

# 基于组合特征选择的随机森林信用评估<sup>①</sup>



饶姗姗, 冷小鹏

(成都理工大学 计算机与网络安全学院 (牛津布鲁克斯学院), 成都 610051)

通信作者: 冷小鹏, E-mail: lengxiaopeng@cdut.edu.cn

**摘要:** 构建个人信用风险评估模型的过程中, 特征工程很大程度上决定了评估器的性能, 传统的特征选择方法无法全面的考虑高维度指标对评估结果的影响, 且大多数研究在构建模型的过程中人为决定特征集大小, 导致随机性强、可信度低; 基于此, 提出基于传统风控指标优化 XGBoost 的随机森林模型 (IV-XGBoostRF), 将传统风控指标 IV 与 XGBoost 相结合对原始特征集进行筛选, 建立较为完善的信用评估模型. 通过对比实验的结果显示改进后的随机森林模型准确度提高了 0.90%, 且其他各项评估指标均优于传统信用评估模型, 证明了该组合特征选择方法的可行性, 有一定的应用价值.

**关键词:** 信用评估; 信息价值; 组合特征选择; 随机森林; XGBoost

引用格式: 饶姗姗, 冷小鹏. 基于组合特征选择的随机森林信用评估. 计算机系统应用, 2022, 31(3): 345-350. <http://www.c-s-a.org.cn/1003-3254/8371.html>

## Random Forest Credit Evaluation Based on Combination Feature Selection

RAO Shan-Shan, LENG Xiao-Peng

(School of Computer and Network Security (Oxford Brookes College), Chengdu University of Technology, Chengdu 610051, China)

**Abstract:** In the process of building a personal credit risk evaluation model, feature engineering largely determines the performance of the evaluator. Traditional feature selection methods cannot fully consider the impact of high-dimensional indicators on the evaluation results, and most studies artificially determines the size of the feature set in the process of building the model, leading to high randomness and low credibility. Therefore, a random forest model (IV-XGBoostRF) based on traditional risk control indicators to optimize XGBoost is proposed. The traditional risk control indicators IV and XGBoost are combined to screen the original feature set to build a relatively complete credit evaluation model. The results of comparison experiments show that the accuracy of the improved random forest model is increased by 0.90%, and other evaluation indicators are better than the traditional credit evaluation model, which proves the feasibility of the feature selection method and has certain application value.

**Key words:** credit evaluation; information value; combination feature selection; random forest; XGBoost

当下是金融经济的飞速发展时代, 个人信用评估数据普遍表现为高维度、高复杂度等特点, 无关、冗余的评价因子会直接影响信用评估模型的准确性, 而评价因子的优劣取决于特征选择方法的选择是否合理、全面. 目前常用的特征选择方法包括过滤法

(filter)、包裹法 (wrapper)、嵌入法 (embedded)<sup>[1]</sup>. Filter 方法利用统计学方法评估变量与预测变量间的关系, 该方法优势在于计算简单、速度快; wrapper 方法通过对生成的子特征组合与其他组合进行比较, 依赖于学习算法的准确度; embedded 方法是通过在训练

<sup>①</sup> 基金项目: 四川省科技厅应用基础研究项目 (2021YJ0335)

收稿时间: 2021-05-17; 修改时间: 2021-06-14; 采用时间: 2021-06-30; csa 在线出版时间: 2022-01-24

模型的过程中根据准确度学习并选择出最优的特征组合. 本文选用的信息量模型 (information value) 与 XGBoost (extreme gradient boosting) 分别是 Filter 和 Embedded 方法的实现. 结合两种不同的特征选择方法不仅能够提高输入变量的有效性同时增强了整个模型的可解释性.

