计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(7):133−138 [doi: 10.15888/j.cnki.csa.006958]

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

基于 C4.5 算法优化 SVM 的个人信用评估模型①

刘潇雅, 王应明

(福州大学 经济与管理学院, 福州 350108)
通讯作者: 刘潇雅, E-mail: 549964064@qq.com

摘　要: 支持向量机作为非参数方法已经广泛应用于信用评估领域. 为克服其训练高维数据不能主动进行特征选择导致准确率下降的缺点, 构建 C4.5 决策树优化支持向量机的信用评估模型. 利用 C4.5 信息熵增益率方法进行属性选择, 减少冗余属性. 模型通过网格搜索确定最优参数, 使用 *F-score* 和平均准确率评价模型性能, 并在两组公开数据集上进行验证. 实证分析表明, C4.5 决策树优化支持向量机的信用评估模型有效减少了数据学习量, 较于传统各类单一模型有较高的分类准确率和实用性.

关键词: 个人信用评估; 支持向量机; C4.5 决策树; 属性选择; 信息熵增益率

引用格式: 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. http://www.c-s-a.org.cn/1003-3254/6958.html

Evaluation Model for Personal Credit Risk Based on C4.5 Algorithm for Optimizing SVM

LIU Xiao-Ya, WANG Ying-Ming

(School of Economics and Management, Fuzhou University, Fuzhou 350108, China)

Abstract: Support Vector Machine (SVM) has been widely used in the field of credit evaluation as non-parametric method. However, it cannot actively select attributes when processing high-dimensional data which may cause a drop in accuracy. In order to overcome this shortcoming, credit evaluation model of C4.5 decision tree optimized SVM is constructed to select attributes, and reduce redundant attributes. The model determines the optimal parameters through grid search, uses *F-score* and average accuracy to evaluate model performance on two sets of public data sets. Empirical analysis shows that the proposed model effectively reduces data learning process, and has higher classification accuracy and practicability than the various traditional types of single models.

Key words: personal credit evaluation; support vector machine; C4.5 decision tree; attribute selection; information entropy gain rate

在无抵押纯信用小额个人贷款越来越有热度的当今社会, 银行等金融行业越来越重视个人消费型贷款业务, 信用评估也成为了大家关注的焦点和金融机构评估信贷风险、增加现金流量、降低违约率的主要方法[1]. 个人信用评估的原理是根据贷款申请人的收入, 支出, 工作性质等基本信息和过去的表现等特征建立信用评估模型, 并用该模型对具有相同特征的未来申请者的信用进行预测[2], 区分出来“好”的贷款和“坏”的贷款, 从而协助银行等金融机构做出是否放贷的决策. 因此, 如何能够在现有信用环境下选取科学、高效的信用评估方法, 从而对贷款申请人做出有效的信用评估, 显得尤为重要[3].

传统用于信用评估的主要是统计学方法比如 logistic 回归, 判别分析等, 这些方法虽简单, 但处理非

① 基金项目: 国家自然科学基金 (61773123)
Foundation item: National Natural Science Foundation of China (61773123)
收稿时间: 2018-12-29; 修改时间: 2019-01-18; 采用时间: 2019-01-22; csa 在线出版时间: 2019-07-01

线性问题时效果较差. 随着计算机的进步, 人工智能等方法已经被用来进行信用评估, 比如神经网络(ANN)[4]、支持向量机 (SVM)[5]、决策树 (DT)[6]等. 人工智能的方法可以有效解决非线性问题, 但存在一定缺陷. 例如神经网络基于经验风险最小化原则常常会出现“过拟合”现象, 泛化能力比较差. 此时基于结构风险最小化原则的支持向量机由于很强的泛化推广能力, 且在解决小样本、非线性识别问题中表现出许多特有的优势, 为信用评估提供了更佳的选择[7]. 吴冲[8]等利用基于模糊积分的支持向量机集成方法对客户信用进行评估, 结果表明支持向量机具有较高的预测准确率. 肖智[9]等利用支持向量机建立了大学生助学贷款个人信用评价分析模型, 通过实证体现了支持向量机方法的优越性. 然而现有大多数支持向量机作为基分类器信用评估时, 面向高维或者大规模样本, 存在不能主动进行特征选择和组合的问题, 因此准确率会受到无关维度的影响, 甚至产生维度灾难.

为解决单一模型的缺陷, 取长补短, 模型的组合应用已经成为提高信用评估准确率和稳定性的一大趋势. 文献[10]将主分量分析和神经网络 (PCA—NN) 模型组合进行个人信用评估取得了更好的预测分类能力. 文献[11]综合比较了多个组合模型在信用评估应用中的效果, 得出了组合模型比单一模型性能更好的结论.

