社会信用体系建设的深入推进,社会信用标准建设飞速发展,相关的标准相继发布。但是,一个包括信用服务标准、信用数据采集和服务标准、信用修复标准、城市信用标准、行业信用标准等在内的多层次标准体系亟待出台,社会信用标准体系有望快速推进。社会信用体系建设是一个系统工程,完善信用评分体系有助于推动整个社会的信用体系升级。个人信用评估构成是社会信用评估体系的基础,构建科学的个人信用评估体系是构建科学社会信用评估的基础,而移动用户信用评估,则是个人信用评估中最重要组成部分之一。随着科技的进步、社会的发展,个人信用分值对于个人愈加重要,而传统的信用评分主要以个人消费能力等少数的维度来衡量,难以全面、客观、及时地反映个人的信用。如今电子商务和互联网金融蓬勃发展,在大数据背景下个人信用评价也需满足时代要求向大数据方向转变。

文中算法旨在解决面向大样本、高维度数据环境

收稿日期:2019-11-11 **修回日期**:2020-03-12
基金项目:国家社科基金(19BYY076);教育部人文社科基金(14YJC860042);山东省社科规划项目(19BJCJ51,18CXWJ01,18BJYJ04,17CHLJ07C)
作者简介:国强强(1995-),男,硕士研究生,研究方向为数据分析与自然语言处理;通讯作者:朱振方,副教授,研究方向为自然语言处理。

下的信用分预测问题,提出一种基于 LightGBM 算法的移动用户信用评分:K-LGB 模型,实现移动用户信用评分。通过该算法可以有效提高信用分预测的准确性,同时又可以提高算法执行效率。

1　相关研究

评分预测问题[1]属于推荐系统中的一个分支,推荐系统的性能很大程度上受评分预测准确性的影响。随着国内外学者的深入研究,信用评估发展出来统计方法和非统计方法两大类[2]。非统计方法包括神经网络、遗传算法、专家系统等,统计方法包括逻辑回归、线性回归、非线性回归、近邻估计等。很多学者早期通过用户历史评分行为和物品属性特征进行建模[3]来解决评分预测问题,在已有研究中,Maher Alarajden 等人[4]将神经网络、支持向量机、随机森林、决策树、Logistic 回归和朴素贝叶斯与 LR 结合使用,达到了很好的效果。到目前为止,Maher Alarajden 所提出的信用评估体系,仍然被认为是信用评分模型的行业标准模型。Maysam F. Abbod 等人[5]提出在数据预处理上将 Gabriel 近域图编辑和多变量自适应回归样条方法融合的算法来实现预测信用分,另外,还提出了一种基于集合建模阶段不同分类算法的共识方法的新分类器组合规则。Luo Cuicui 等人[6]将信念网络与限制玻尔兹曼机器等深度学习算法与当前流行机器学习算法(如逻辑回归、支持向量机、多层感知机)进行比较,发现使用分类精度和接收器工作特性曲线下的面积评估性能中 DBN 的性能最佳。Leong C K 等人[7]提出了一种贝叶斯网络模型,用于解决信用风险评分中的截尾样本、样本不平衡、实时实现等问题,相较于竞争模型(逻辑回归与神经网络)在精度、灵敏度等几个维度上表现更佳。

随着机器学习技术的快速发展,国内学者的研究更侧重对这些模型的组合及应用。综合应用多种机器学习方法进行信用评分,正逐渐成为主要手段,能够解决单个算法结果准确率不足的问题,获得更优的预测结果。例如,姜明辉[8]、王磊等人[9]通过改进 Logistic 模型,建立信用评分模型,取得了较好的效果。近年来,随着信用评估研究的深入,引入了人工智能等非统计方法,学者们的研究重心转向了集成学习算法和神经网络(NNs)、支持向量机(VSM)等算法。现有研究结果显示,根据训练数据构建一组个体学习器,并采用某种策略将多个学习器进行集成的学习方法,比较逻辑回归、决策树等单一分类器和神经网络评估模型[10]和模糊分析评估模型,具有更高的准确度和更好的稳健性[11]。