近二十年来, 信用风险评估方法主要分为两部分: 统计学方法和机器学习方法. Fernandes 等<sup>[2]</sup> 利用逻辑回归算法建立了信用评估模型, 并通过实验证实了 logistic 算法在信贷评估业务的可行性, 成为了主流的统计学方法之一. 郭畅<sup>[3]</sup> 将 IV 与 Lasso-Logistic 结合, 通过 IV 排除风险识别能力、稳定性较差的变量, 整体提升了信用预测模型的效果. 然而随着信息技术的迅猛发展, 传统的统计学评分模型暴露出 3 个主要问题: 一是模型单一; 二是处理数据维度较小; 三是主要以专家评价法为主, 预测结果缺乏客观性<sup>[4]</sup>; 为了解决这些问题, 学者们尝试将机器学习建模技术应用在个人信用评估业务中, 其中主要包括决策树<sup>[5]</sup>、BP 神经网络<sup>[6]</sup>、支持向量机 (SVM)<sup>[7]</sup> 等模型, 但单一机器学习模型处理的数据维度和预测精度有限, 为决定该问题, 集成算法逐渐应用在个人信用评估领域中. Twala<sup>[8]</sup>、Zhu 等<sup>[9]</sup> 使用 5 类信用数据证实了集成算法在信用评估领域表现出更优的预测效果并且多个分类器组合显著提高了整个模型的学习能力; 萧超武等<sup>[10]</sup> 基于组合分类模型随机森林 (RF), 发现 RF 模型分类准确率、稳定性更高, 并且噪声容忍度高, 训练过程中能够避免过拟合现象. 周永圣等<sup>[11]</sup> 首次将 XGBoost 和随机森林模型两种不同的集成算法融入到信用评估中, 证实了 XGBoost 算法基于特征重要性能有效剔除信用数据冗余变量, 但该实验缺少对比实验且变量选取较随意、可靠性较低. 李欣等<sup>[12]</sup> 提出一种基于改进网格搜索优化的 XGBoost 模型, 实验结果显示 *F-score* 和 *G-mean* 均优于其他机器学习模型, 进一步表明集成算法的优越性.

鉴于个人信用数据维度高、变量冗余度高的特点, 本文采用 IV-XGBoost 组合特征选择方法对随机森林模型进行优化, 不仅避免了传统特征选择方法的单一性同时能够更好解决数据维度过高的问题; IV 值剔除冗余、无预测能力的变量, XGBoost 利用变量打分机制筛选变量, 采样逐步排除法输出最优特征集, 相比根据特征重要性排名随意选取的方法, 该组合特征选择

模型更加合理并且可信度高; 此外超参数组合会直接影响 RF 模型的预测效果, 利用网格-5 折对 RF 中的重要参数进行参数寻优. 最后实验结果表明基于新型的组合特征选择方法的随机森林相比其他单一的机器学习方法以及原始的集成算法有着更高的稳定性和预测准确率.

## 1 特征选择理论与方法

### 1.1 证据权重 (WOE) 与信息价值 (IV)

WOE (weight of evidence) 即为证据权重, 通过编码的方式将自变量表示成其对目标变量的区分程度的形式. 根据 WOE 值大小反映某属性对目标的影响, 若同一属性的不同划分标准计算出的 WOE 值越大, 则表明该属性对目标属性的区分度越大, 其计算公式:

$$WOE(x) = \ln\left(\frac{py_i}{pn_i}\right) = \ln\left(\frac{y_i}{y_T} \frac{n_i}{n_T}\right) \quad (1)$$

其中,  $WOE(x)$  是变量进行分箱处理后第  $i$  组的 WOE 值, 其中  $py_i$  是该组未响应样本 ('未违约客户') 占比;  $pn_i$  是响应样本 ('违约客户') 占比;  $y_i$  是某变量第  $i$  个属性对应的未响应样本数,  $y_T$  是所有未响应样本数;  $n_i$  是某变量第  $i$  个属性对应的响应样本数,  $n_T$  是所有响应样本数.