综上所述, 考虑到决策树算法本身以属性的差异性为依据进行分支和最优树的生成, 为优化支持向量机会受冗余属性影响导致准确率下降的缺陷, 本文将基于信息熵增益率分类原理的 C4.5 最优决策树和 SVM 模型优化整合, 提出基于 C4.5 算法优化 SVM 的个人信用评估模型. 实验部分, 为检测模型效果, 在两个公开数据集上比较了本文提出的模型与常见单一模型的性能, 并用 *F-score* 和平均准确率两个指标对模型效果进行评估. 实验结果表明, 基于 C4.5 算法优化 SVM 的个人信用评估模型可以取得更好的性能, 能够成为一种有效的信用评估模型.

1 理论简介

1.1 支持向量机

支持向量机 (Support Vector Machine, SVM) 是在 Vapnik 等人所建立的统计学习理论 (Statistical Learning Theory, STL) 基础上发展起来的一种新的学习算法, 基于 VC 维理论和结构风险最小化原理, 根据有限的样本信息在模型的复杂性和学习能力之间寻求最佳折衷[12].

设训练样本集 D 为(x_i,y_i),能使分类间隔 $(2/\|\omega\|^2)$ 最大的超平面为最优超平面. 在分类中, 支持向量机尝试找到一个使得期望分类误差最小化的分类器$f(x)$,找这个分类器的过程等同为求解下例凸二次规划化问题:

$$\begin{cases} \min \dfrac{1}{2}\left\|\omega^2\right\|; \\ \text{s.t. } y_i[\omega^{\mathrm{T}}x+b]\geq 1 \quad (i=1,2,\cdots,n) \end{cases} \tag{1}$$

上述二次规划可以用对偶理论求解, 最终线性可分情况下的决策函数为:

$$f(x)=\operatorname{sgn}\left\{\sum_{i=1}^{n}\alpha_i^{*}y_i(x_i^{\mathrm{T}}x)+b^{*}\right\} \tag{2}$$

对于线性不可分的问题, 通过核函数将向量映射到一个更高的特征空间, 在高维空间输入的向量可以被超平面成功分开. 通过核函数可以简化内积的运算, 常用核函数有高斯核函数、线性核函数和多项式核函数, 引入松弛变量ξ_i和惩罚函数 C, 线性不可分情况下凸二次规划问题变为:

$$\begin{cases} \min \dfrac{1}{2}\|\omega\|^2+C\left(\sum\limits_{i=1}^{n}\xi_i\right) \\ \text{s.t. } y_i[\omega^{T}x+b]\geq 1(i=1,2,\cdots,n) \end{cases} \tag{3}$$

根据对偶理论求解可得决策函数为:

$$f(x)=\operatorname{sgn}\left\{\sum_{i=1}^{n}\alpha_i^{*}y_iK(x_i,x)+b^{*}\right\} \tag{4}$$

本文采用高斯核函数:

$$K(x_i,x_j)=\exp\left(\gamma\left\|x_i-x_j\right\|^2\right)\gamma=-\frac{1}{2\sigma^2} \tag{5}$$

1.2 C4.5 决策树

决策树学习是应用统计, 机器学习和数据挖掘领域中一类重要的监督学习算法[13,14]. 采用自顶向下的递归方式, 从树根节点开始在内部进行属性的测试比较, 根据属性值确定分支, 最后在决策树的叶子节点得到分类的结论, 整个过程在以新的节点为根的子树上重复, 直到训练停止得到最优决策树. 影响最大的决策树算法是 ID3 算法, 它以信息增益来选择属性. 为克服 ID3 算法信息增益选择属性时偏向于选择取值多的属性和其只能处理离散型和完整性属性等缺点, 1993 年提出了 C4.5 决策树算法, 以信息熵增益率方法测试属性[15].

信息增益率计算公式为:

$$GainRatio(D,A)=\frac{Gain(D,A)}{Split\_\inf o(D,A)} \tag{6}$$

其中, D 为数据集, A 是数据集属性, $Gain(D,A)$为属性 A 的信息增益, $Split_\inf o(D,A)$为属性 A 的分裂信息量.

2 基于 C4.5 算法优化 SVM 的个人信用评估模型

信用评估领域中, 银行等金融机构为了从众多信用数据中归纳出信用“好”和信用“差”的顾客的一般规律从而降低误判率, 会收集和积累大量的数据, 但是对于数据集本身而言, 并不是所有的样本属性均包含相同的对结果有影响的信息量. 冗余属性较多反而会出现“维数灾难”, 增加了模型计算的复杂度降低模型效率. 基于此, 本文提出 C4.5 决策树利用自身属性筛选方法优化支持向量机无法主动降维缺陷的个人信用评估模型.

