计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(11):145−150 [doi: 10.15888/j.cnki.csa.007624]

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

基于改进 GS-XGBoost 的个人信用评估①

李　欣, 俞卫琴

(上海工程技术大学 数理与统计学院, 上海 201620)
通讯作者: 李　欣, E-mail: lixinsues@163.com

摘　要: 信用评估分类器的好坏能够直接影响信贷金融机构的盈利能力. 传统的网格搜索法进行参数寻优时会耗费大量的时间, 基于此提出改进的网格搜索法优化 XGBoost (GS-XGBoost) 的个人信用评估算法. 该算法利用随机森林进行特征选择后, 将改进的网格搜索法对 XGBoost 中的 n_estimators 和 learning_rate 进行参数寻优, 建立评估模型. 从 UCI 数据库中选取信贷数据进行分析, 分别与支持向量机、随机森林、逻辑回归、神经网络以及未改进的 XGBoost 进行比较. 实验结果表明, 该模型的 *F-score* 和 *G-mean* 的值均有提高.

关键词: 网格搜索; 信用评估; GS-XGBoost; 参数寻优

引用格式: 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. http://www.c-s-a.org.cn/1003-3254/7624.html

Personal Credit Evaluation Based on Improved GS-XGBoost

LI Xin, YU Wei-Qin

(School of Mathematics, Physics and Statistics, Shanghai University of Engineering Science, Shanghai 201620, China)

Abstract: The quality of the credit evaluation classifier can directly affect the profitability of credit financial institutions. The traditional grid search takes a lot of time for parameter optimization. Based on this, we propose an improved grid search to optimize the XGBoost (GS-XGBoost) personal credit evaluation algorithm. After using the feature selection based on random forest, the algorithm uses the improved grid search method to optimize the parameters of n_estimators and learning_rate in XGBoost to establish an evaluation model. We analyze the credit data selected from the UCI database to compare with support vector machines, random forests, logistic regression, neural networks, and unimproved XGBoost. Experimental results show that the *F-score* and *G-mean* values of the model are improved.

Key words: grid search; credit evaluation; GS-XGBoost; parameter optimization

在信用贷款不断发展的今天, 信用评分已成为金融机构日益关注的问题, 目前已成为研究的热门问题. 信用评分是金融业的重要组成部分, 在信用客户选择、风险计量、贷款前后监管、综合绩效评估和资产组合风险管理等现代事务中发挥着重要作用[1]. 在银行、金融机构以及基于互联网的金融公司中, 强大的信用风险预测能力可以更好地巩固市场上的可持续利润. 信用评分的目的是将申请人分为两类: 信誉良好的人和信誉不良的人[2]. 信誉良好的人很有可能还清财务义务. 信誉不良的人极有可能发生违约. 信用评分的准确性对金融机构的盈利能力至关重要. 即使将信用不良的申请人的信用评分准确性提高 1%, 也将减少金融

① 基金项目: 国家自然科学基金 (11602134, 11772148); 全国统计科学研究项目一般项目 (2018LY16)
Foundation item: National Natural Science Foundation of China (11602134, 11772148); General Project of National Statistical Science Research Program (2018LY16)
收稿时间: 2020-03-04; 修改时间: 2020-03-27; 采用时间: 2020-04-14; csa 在线出版时间: 2020-10-29

机构的巨大损失.

针对信用评估的方法主要有逻辑回归[3]、支持向量机[4]、神经网络[5] 和决策树[6] 等. 传统的对信用评估的模型主要采用单一模型. 如王黎等[7] 直接采用 Gradient Boosted Decision Tree (GBDT) 的方法对个人信用进行评估. 罗方科等[8] 运用逻辑回归模型对小额贷款风险进行评估. 然而, 单一模型在处理非线性问题时效果并不十分理想.

为了解决单一模型的问题, 对模型进行组合应用逐渐成为提高信用评估准确率以及稳定性主要方法. Wang 等[9] 将逻辑回归分析、决策树、人工神经网络以及支持向量机多个分类器 (即集成学习) 结合使用, 显著提高单个基础学习者的学习能力. Koutanaei 等[10] 提出特征选择算法和集成学习分类器的混合数据挖掘模型引用于信用评估, 将 4 种特征选择算法进行比较得出 PCA 算法较好. He 等[11] 为信用评分生成一个新颖的集成模型, 使用粒子群优化算法进行基本分类器的参数优化, 减少了数据不平衡带来的负面影响, 提高了信用评分方面预测模型的综合性能. 刘潇雅等[12] 应用 C4.5 信息熵增益率方法进行特征选择, 减少了数据的冗余属性. 王名豪等[13] 对混沌粒子群法进行改进, 并应用于 XGBoost 算法中进行参数优化, 提高了信用评估的准确性.