集成学习方法主要分为两大类,即 Bagging 方法[12](如 RF 算法等)与 Boosting 方法[13](如 LightGBM[14])。其中,RF[15]算法利用样本扰动和属性扰动实现基学习器的多样性,虽然提升了算法的泛化性能,但该算法需要存储每棵决策树及其每个节点不同的样本集合,内存开销较大,导致模型训练速度较慢。相比之下,LightGBM 具有更快的训练速度、更低的内存消耗、更好的模型精度、支持并行学习、可以快速处理海量数据等优点[16]。鉴于此,文中基于 LightGBM 算法构建信用评分模型,进行中国移动用户信用分预测。

2　基于 LightGBM 算法的移动用户信用评分研究

现有的信用评分模式往往只采用集成学习中的 Bagging 方法(如 RF 算法)或者 Boosting 方法(如 LightGBM),在多维度特征提取、线性关系挖掘等方面存在很大的局限性。鉴于此,在面对大样本、多维度的数据环境下,为了解决模型过拟合问题,构造有效的特征信息、提高模型信用评分准确性,文中提出一种 K-LGB 模型,实现移动用户信用评分。首先通过分析线性相关性来构建特征集合,然后通过 K-means 算法对特征集合进行聚类分析,将特征集合聚类分析结果作为有效特征信息加入数据集,最后将加入有效特征信息的数据集作为 LightGBM 模型的输入,通过 LightGBM 模型得出信用评分。算法流程如图 1 所示。

2.1　线性相关性分析

经研究发现分析线性相关性不仅可用来解决模型过拟合问题,而且可以解决多维度特征提取、线性关系挖掘的问题。鉴于此,文中采用皮尔逊相关系数来进行线性相关性分析。皮尔逊相关系数(Pearson correlation coefficient)又称皮尔逊积矩相关系数,在统计学中常用来度量两组数据间的相关程度。皮尔逊相关系数的值介于-1 与 1 之间,绝对值越大,线性相关性越强;绝对值越接近于 0,线性相关性越弱。假设给定包含 i 个项的数据集 $X=\{x_1,x_2,\cdots,x_i\}$ 和 $Y=\{y_1,y_2,\cdots,y_i\}$,则皮尔逊相关系数公式如下:

$$r_{xy}=\frac{n\sum x_iy_i-\sum x_i\sum y_i}{\sqrt{n\sum x_i^2-(\sum x_i)^2}-\sqrt{n\sum y_i^2-(\sum y_i)^2}} \tag{1}$$

其中,n 为变量取值个数,r_{xy} 为数据集 X、Y 的皮尔逊相关系数值。

具体到本次评测,首先分别计算特征之间、特征与信用分之间的皮尔逊相关系数,确定它们的线性相关性,然后选择与信用分线性相关性比较强的特征,最后将经过线性相关性分析的特征集合作为下一步 K-

means 聚类算法的输入。部分数据特征与信用分线性相关性如表 1 所示。

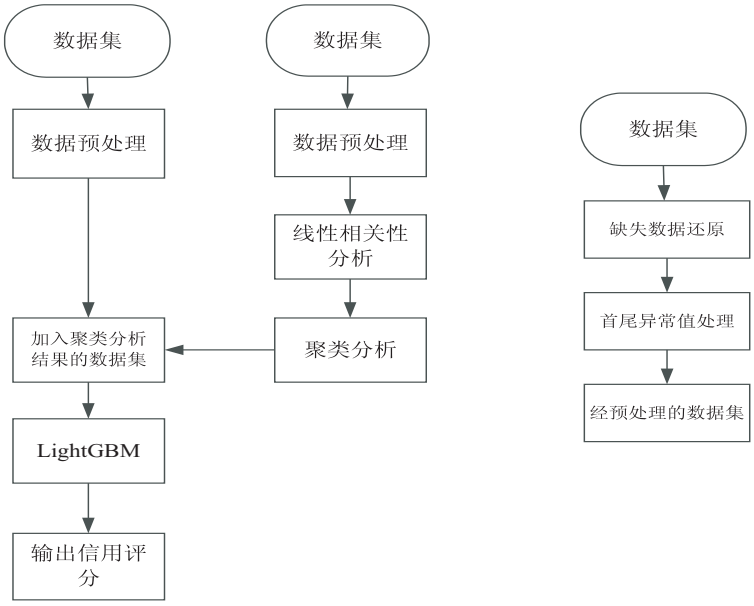

图 1 算法流程

表 1 部分数据特征与信用分线性相关性

数据特征	线性相关性值
用户网龄(月)	0.55
用户近 6 月平均消费值(元)	0.49
当月通话交往圈人数	0.48
……	……
当月是否景点旅游	0.27
是否 4G 不健康用户	-0.15
当月话费敏感度	-0.24