信息量 (IV) 是基于 WOE 算法改进的衡量指标, 其计算本质是某变量所有属性的 WOE 值加权求和, 其值大小反映了变量对目标变量的预测能力. IV 通常应用在风控模型中评价因子的选择, 其计算公式:

$$\begin{aligned} IV &= \sum_i \left( \frac{y_i}{y_T} - \frac{n_i}{n_T} \right) \times \ln \left[ \frac{y_i}{y_T} \frac{n_i}{n_T} \right] \\ &= \sum_i \left( \frac{y_i}{y_T} - \frac{n_i}{n_T} \right) \times WOE_i \end{aligned} \quad (2)$$

其中,  $WOE_i$  是某变量进行离散化处理后第  $i$  组的 WOE 值; 其中  $y_i$  是某变量第  $i$  个属性对应的未响应样本数;  $T$  是所有未响应样本数;  $n_i$  是某变量第  $i$  个属性对应的响应样本数;  $n_T$  所有响应样本数.

### 1.2 XGBoost 特征选择

#### 1.2.1 算法原理

XGBoost<sup>[13]</sup> 是基于梯度提升算法 GBDT 改进的新型集成学习算法, 其主要算法思想是将多个分类精度较低的子树模型进行迭代组合从而构建出准确度、稳定性更强的模型. XGBoost 在 GBDT 的原目标函数上

加入了正则项,因此加快了收敛效率的同时降低过拟合风险.其变换后的公式如下:

$$H^{(i)} = \sum_{i=1}^n (l(y_i, \hat{y}_i) + \Omega(f_i)) \quad (3)$$

$$\Omega(f) = \gamma N + \frac{1}{2} \lambda \|O\|^2 \quad (4)$$

式(3)中 $l(y_i, \hat{y}_i)$ 是预测值 $\hat{y}_i$ 与 $y_i$ 之间的平方差损失函数.

式(4)计算所有子树的复杂度总和,其中 $\Omega(f)$ 是正则化项; $N$ 代表子树中叶子节点个数; $\lambda$ 代表了叶子节点权重 $O$ 的惩罚系统值; $\gamma$ 是衡量树的分割难度大小,用于控制树生长.

XGBoost与GBDT的不同之处在于前者是以泰勒公式二阶导展开,从而加快了函数收敛速度且提高了模型预测准确度,其变换后目标函数为:

$$H^i = -\frac{1}{2} \sum_{j=1}^N \frac{\left( \sum_{i \in I_j} g_i \right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma N \quad (5)$$

其中, $I_j \in \{q(X_i) = j\}$ ,  $h_i$ 为 $[l(\alpha)]''$ ,  $g_i$ 为 $[l(\alpha)]'$ .

### 1.2.2 XGBoost 特征选择原理

机器学习中特征选择是建模工作中至关重要的环节,XGBoost采用梯度提升的原理对样本进行分类,该模型是根据计算各个变量的重要性来进行特征选择,其主要原理是子树节点在分裂的同时计算其信息增益量,并选择差值最大的指标作为下一次的分裂属性,直到完成全部计算.本文根据XGBoost计算后输出的特征重要性排名,由低到高逐个加入到特征集并计算RF模型准确度,选取准确度最高的特征集作为最终特征选择的结果.

## 2 改进的随机森林个人信用评估模型

### 2.1 随机森林算法

随机森林是Breiman<sup>[14]</sup>提出的基于树的集成学习算法,根据特征数对每个样本选取分裂指标进而构建单棵子树.随机森林旨在集成多个弱分类器来构建一个强分类器,各个基分类器之间相互互补,降低了方差以及过拟合的风险,从而提高模型的性能.

RF是在Bagging集成学习和随机子空间的基础上进一步优化的集成学习算法,由服从独立同分布随

机向量 $\theta_i$ 生成的 $i$ 棵树 $\{W_i(x, \theta_i), i = 1, 2, 3, \dots\}$ , $i$ 棵子树最终形成集成树模型.RF模型的最终结果采用基分类器中平均票数最多的结果作为输出.随机森林模型算法流程如下:

(1)采用Bootstrap方法有放回的从总量为 $W$ 的训练集中随机抽取形成 $N$ 个子训练集 $\{N_i\}$ , $i \in \{1, 2, 3, 4, 5, \dots, k\}$ 每个训练子样本对应一棵CART树.