基于上述研究进展, 本文提出基于改进的 GS-XGBoost 的个人信用评估研究, 用改进的网格搜索法寻找分类器的最优参数. 实验部分, 在 UCI 机器学习数据库中的信贷数据集上比较了本文提出 GS-XGBoost 与其他常用算法的性能. 实验结果表明, 本文算法具有较高的预测准确率, 是进行信用风险评估的有效模型.

1 理论与方法

1.1 XGBoost 模型

XGBoost (eXtreme Gradient Boosting) 是极限梯度提升算法, 由 Chen 等[14] 设计, 主要使提升树突破自身的计算极限, 来实现运算快速, 性能优秀的工程目标.

XGBoost 的目标函数为:

$$\begin{cases} L = \sum\limits_i l(\hat{y}_i, y_i) + \sum\limits_k \Omega(f_k) \\ \Omega(f) = \gamma T + \dfrac{1}{2}\lambda \|w\|^2 \end{cases} \tag{1}$$

其中, l是损失函数, 用于测量预测值$\hat{y}_i$与真实值y_i之间的差, 第二项Ω是惩罚函数, 即惩罚了模型的复杂性. 在惩罚函数中, γ是复杂度参数, T为叶子节点数, λ是叶子权重w的惩罚系数[15]. 惩罚函数Ω有助于平滑最终学习的权重, 以避免过度拟合.

在 XGBoost 中, 完整的迭代决策树的公式应该写作:

$$\hat{y}_i^{(k+1)} = \hat{y}_i^{(k)} + \eta f_{k+1}(X_i) \tag{2}$$

其中, f_{k+1}为第$k+1$棵树的模型, η是迭代决策树时的步长 (shrinkage), 又称为学习率 (learning rate). η越大, 迭代的速度越快, 算法的极限很快被达到, 有可能无法收敛到真正的最佳. η越小, 越有可能找到更精确的最佳值, 更多的空间被留给了后面建立的树, 但迭代速度会比较缓慢.

式 (1) 中的树集成模型 L 将函数$\hat{y}_i$作为参数, 且无法使用欧几里得空间中的传统优化方法进行优化, 取而代之以附加方式进行优化训练. 通常情况下, 令$\hat{y}_i^{(t)}$表示第 t 次迭代中第 i 个实例的预测值, 我们将需要添加f_t来最小化以下目标:

$$\begin{aligned} L^{(t)} &= \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(X_i)) + \Omega(f_t) \\ &\approx \sum_{i=1}^{n} \left[g_i f_t(X_i) + \frac{1}{2} h_i f_t^2(X_i) \right] + \Omega(f_t) \end{aligned} \tag{3}$$

其中, g_i和h_i分别为损失函数的一阶和二阶梯度统计量.

定义$I_j = \{i|q(X_i) = j\}$为叶子j的实例集, 则可以将等式 (3) 写成如式 (4).

$$\begin{aligned} L^{(t)} &= \sum_{i=1}^{n} \left[g_i f_t(X_i) + \frac{1}{2} h_i f_t^2(X_i) \right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \\ &= \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i \right) w_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) w_j^2 \right] + \gamma T \end{aligned} \tag{4}$$

对于固定结构$q(X)$, 我们可以计算出叶子 j 的最优权重w_j^*以及目标函数的最优值为:

$$w_j^* = -\frac{\sum\limits_{i \in I_j} g_i}{\sum\limits_{i \in I_j} h_i + \lambda} \tag{5}$$

$$L^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^{T} \frac{\left(\sum\limits_{i \in I_j} g_i \right)^2}{\sum\limits_{i \in I_j} h_i + \lambda} + \gamma T \tag{6}$$

式 (6) 用来对结构树 q 的质量进行评分. 该分数类似于评估决策树的杂质系数, 不同之处在于它是针对更广泛的目标函数而得出的. XGBoost 本身的核心是基于梯度提升树实现的集成算法, 整体来说可以有 3 个核心部分: 集成算法本身, 用于集成的弱评估器, 以及应用中的其他过程.