C4.5 算法优化 SVM 的个人信用评估模型包含了两个子系统: 一个是基于 C4.5 决策树的属性筛选和 SVM 参数优化系统; 一个是训练和测试 SVM 分类器性能系统.

2.1 SVM 参数优化

本文采用高斯核函数作为 SVM 的核函数解决线性不可分的问题. 惩罚参数 C 的作用是为了权衡经验风险和结构风险, C 值越大, 模型对离群点越重视, 模型越复杂容易出现过拟合; C 值越小, 模型对离群点越不重视, 容易出现欠拟合现象. 而高斯核函数的参数 γ的改变实际上是隐含的改变了样本空间的复杂程度, 若太大会将样本类别分的太细, 太小会将样本类别分的太粗. 因此, 参数 C 和核函数的参数γ是影响支持向量机分类器性能的至关重要的因素. C4.5 算法优化 SVM 的个人信用评估模型采用网格搜索与交叉验证的方法挑选 SVM 参数, 确保 SVM 作为基分类器达到较佳状态.

2.2 C4.5 算法优化 SVM 的个人信用评估模型

C4.5 算法优化 SVM 的个人信用评估模型流程图如图 1 所示, 具体步骤如下:

步骤 1: 数据预处理. 所有的连续变量都运用公式 $\frac{x_{ij}-\min(x_j)}{\max(x_j)-\min(x_j)}$ 归一化处理. x_{ij} 是第 i 个样本的第 j 个属性值, $\max(x_j)$ 和 $\min(x_j)$ 是所有的样本点之中属性 j 的最大值和最小值.

步骤 2: 确定初始训练集, 测试集.

步骤 3: C4.5 决策树特征筛选.

(1) 设置损失比例. 实际中, 将信用“好”的客户误判为信用“差”的客户损失的可能仅仅是贷款利息, 而将信用“差”的客户误判为信用“好”的客户则可能遭受巨大的违约风险, 二者所造成的损失不对等, 决策树模型通过设置损失比例将可能导致的损失引入系统分析过程.

(2) 设置 *Boosting* 迭代次数. 反复 *Boosting* 迭代, 不断增大误判样本被抽为训练集的可能性, 提高模型精度.

(3) 确定决策树的修剪严重性. 对比不同修剪度, 确定决策树最佳修剪程度.

(4) 特征筛选. 在最优树下计算特征贡献率, 筛选对分类结果有较大影响属性.

步骤 4: 训练 SVM 模型.

(1) 根据步骤 3 特征筛选的结果, 组成新数据集. 采用 k 折交叉验证方法, 将全部数据集分成 k 个不相交的子集, 假设样本数为 m, 则子集就有 m/k 个样例, 每次从分好的子集中里面, 拿出一个作为测试集, 其它 k–1 个作为训练集.

(3) 训练分类器. 利用网格搜索法优化 SVM 参数 C 和核函数参数.

步骤 5: 评估 C4.5 算法优化 SVM 的个人信用评估模型效果. 选取评价指标, 并取 k 次实验结果的平均值.

2.3 模型评价指标

选择两个指标来评估模型的效果, 分别是 *F-score* 和平均准确率 *accuracy*, 这两个指标可以综合常用于信用评估的 *precision* 查准率与 *recall* 召回率. 根据混淆矩阵, 指标的计算方法如下:

$$F-score=\frac{2*recall*precision}{recall+precision} \tag{7}$$

$$precision=\frac{TP}{TP+FP} \tag{8}$$

$$recall=\frac{TP}{TP+FN} \tag{9}$$

$$accuracy=\frac{TP+TN}{TP+FP+FN+TN} \tag{10}$$

3 实证分析

3.1 数据集介绍

本文选取 UCI 机器学习库中的两组公开数据集验

证模型效果, 分别是德国信贷数据集和澳大利亚信贷数据集. 数据集的具体信息如表 2 所示, 德国信贷数据属性的详细描述如表 3 所示, 每个样本包含 20 个属性, 其中 4 个被转换为 8 个虚拟的变量最终表现为 24 维的数字变量和一个类别标签. 澳大利亚信贷数据集共有 14 个属性特征和一个类别标签.