经过线性相关性分析,发现“用户网龄(月)”、“用户近 6 月平均消费值(元)”、“当月通话交往圈人数”、“当月是否景点游览”等 7 个特征与信用分具有较强的线性相关性。因此,选择这部分特征集合进行进一步的分析。

2.2 基于特征集合的 K-means 聚类

2.2.1 K-means 聚类分析

聚类算法可以分为基于划分、层次、密度的方法。其中,基于层次的聚类方法,如 hierarchical methods,有两种类型:合并的层次聚类和分裂的层次聚类,该方法可解释性好,时间复杂度高,较为适用于小数量级聚类分析。基于密度的聚类方法,如 DBSCAN[17],解决了不规则形状的聚类问题,对于噪声数据不敏感,能发现任意形状的聚类结果,但是该方法对于参数设置非常敏感。基于划分的聚类方法,如 K-means 方法[18](K-均值),虽然对数据集中噪声、离群值、初始值设置较为敏感,但是该方法较为适合欧氏空间中按向量和欧氏距离定义的样本聚类,对于处理大型数据较为高效(时间复杂度、空间复杂度),因此,文中采用 K-means 算法作为聚类分析的方法。

假设给定的数据集 $X = \{x_m \mid m = 1,2,\cdots,h, h \in R\}$,$Y$ 中样本有 n 个属性(维度)$A_1, A_2, \cdots, A_n$,则欧氏距离公式如下:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2} \tag{2}$$

$d(x_i, x_j)$ 距离越小,样本 x_i 和 x_j 相似度高,差异度小;$d(x_i, x_j)$ 距离越大,样本 x_i 和 x_j 相似度低,差异度大。

K-means 聚类算法一般使用误差平方和作为标准测度函数,具体定义如下:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} | p - m_i |^2 \tag{3}$$

其中,p 为代表对象的空间的一个点,m_i 为聚类 C_i 的均值(p 和 m_i 均为多维的)。其中 E 为数据集中所有对象的平方误差和,对于不同聚类 E 的大小也会不同,因此算法需要将 E 调整到最小,使得聚类达到最优。

K-means 是属于划分方法的聚类算法,是一种经典的聚类算法。由于算法简单快捷,所以在工业界中应用比较广泛。其优点主要为:算法尽量使确定的 K 个划分达到平方和误差最小;当聚类的数据是密集的(凸型的),并且簇与簇之间的数据差异较大,算法的聚类效果较好;当处理大量数据集时,算法高效并且相对可以伸缩。

2.2.2 基于线性相关性分析结果的聚类分析

如前所述,构造有效特征信息方法流程如下:

(1)聚类算法的选择:不同的聚类算法有不同的优劣,将数据的属性(算法是否独立于数据输入顺序;数据维度)、算法处理能力(算法复杂度)作为聚算法

选择依据。对比聚类算法中基于层次的方法(hierarchical methods)、基于划分的方法(K-means)、支持向量机(SVM)等,最终选取基于划分的方法(K-means)作为文中模型的聚类算法。

(2)K-means 聚类算法的输入:线性相关性分析结果(与信用分具有较强的线性相关性的 N 维特征集合)、聚类簇的个数 K(K 值为 4)。

(3)K-means 聚类算法的输出:有效特征信息(1 维),K-means 聚类算法结果样例如表 2 所示。

表 2　K-means 聚类分析结果

用户网龄(月)	用户近 6 月平均消费值(元)	当月通话交往圈人数	…	K-means 聚类分析结果
186	163.86	83	…	3
145	109.64	70	…	3
62	162.98	77	…	2
78	98.33	45	…	1
10	78.86	29	…	0