梯度提升算法是 XGBoost 算法的基础, 它是实现模型预测的有力技术之一, 在 Boosting 算法中处于重要位置. 集成算法主要通过在数据集中生成弱评估器 (n_estimators), 并将弱评估器集合起来, 效果优于单一的模型. n_estimators 过小容易造成数据的欠拟合, 过多容易造成数据的过拟合问题. 所以如何选择合适的 n_estimators 是一个重点.

1.2 改进的网格搜索法

网格搜索法 (grid search)[16] 是指将指定参数进行枚举, 通过将评估函数中的参数进行交叉验证得到最优参数的算法. 即把指定优化的参数在一定范围内依次排序, 并将这些数据排列成组合形成网格, 依次将数据放入分类器中进行训练, 并采用交叉验证方法对参数的表现进行评估, 在分类器遍历了所有的参数组合后, 返回一个最优的分类器, 同时获得最优的参数组合.

本文中对网格搜索法的实际应用是让 eta 与 n_estimators 在一定的范围内划分网格并遍历网格内所有点进行取值 (数据为本文借贷数据), 其中 eta 的范围为 [0.05, 1],步长为 0.01, n_estimators 的范围为 [1, 300], 步长为 3. 在此范围内得到 eta 与 n_estimators 下训练集分类准确率, 通过比较准确率来确定最优的参数组合. 参数的寻优如图 1 所示.

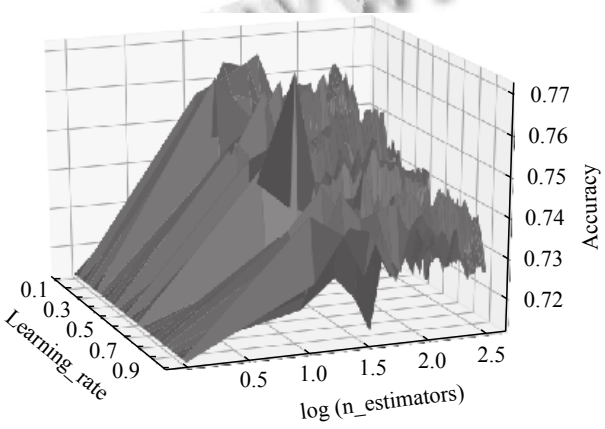

图 1 网格搜索法

从图 1 中能够知道, 参数 eta 和 n_estimators 在一定的区间范围内能够取得比较高的准确率, 但在其他的多数范围内的准确率并不高, 使得在进行参数寻优的过程中消耗大量的时间.

针对上述问题提出改进方法. 首先, 在参数区间上选择大步长进行参数寻优, 得到准确率高的最优局部参数. 再次, 在局部最优参数范围内采用小步长在该范围内进行二次寻优, 寻找最优参数. 改进的方法减少了不必要的计算, 节省了大量的时间.

2 改进的 GS-XGBoost 的个人信用评估模型

改进的个人信用评估模型分为 2 部分, 第 1 部分为数据预处理过程, 首先将数据集进行极差标准化处理后进行特征选择, 筛选出重要性高的特征属性. 第 2 部分为模型的优化过程, 将筛选出的特征数据集进行改进的网格搜索法处理, 寻找最优参数 n_estimators 和 learning rate. 随后将模型进行评估, 采用 5 折交叉验证法并取均值进行对比.

2.1 算法流程

改进的 GS-XGBoost 的个人信用评估模型的流程图如图 2 所示.

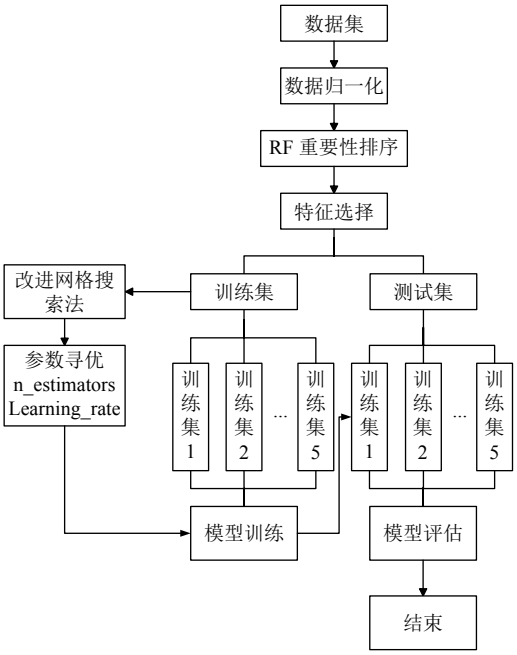

图 2 算法流程图

具体步骤如下:

步骤 1. 数据预处理. 对数据进行建模分析之前, 需

要对数据中的缺失值进行填补或删除. 之后, 对处理后的数据进行极差标准化处理, 公式如下:

$$x^* = \frac{x_{ij} - \min(x_j)}{\max(x_j) - \min(x_j)} \tag{7}$$

其中, x_{ij}为样本点, $\min(x_j)$与$\max(x_j)$是第j属性下样本数据的最小与最大值.