(2)随机森林由 $i$ 棵分类树构成,每棵分类树的子节点在进行分裂时随机选择分裂指标数 $n(n \leq M)$ ,其中 $M$ 为总样本的指标个数,根据衡量指标大小选择最优分割指标进行划分.

(3)不断重复步骤(2),直至森林中所有的子树构建完成.

(4)由 $i$ 棵子树形成最终随机森林,将待测试样本引入构建好的随机森林,最终结果采用投票选举的方式产生.其最终的决策函数 $P_{rf}(X)$ 由式(6)得出:

$$P_{rf}(X) = \arg \max_Y \sum_{i=1}^k I(w(X, \theta_i) = Y) I(\cdot) Y_c \quad (6)$$

其中, $w(X, \theta_i)$ 为单个分类决策树; $I(\cdot)$ 为指标函数表示满足式子的样本总数; $k$ 为待建子树棵数; $Y$ 为目标变量,解释为是否违约; $\theta_i$ 是随机变量.

随机森林的决策结果取决于每一棵子树的训练结果,分裂指标的选取决定了分裂标准,随机森林一般采用基尼指数(Gini),其大小衡量了各节点混乱程度,其计算如下:

$$\begin{aligned} Gini(c) &= \sum_{y=1}^u \sum_{y' \neq y} p_y(y|c) p_{y'}(y'|c) \\ &= 1 - \sum_{y=1}^u p(y|c)^2 \end{aligned} \quad (7)$$

其中, $p(y|c)$ 为客户类别 $y$ 在子树 $c$ 节点的条件概率;一般来说基尼指数越大,表明在该节点处的数据越趋向均匀分布,样本越纯;当Gini指数为0时,表明该节点所有样本均为一个类.

### 2.2 改进的RF模型流程

基于IV-XGBoostRF的个人信用评估算法模型流程图(如图1),其运行步骤如下:

步骤(1)剔除异常值、重复值、缺失值,对数据中的连续型变量采用最优决策树分箱进行离散化、归一化.

步骤(2)采用皮尔逊相关分析排除变量间的强相



关变量进而排除变量间共线性可能; 同时进行 WOE 编码以及计算分箱后各变量属性的 IV 值, 并通过 IV 值筛选产生初步的特征集; 根据 XGBoost 输出该特征集的重要性完成组合特征选择。

步骤 (3) 利用超参数曲线确定森林大小参数  $n\_estimators$ 、网格搜索确定其他重要参数; 为了减少训练集、测试集划分的随机性, 实验过程采用五折交叉验证, 将数据集  $A$  随机分为 5 份训练样本,  $A_1, A_2, \dots, A_5$ , 每一个样本  $A_i$  都逐次作为训练数据, 其余为测试数据。

步骤 (4) 模型评估。

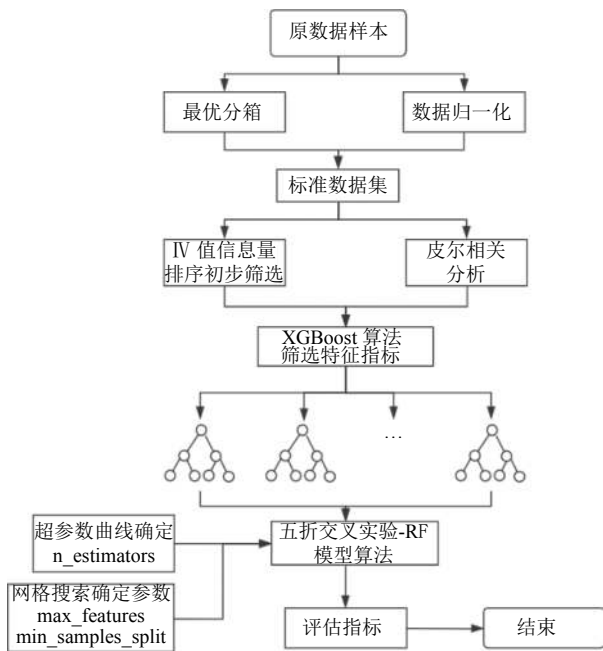


图 1 IV-XGBoost-RF 模型

### 3 实证分析

#### 3.1 实验数据

实验数据集来自 UCI 德国某银行信用卡数据, 利用该数据集验证改进的随机森林模型的可行性, 数据集主要从个人基本信息、账户信息、贷款信息 3 个方面对客户进行描述, 样本大小为 1 000 个样本, 其中正例 700 例, 反例 300, 不平衡率为 2.3, 具体的指标信息如表 1 所示。