2.3　LightGBM

假设训练集样本为 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_m,y_m)\}$,其中 $x_i \in T$ 为训练集中第 i 个样本,$\hat{y}_i$ 为预测值,y_i 为真实值,l 为损失函数。在 t 步模型对 x_i 进行预测,如式(4)所示,其目标函数如式(5)所示:

$$\hat{y}_i^t=\hat{y}_i^{t-1}+f_t(x_i) \tag{4}$$

$$\mathrm{Obj}^{(t)}=\sum_{i=1}^{n} l(y_i,\hat{y}_i^t)+\sum_{i=1}^{t}\Omega(f_i) \tag{5}$$

其中,$\Omega(f_i)$ 为正则项,f_i 为一棵决策树。

将损失函数设为平方损失,则目标函数为:

$$\mathrm{Obj}^{(t)}=\sum_{i=1}^{n}(y_i-(\hat{y}_i^{t-1}+f_t(x_i))^2)+\Omega(f_t)+\mathrm{constant} \tag{6}$$

如前所述,无关和冗余变量会对模型预测的准确性造成不利影响,选择有效的特征信息,直接决定了信用评分模型的准确性。鉴于此,将 K-means 聚类算法输出(构造的有效特征信息)手动加入数据集,作为新的特征列。融入新特征列的数据集作为 LightGBM 模型的输入,具体 LightGBM 信用评分模型训练流程如下所示:

输入:K-means 聚类算法的输出作为有效特征信息,作为新特征列,手动加入到数据集中。加入新特征列的数据集,作为 LightGBM 模型输入。

输出:移动用户预测信用分。

算法步骤:

(1)算法确定目标函数,将损失函数设为平方损失,通过贪心策略生成决策树的每个节点,找到最佳树结构。

(2)算法每次迭代前计算损失函数样本点的一阶导数和二阶导数,生成新的决策树并计算每个节点的预测值。

(3)将迭代生成的 N 棵决策树迭代加入模型中,初始化 N 棵决策树,平均分配训练样例权重。

(4)训练弱分类器,更新权重得到最终分类器,输出移动用户预测信用分。

3　实验及分析

3.1　实验数据与预处理

3.1.1　实验数据与实验设定

实验采用的是 2019 数字中国创新大赛(https://www.datafountain.cn/)中赛题“消费者人群画像—信用智能评分”的数据集,该数据集是中国移动福建公司提供的 2018 年 x 月份的样本数据(脱敏),包括客户的各类通信支出、欠费情况、出行情况、消费场所、社交、个人兴趣等丰富的多维度(30 维度)数据。其中训练集 50 000 条,测试集 50 000 条。实验配置与环境如表 3 所示。

表 3　实验配置与环境

实验环境	环境配置
操作系统	Windows10
CPU	Intel(R) Core(TM) i7-5500U 2.40 GHz
RAM	8.00 GB
编程语言	Python 3.6

3.1.2　数据分析预处理

在数据集中,不同维度的特征虽然具有不同的量纲,但是特征数值应该具有正确性和有效性。通过对数据集的统计分析,发现数据集中存在数据缺失[19]和首尾异常值的问题,导致特征数值失去有效性和正确性,因此需要对数据集进行缺失数据还原和首尾异常值处理。

3.2　评测指标

评价用户信用评分模型有很多指标,如准确率(Accuracy)、查全率(Recall)、F 得分、MAE、ROC 曲线和精确度(Precision)。为了验证该模型的性能,选择 MAE 和 ROC 曲线和 AUC(area under curve)作为该模

型的评价指标。将 MAE 转换成了 Score 指标,具体公式如下所示:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} | \mathrm{pred}_i - y_i | \tag{7}$$

$$\mathrm{Score} = \frac{1}{1 + \mathrm{MAE}} \tag{8}$$

其中, pred_i 为预测样本, y_i 为真实样本。MAE 的值越小,说明预测数据与真实数据越接近,所有 Score 的值越高评测效果越好。

3.3 实验结果与分析

3.3.1 *K* 值的选取

聚类结果依赖于初始值的设定,但是值的选定往往要经过很多次实验才能找到最佳聚类个数。目前 *K* 值的确定主要通过以下几种方法:

(1)凭经验选代表点,根据问题的性质、数据分布,从直观上找到较合理的 *K* 值。

(2)将全部样本随机分成类,计算每类重心,把这些重心作为每类的代表点,然后选取 *K* 值。

(3)按密度大小选取 *K* 值。

实验使用不同的 *K* 值进行评测结果对比,经实验结果发现, *K* 值为 4 时该模型评测结果为最优。

3.3.2 LightGBM 参数调整

LightGBM 模型参数虽然包含多类参数但是构造相对简单,参数设置与模型效果成正比关系,参数调节的越优模型效果越好。LightGBM 模型为用户提供了多类参数,并提供了便捷的 CV 函数供用户进行调参。在调整模型参数的过程中,文中将训练集拆分出 80% 作为新的训练集,剩余的 20% 数据作为新的测试集。依据新测试集的预测结果与真实结果误差微调参数,同时采用了 CV 函数,得到 LightGBM 模型最优参数。LightGBM 参数如表 4 所示。

表 4 LightGBM 参数

参数	值
learning_rate	0.01
objective	regression_l1
metric	mae
feature_fraction	0.6
bagging_fraction	0.8
bagging_freq	2
num_leaves	31
verbose	-1
max_depth	5
lambda_l2	5
lambda_l1	0

3.3.3 模型效果对比分析

为了验证文中方法的优越性,采用了评测指标 Score、预测准确度 ROC 曲线和 AUC。使用 LightGBM、XGBoost[20]、K-LGB、K- XGB 四种模型,通过评测指标 Score、执行效率、准确度进行实验结果对比,评测指标 Score 结果如表 5 所示。

表 5 模型评测 Score 结果与效率

模型	Score	执行效率/minute
K-LGB	6.412	8
XGBoost	6.340	15
LightGBM	6.276	6
K-XGB	6.391	20

由表 5 的实验结果显示,文中算法 Score 得分为 6.412,模型运行时间为 8 分钟,对比 LightGBM 模型 Score 提高了 5.412 个百分点。为了进一步对比预测准确度,对预处理后的 40 000 条有效数据采用 5 次五折交叉验证[21],分别建立信用评分模型,结果如表 6 所示。

表 6 五折交叉验证的预测准确度对比 %

次数	K-LGB	XGBoost	LightGBM	K-XGB
1	69.61	68.97	67.82	68.89
2	68.92	68.20	67.97	68.54
3	69.53	66.60	66.86	68.23
4	69.82	67.63	68.01	69.11
5	70.06	67.54	64.23	68.32
平均	69.58	67.78	66.97	68.61

图 2 为 4 种模型的 ROC[22] 曲线图。在 ROC 空间中,ROC 曲线下的面积为 AUC 值,AUC 值介于 0 和 1 之间,AUC 的值越高则模型信用评估性能越好。从图中可以看出,在相同的数据集与实验设备下,K-LGB 模型表现出了较好的信用评估性能,AUC 值为 0.85,较 LightGBM 模型提高了 0.15。

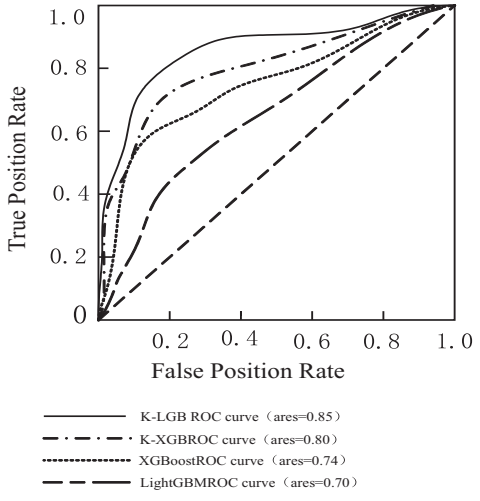

图 2 模型 ROC 曲线

该实验结果表明,文中算法评测结果和预测准确度优于其他算法,证实了算法的可行性和有效性。为了方便观察实验结果,执行效率以分钟为单位,由于评

测结果值为百分位小数,评测结果值放大 100 倍。把 K-means 算法与 LightGBM 算法相融合的模型称为 K-LGB,K-means 算法与 XGBoost 算法相融合的模型称为 K-XGB。