步骤 2. 特征选择. 数据集中的维数过高时, 不相关的属性特征对个人信用的评估之间并没有相互关联性, 影响个人评估的准确率. 通过特征选择可以提高模型的精确度, 预防过拟合. 本文采用随机森林算法 (RF) 对数据集进行训练, 取得每个特征的重要性排名并移除重要度低的特征属性.

步骤 3. 参数寻优. 使用改进的网格搜索法对分类器的参数进行寻优.

1) 将数据集 D 分为训练集D_{train}与测试集D_{test}, 比率为 7/3. 在数据集中采用五折交叉验证法 (5-fold Cross Validation) 将训练集分为 5 份$D_{\text{train1}}, D_{\text{train2}}, \cdots, D_{\text{train5}}$.

2) 将产生的 5 组数据对分类器中的弱评估器 (n_estimators) 以及学习速率 (learning_rate) 进行训练, 使用改进的网格搜索法对 XGBoost 模型进行寻优, 得到最优参数.

步骤 4. 模型评估. 将最优参数与特征子集代入模型中进行评估, 并与其它分类器进行比较.

2.2 评价指标

本文选择$F-value$与$G-mean$值来对信用评估进行评价. 混淆矩阵如表 1 所示.

表 1 混淆矩阵

类别	预测为多数类	预测为少数类
实际为多数类	T_p	F_n
实际为少数类	F_p	T_n

$$\begin{cases} R_{\text{call}} = \dfrac{T_p}{T_p + F_n} \\ P_{\text{recision}} = \dfrac{T_p}{T_p + F_n} \\ F-value = \dfrac{(1+\beta^2) \times R_{\text{call}} \times P_{\text{recision}}}{\beta^2 \times R_{\text{call}} + P_{\text{recision}}} \end{cases} \tag{8}$$

$$\begin{cases} N_{\text{accrance}} = \dfrac{T_n}{T_n + F_p} \\ G-mean = \sqrt{R_{\text{call}} \times N_{\text{accrance}}} \end{cases} \tag{9}$$

3 实证分析

3.1 数据预处理

为了检验本文改进算法的有效性, 对本文算法进行实证分析, 从 UCI 国际机器学习库中挑选出信用卡借贷数据. 数据的相关信息如表 2 所示. 属性相关信息如表 3 所示.

表 2 样本信息

数据名称	指标属性	样本数量	不平衡率(正/负)
German	20	1000	2.33

表 3 特征属性

属性	内容	属性	内容
A1	现有支票账户状态	A12	固定资产
A3	信用记录	A14	商品分期付款情况
A4	贷款使用目的	A15	住房情况
A6	储蓄账户/债券情况	A17	工作职务
A7	工作持续年限	A19	电话
A9	性别及婚姻状态	A20	是否外籍工人
A10	担保人状态	其他	持续性数据

对数据集进行特征选择, 利用随机森林对数据集进行特征重要性排名, 结果如图 3 所示. 选取排名前 12 的特征属性数据集.

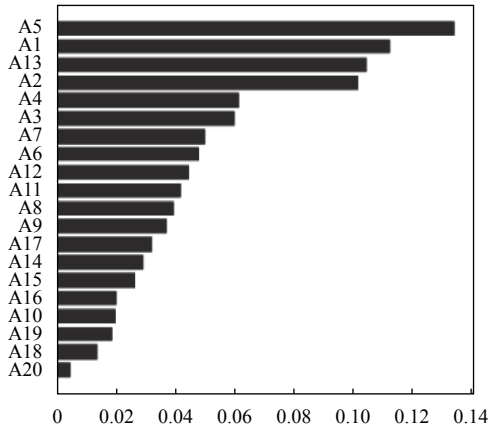

图 3 特征重要性排名

3.2 改进网格搜索法

为了比较改进算法的优越性, 将特征选择后的数据集进行参数寻优. 网格搜索法的变量为 n_estimators 和 learning rate, 设定不同的步长范围, 分为 4 组数据进行比较. 第 1 组数据 n_estimators 的范围为 (1, 300), 步长为 5, learning rate 的范围为 (0.05, 1), 步长为 0.05. 第 2 组数据 n_estimators 的范围为 (1, 300), 步长为 10, learning rate 的范围为 (0.1, 1), 步长为 0.1. 第 3 组数据 n_estimators 的范围为 (1, 300), 步长为 20, learning