#### 3.2 数据预处理与分析

##### (1) 变量分箱

选用最优决策树分箱, 通过比较 Gini 系数大小决定分箱点, 对数据集中 R2 (贷款期限)、R5 (贷款金

额)、R13 (年龄) 3 个连续型变量进行决策树分箱同时计算 IV、WOE 值, 结果如表 2 所示。

表 1 数据集信息说明

数据类型	指标序号	含义
定类变量 (10个)	R3	信用记录 (5类)
	R4	贷款目的 (11类)
	R9	性别及婚姻状况 (5类)
	R10	担保人 (3类)
	R12	资产状况 (4类)
	R14	分期付款情况 (3类)
	R15	住房情况 (3类)
	R17	职业 (4类)
	R19	联系方式 (2类)
	R20	是否外籍 (2类)
定序变量 (3个)	R1	现有支票账户情况 (11类)
	R6	储蓄账户/债券情况 (5类)
	R7	工作年限 (5类)
连续变量 (7个)	R2	贷款期限
	R5	贷款金额
	R8	分期付款率占可支配收入比
	R11	居住时间
	R13	年龄
	R16	现有信贷额度
	R18	需赡养人数
目标变量	target	是否违约 (二分类)

表 2 贷款期限 R2 与贷款金额 R5 及年龄 R13 证据权重

指标序号	分组区间	WOE
R2	[4.0, 8.5)	1.280 9
	[8.0, 15.5)	0.336 214
	[15.5.0, 34.5)	-0.108 688
	[34.5, 43.5)	-0.524 524
R5	[43.5, 72.0)	-1.134 980
	[250.0, 1 373.0)	-0.076 033
	[1 373.0, 1 806.0)	0.772 090
	[1 806.0, 3 913.5)	0.288 083
	[3 913.5, 6 750.0)	-0.336 472
R13	[6 750.0, 18 424.1)	-0.812 812
	[19.0, 25.5)	-0.528 844
	[25.5, 34.5)	-0.060 465
	[34.5, 75)	0.311 213

##### (2) 皮尔逊相关分析

根据计算得到各变量间相关性系数范围为 0.01-0.4, 各指标间关联性均呈弱相关, 因此排除变量间共线性可能。同时结果显示支票账户情况、贷款期限、信用记录特征与是否违约相关度较高。

#### 3.3 特征选择

基于 IV-XGBoost 进行特征筛选, 具体步骤如下:

(1) 根据信息价值计算公式分别计算 20 个指标对应 IV 值 (如表 3), 剔除信息量小于 0.02 的特征, 即 R17、R19、R11、R18, 剩余 16 个待筛选特征。

表 3 各指标 IV 值大小及排名

指标	信息价值IV		指标	信息价值IV	
	值大小	排名		值大小	排名
R1	0.666 0	1	R14	0.057 0	11
R3	0.292 7	2	R20	0.044 8	12
R2	0.284 3	3	R9	0.044 6	13
R6	0.194 7	4	R10	0.032 2	14
R5	0.192 4	5	R8	0.026 4	15
R4	0.168 0	6	R16	0.022 7	16
R12	0.112 5	7	R17	0.008 5	17
R13	0.100 4	8	R19	0.006 3	18
R7	0.086 1	9	R11	0.003 5	19
R15	0.083 5	10	R18	0.003 0	20

(2) 根据图 2 中 XGBoost 对 16 个特征的重要性排名结果, 采用逐步减小变量个数 (如表 4), 以准确度作为衡量标准最终选择 XGBoost 重要性排名前 14 的特征 (加粗为准确度最高)。