4　结束语

基于线性相关性分析结果进行聚类分析,充分挖掘数据特征,以 LightGBM 算法为典型的大数据技术,进行中国移动用户信用分预测。在数据预处理方面,针对数据缺失问题采用还原为 NaN 的方法,针对数据首尾异常值问题采用设置上下限的方法。在数据集大样本、高维度的环境下,与 GBDT、XGBoost 等算法进行对比,结果表明该算法具有较好的预测准确度和计算效率,适合处理大规模数据。

参考文献:

[1] 杨贵军,徐　雪,赵富强. 基于 XGBoost 算法的用户评分预测模型及应用[J]. 数据分析与知识发现,2019,3(1):118-126.

[2] 金　欣. 个人信用评估指标及模型研究—基于 XGBoost-BOA 集成分类模型[D]. 杭州:浙江财经大学,2017.

[3] 邓晓懿,金　淳,韩庆平,等. 基于情境聚类和用户评级的协同过滤推荐模型[J]. 系统工程理论与实践,2013,33(11):2945-2953.

[4] ALARAJ M,ABBOD M F. A new hybrid ensemble credit scoring model based on classifiers consensus system approach [J]. Expert Systems with Applications,2016,64:36-55.

[5] ALARAJ M,ABBOD M. Classifiers consensus system approach for credit scoring[J]. Knowledge-Based Systems,2016,104:89-105.

[6] LUO C,WU D,WU D. A deep learning approach for credit scoring using credit default swaps[J]. Engineering Applications of Artificial Intelligence,2017,65:465-470.

[7] LEONG C K. Credit risk scoring with Bayesian network models [J]. Computational Economics,2016,47(3):423-446.

[8] 姜明辉,许　佩,韩旖桐,等. 基于优化 CBR 的个人信用评分研究[J]. 中国软科学,2014(12):148-156.

[9] 王　磊,范　超,解明明. 数据挖掘模型在小企业主信用评分领域的应用[J]. 统计研究,2014,31(10):89-98.

[10] 冯　婧. 基于 BP 神经网络的个人信用风险评估模型的研究[D]. 太原:太原理工大学,2017.

[11] BROWN I,MUES C. An experimental comparison of classification algorithms for imbalanced credit scoring data sets[J]. Expert Systems with Applications,2012,39(3):3446-3453.

[12] 肖连杰,郜梦蕊,苏新宁. 一种基于模糊 C-均值聚类的欠采样集成不平衡数据分类算法[J]. 数据分析与知识发现,2019,3(4):90-96.

[13] MA X,SHA J,WANG D,et al. Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning[J]. Electronic Commerce Research and Applications,2018,31:24-39.

[14] 张丹峰. 基于 LightGBM,XGBoost,ERT 混合模型的风机叶片结冰预测研究[D]. 上海:上海师范大学,2018.

[15] BREIMAN L,BREIMAN L,CUTLER R A. Random forests machine learning[J]. Journal of Clinical Microbiology,2001,2:199-228.

[16] 沙靖岚. 基于 LightGBM 与 XGBoost 算法的 P2P 网络借贷违约预测模型的比较研究[D]. 大连:东北财经大学,2017.

[17] 宋金玉,郭一平,王　斌. DBSCAN 聚类算法的参数配置方法研究[J]. 计算机技术与发展,2019,29(5):44-48.

[18] 吴夙慧,成　颖,郑彦宁,等. K-means 算法研究综述[J]. 现代图书情报技术,2011(5):28-35.

[19] 肖　江,陈璐瑜. 改进的 P2P 信贷借款人信用风险的研究[J]. 信息技术,2016,40(11):212-214.

[20] CHEN T,GUESTRIN C. Xgboost:a scalable tree boosting system[C]//Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. San Francisco:ACM,2016:785-794.

[21] LITTLE M A,VAROQUAUX G,SAEB S,et al. Using and understanding cross-validation strategies. Perspectives on Saeb et al. [J]. Gigascience,2017,6(5):1-6.

[22] 李晓刚. 个人信用风险评估的一种基于 XGBoost 的集成学习方法[D]. 合肥:中国科学技术大学,2018.