rate 的范围为 (0.1, 1), 步长为 0.1. 第 4 组数据 n_estimators 的范围为 (1, 300), 步长为 50, learning rate 的范围为 (0.1, 1), 步长为 0.1, 结果如表 4 所示. 选择最优参数 n_estimators 为 16, learning rate 为 0.44.

3.3 模型评估

本文将改进后的模型进行评估. 经过特征选择后的数据集从 20 维下降为 12 维, 将特征选择后筛选的数据集作为分类器 XGBoost 的训练集进行训练, 使用改进的网格搜索法寻找 XGBoost 的最优参数 n_estimators 和 learning rate. 本文使用软件为 Python3.7, 使用 5 折交叉验证法对数据集进行训练来减少随机性对分类结果的影响. 本文算法模型 (GS-XGB) 与支持向量机 (SVM)、随机森林 (RF)、逻辑回归 (LOG)、神经网络 (BP) 以及未改进的 XGBoost (XGB) 进行比较, 实验结果如表 5 和表 6 所示, F 和 G 分别为$F-value$和$G-mean$的值, 其中加粗部分为相同条件下的模型的最大数值.

表 4 参数寻优结果比较

组别	n_estimators步长	learning rate步长	准确率	时间(s)
第1组	5	0.01	0.776	17337
第2组	10	0.01	0.771	2552
第3组	20	0.01	0.77	1036
第4组	50	0.01	0.768	377

表 5 少数类准确率

模型	SVM	RF	LOG	BP	XGB	GS-XGB
1-k	0.823	0.811	0.816	0.867	0.846	**0.886**
2-k	**0.879**	0.805	0.866	0.852	0.839	0.866
3-k	0.854	0.878	0.862	0.939	0.947	**0.954**
4-k	0.899	0.804	0.85	0.862	0.862	**0.891**
5-k	0.862	0.856	0.871	0.849	0.863	**0.895**
mean	0.863	0.83	0.853	0.873	0.871	**0.898**

表 6 模型实验结果对比

	模型	SVM	RF	LOG	BP	XGB	GS-XGB
F	1-k	0.854	0.811	0.848	0.835	0.826	**0.937**
	2-k	0.853	0.828	0.846	0.844	0.872	**0.889**
	3-k	0.822	0.816	0.832	0.828	0.821	**0.873**
	4-k	0.818	0.79	0.816	0.812	0.842	**0.924**
	5-k	0.844	0.81	0.843	0.84	0.83	**0.896**
	mean	0.838	0.811	0.837	0.831	0.8382	**0.897**
G	1-k	0.61	0.653	0.595	0.641	0.645	**0.689**
	2-k	0.643	0.688	0.638	0.659	**0.748**	0.666
	3-k	0.539	0.648	0.566	0.595	0.549	**0.588**
	4-k	0.552	0.624	0.604	0.601	0.668	**0.676**
	5-k	0.662	0.592	0.697	**0.718**	0.662	0.673
	mean	0.601	0.641	0.62	0.642	0.654	**0.659**

表 5 为各模型下少数类的准确率, 少数类为信用较差的用户. 从表中可以得到如下结论: 1) 总体上每个模型下的少数类准确率都比较高, 差异较小, 但与其他模型相比, 该方法对信用评估的分类效果优于其他算法, 能够有较大的准确率识别信用不良人员. 2) 与随机森林模型 (RF) 相比, 少数类分类的平均准确率提高了 6.8%, 与未经过改进的 XGBoost 相比, 平均准确率提高了 2.7%.

表 6 为各个评估模型在信贷数据集上的$F-value$和$G-mean$值, 从实验结果可以得到如下结论: 1) 改进的网格搜索法优化 XGBoost 算法的评估效果优于其他算法. 2) 相比没有进行改进的 XGBoost 算法, 改进的 XGBoost 算法$F-value$平均值提高了 5.88%, $G-mean$平均值提高了 0.5%. 这是因为在对数据进行特征选择后摒弃了无关的数据特征.