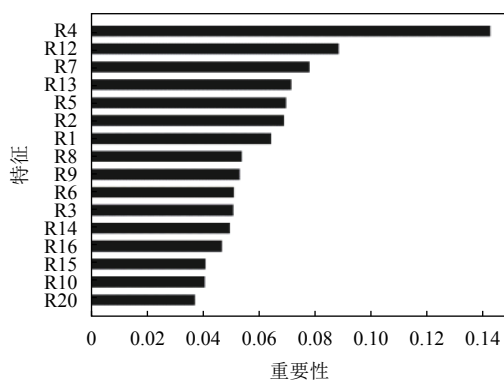


图 2 XGBoost 特征重要性排序

表 4 特征数量选择及模型准确度 (%)

特征数量	准确度
20	75.50
16	75.70
15	76.00
14	<b>76.40</b>
13	75.81
12	75.60
11	74.10

### 3.4 模型性能评估

经过信息值 IV 以及 XGBoost 特征选择后保留了 14 个信用评估特征, 使用超参数学习曲线结合网格搜索确定 RF 模型重要参数  $n\_estimators=104$ 、 $max\_$

$features=4$ 、 $min\_samples\_split=3$ , 该参数组合下的 RF 评估性能最优. 为了进一步验证改进后的随机森林模型的性能, 实验分别与其他五种机器学习模型相比较; 实验过程采用五折交叉验证减小随机性对结果的影响, 据此作以下分析:

(1) 根据表 5 实验结果, 改进后的随机森林的准确度分别高于默认参数下的随机森林、逻辑回归、支持向量机、BP 神经网络模型 0.90%、3.80%、2.70%、1.30%。

表 5 5 种模型实验结果

模型	Accuracy (%)	AUC
RF	75.50	0.968 8
Logistic	72.60	0.760 8
SVM	73.70	0.865 9
BP	75.10	0.864 4
改进后的RF	<b>76.40</b>	<b>0.971 2</b>

(2) ROC 曲线用来衡量模型分类性能优劣的一种图像, ROC 曲线越靠近左上方模型分类效果越好; 其 ROC 曲线下方于 X、Y 轴围成的面积为 AUC 值, 其范围在 0-1 之间, AUC 值越高表明模型性能越优. 图 3 显示改进的随机森林模型 AUC 值高于其他模型, 其 AUC 值为到 0.971, ROC 曲线更靠近左上方。

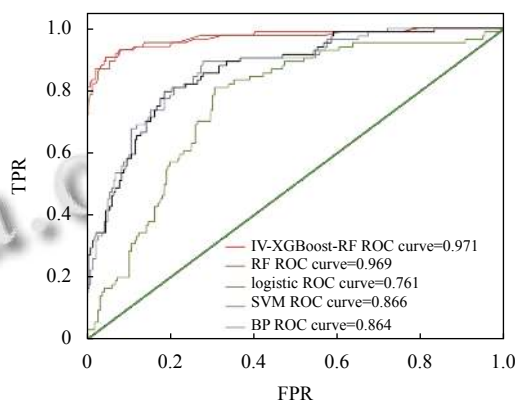


图 3 5 种模型 ROC 曲线

(3) 未改进的随机森林模型  $F$ -score 为 0.874 2, 改进后随机森林  $F$ -score 为 0.895 7, 有明显的提升。

## 4 结论与展望

本文提出了组合特征选择的方法, 首次将传统的风控指标-信息价值 (IV) 和新型集成学习方法 XGBoost 相结合, 以随机森林作为信用评估器. 现针对提出的改进方法作以下几点总结:

(1) 根据组合特征选择结果总结出银行系统应该更关注贷款与账户相关信息, 如支票账户状态、信用记录、贷款期限、贷款目的、储蓄账户情况、固定资产等银行账户等因素; 个人基本信息中更关注就业情况、婚姻状况因素. 符合常理.