4 结语

随着个人信用贷款消费愈来愈普及, 个人信用良好与否直接导致信贷金融机构的损失, 因此对个人信用评估的研究非常重要. 本文提出了基于改进的 GS-XGBoost 的个人信用评估研究, 该方法将改进的网格搜索法应用于 XGBoost 模型, 筛选出最优参数 n_estimators 和 learning rate. 选用 UCI 公开数据集进行评估, 使用$F-value$和$G-mean$以及少数类准确率作为评估指标. 实验结果表示, 该算法对个人信用借贷的评估性能优于其他算法.

未来需要解决的问题有: 1) 本文属于二分类问题, 对于多分类还需要进一步的研究. 2) 该算法对本文数据集的有效性是否对其他数据也有效. 3) 在特征选择上, 如何将该算法与其他方法相结合 (如神经网络, 支持向量机等) 进一步提高算法精确度.

参考文献

1 Kozeny V. Genetic algorithms for credit scoring: Alternative fitness function performance comparison. Expert Systems with Applications, 2015, 42(6): 2998–3004. [doi: 10.1016/j.eswa.2014.11.028]

2 Ala'Raj M, Abbod M. Classifiers consensus system approach for credit scoring. Knowledge-Based Systems, 2016, 104: 89–105. [doi: 10.1016/j.knosys.2016.04.013]

3 Feng JZ, Wang Y, Peng J, *et al*. Comparison between logistic regression and machine learning algorithms on survival prediction of traumatic brain injuries. Journal of critical care, 2019, 54: 110–116. [doi: 10.1016/j.jcrc.2019.08.010]

4 谢娟英, 谢维信. 基于特征子集区分度与支持向量机的特征选择算法. 计算机学报, 2014, 37(08): 1704–1718.
5 黄卿, 谢合亮. 机器学习方法在股指期货预测中的应用研究——基于 BP 神经网络、SVM 和 XGBoost 的比较分析. 数学的实践与认识, 2018, 48(08): 297–307.
6 唐耀先, 余青松. 消除属性间依赖的 C4.5 决策树改进算法. 计算机应用与软件, 2018, 35(03): 262–265, 315. [doi: 10.3969/j.issn.1000-386x.2018.03.050]
7 王黎, 廖闻剑. 基于 GBDT 的个人信用评估方法. 电子设计工程, 2017, 25(15): 68–72. [doi: 10.3969/j.issn.1674-6236.2017.15.018]
8 罗方科, 陈晓红. 基于 Logistic 回归模型的个人小额贷款信用风险评估及应用. 财经理论与践, 2017, 38(01): 30–35.
9 Wang G, Hao J, Ma J, *et al*. A comparative assessment of ensemble learning for credit scoring. Expert Systems with Applications, 2011, 38(1): 223–230. [doi: 10.1016/j.eswa.2010.06.048]
10 Koutanaei FN, Sajedi H, Khanbabaei M. A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring. Journal of Retailing and Consumer Services, 2015, 27: 11–23. [doi: 10.1016/j.jretconser.2015.07.003]
11 He HL, Zhang W, Zhang S. A novel ensemble method for credit scoring: Adaption of different imbalance ratios. Expert Systems with Applications, 2018, 98: 105–117. [doi: 10.1016/j.eswa.2018.01.012]
12 刘潇雅, 王应明. 基于 C4.5 算法优化 SVM 的个人信用评估模型. 计算机系统应用, 2019, 28(7): 133–138. [doi: 10.15888/j.cnki.csa.006958]
13 王名豪, 梁雪春. 基于 CPSO-XGboost 的个人信用评估. 计算机工程与设计, 2019, 40(07): 1891–1895.
14 Chen TQ, Guestrin C. XGBoost: a scalable tree boosting system. The 22nd ACM SIGKDD International Conference. San Francisco, CA, USA. 2016. 785–794. [doi: 10.1145/2939672.2939785]
15 Xia Y, Liu C, Li YY, *et al*. A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring. Expert Systems with Applications, 2017, 78: 225–241. [doi: 10.1016/j.eswa.2017.02.017]
16 王健峰, 张磊, 陈国兴, 等. 基于改进的网格搜索法的 SVM 参数优化. 应用科技, 2012, 39(03): 28–31.