(2) 基于 IV-XGBoost 的组合特征选择方法相比直接通过分类算法的特征重要性排序剔除特征更加合理且符合业务逻辑; 同时 IV 计算量小且简单, 当数据维度较大时, 可以有效地减少模型训练时间从而进一步提高算法整体性能.

(3) 与未改进的随机森林模型比较, 改进的随机森林模型 Accuracy 平均值提高 0.90%,  $F$ -score 提高了 2.15%, AUC 提高了 0.20%, 证实本文提出的 IV-XGBoost 组合特征选择方法的有效性和可行性.

不足以及未来展望: 1) 本文的实验还有一些不足, 由于条件有限实验数据集较小, 仅能证实特征选择方法和集成模型有效性和可行性, 在大数据集上该组合特征选择的效率高低并未得到证实; 2) 面对大数据集的个人信用数据, 能否将大数据技术与集成学习算法相结合也是未来研究的主要方向.

### 参考文献

- 1 Hoque N, Bhattacharyya DK, Kalita JK. MIFS-ND: A mutual information-based feature selection method. *Expert Systems with Applications*, 2014, 41(14): 6371–6385. [doi: 10.1016/j.eswa.2014.04.019]
- 2 Fernandes GB, Artes R. Spatial dependence in credit risk and its improvement in credit scoring. *European Journal of Operational Research*, 2016, 249(2): 517–524. [doi: 10.1016/j.ejor.2015.07.013]
- 3 郭畅. 基于类别不平衡的企业信用风险违约测度探索——以制造业上市公司为例. *重庆工商大学学报(自然科学版)*, 2021, 38(1): 113–119. [doi: 10.16055/j.issn.1672-058X.2021.0001.017]
- 4 周毓萍, 陈官羽. 基于机器学习方法的个人信用评价研究. *金融理论与实践*, 2019, (12): 1–8. [doi: 10.3969/j.issn.1003-4625.2019.12.001]
- 5 Harris T. Credit scoring using the clustered support vector machine. *Expert Systems with Applications*, 2015, 42(2): 741–750. [doi: 10.1016/j.eswa.2014.08.029]
- 6 Abdou H, Pointon J, El-Masry A. Neural nets versus conventional techniques in credit scoring in Egyptian banking. *Expert Systems with Applications*, 2008, 35(3): 1275–1292. [doi: 10.1016/j.eswa.2007.08.030]
- 7 陈云, 石松, 潘彦, 等. 基于 SVM 混合集成的信用风险评估模型. *计算机工程与应用*, 2016, 52(4): 115–120. [doi: 10.3778/j.issn.1002-8331.1403-0473]
- 8 Twala B. Combining classifiers for credit risk prediction. *Journal of Systems Science and Systems Engineering*, 2009, 18(3): 292–311. [doi: 10.1007/s11518-009-5109-y]
- 9 Zhu Y, Xie C, Wang GJ, et al. Comparison of individual, ensemble and integrated ensemble machine learning methods to predict China's SME credit risk in supply chain finance. *Neural Computing and Applications*, 2017, 28(1): 41–50. [doi: 10.1007/s00521-016-2304-x]
- 10 萧超武, 蔡文学, 黄晓宇, 等. 基于随机森林的个人信用评估模型研究及实证分析. *管理现代化*, 2014, 34(6): 111–113. [doi: 10.3969/j.issn.1003-1154.2014.06.038]
- 11 周永圣, 崔佳丽, 周琳云, 等. 基于改进的随机森林模型的个人信用风险评估研究. *征信*, 2020, 38(1): 28–32. [doi: 10.3969/j.issn.1674-747X.2020.01.006]
- 12 李欣, 俞卫琴. 基于改进 GS-XGBoost 的个人信用评估. *计算机系统应用*, 2020, 29(11): 145–150. [doi: 10.15888/j.cnki.csa.007624]
- 13 Chen TQ, Guestrin C. XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM, 2016. 785–794. [doi: 10.1145/2939672.2939785]
- 14 Breiman L. Random forests. *Machine Learning*, 2001, 45(1): 5–32. [doi: 10.1023/A:1010933404324]