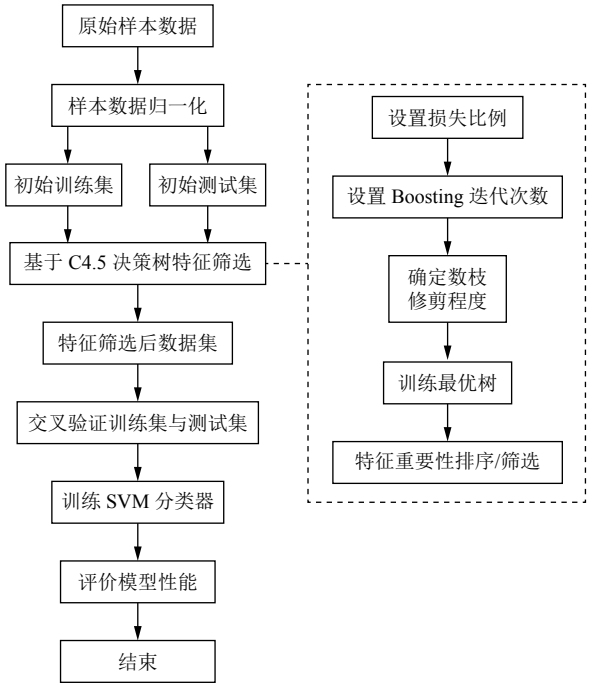

图 1 C4.5 决策树优化 SVM 模型

表 1 混淆矩阵

		实际情况	
		信用“好”	信用“差”
测试情况	信用“好”	真好 (TP)	假好 (FP)
	信用“差”	假差 (FN)	真差 (TN)

表 2 数据集信息

数据库名称	类别数目	信用“好”类别数	信用“差”类别数	样本总数	特征数目
德国数据	2	700	300	1000	20
澳大利亚数据	2	383	307	690	14

3.2 基于 C4.5 决策树算法特征提取

决策树特征提取基于 clementine12.0 平台, 采用保留法建立模型. 在两组数据集上以 4:1 的比例设置训练集和测试集, 按文献[13]研究结论, 将损失比例设为最佳 2:1, Boosting 迭代次数设置为默认值 10, 比对不同修剪程度对分类准确率的影响如表 4 所示, 可知当修剪严重性为 85 时, 德国数据集测试集与训练集分类准确率最高, 当修剪严重性为 65 时, 澳大利亚数据集测试集与训练集分类准确率均最高.

表 3 德国数据集描述

属性	数字变量	属性描述
A1	V1	现有支票账户状况
A2	V2	支票账户持续时间
A3	V3	信贷历史
A4	V4, V5	贷款目的
A5	V6	贷款金额
A6	V7	储蓄账户状况
A7	V8	在职时间
A8	V11	分期付款的比重
A9	V9	性别及婚姻状况
A10	V10	债务人或担保人
A11	V12	现居地的居住时间
A12	V13	资产状况
A13	V14	年龄
A14	V15	其他分期付款计划
A15	V16, V17	住房状况
A16	V18	本银行贷款次数
A17	V19, V20, V21	工作性质
A18	V22	提供维护的人数
A19	V23	电话
A20	V24	是否外国人

表 4 不同修剪程度决策树正确率

修剪程度	德国数据		澳大利亚数据	
	准确率		准确率	
	训练集	测试集	训练集	测试集
55	92.06	77.43	93.14	77.21
60	91.79	80.47	93.47	78.65
65	91.78	79.37	93.95	81.00
70	91.36	78.22	92.24	78.41
75	92.99	79.84	93.02	77.58
80	92.65	80.19	92.59	76.56
85	93.48	80.4	92.21	74.44

按损失比例 2:1, Boosting 迭代次数 10, 修剪严重性 85 设置生成依托德国信贷数据的最优树, 特征相对重要性排序如图 2 所示. 根据贡献度大小靠前的变量分别为: 变量 1(0.2634)、变量 4(0.1478)、变量 2(0.1352)、变量 3(0.1226)、变量 17(0.1122)、变量 5(0.076)、变量 10(0.0631)、变量 21(0.0441)、变量 13(0.02)、变量 24(0.0094)、变量 20(0.0054)、变量 16(0.0008). 按损失比例 2:1, Boosting 迭代次数 10, 修剪严重性 65 设置生成依托澳大利亚信贷数据的最优树, 特征相对重要性排序如图 3 所示. 根据贡献度大小靠前的属性为: 属性 8(0.7678)、属性 5(0.0542) 属性 3(0.052)、属性 2(0.0332)、属性 9(0.0326)、属性

14(0.0091)、属性 13(0.0077)、属性 4(0.0043)、属性 12(0.0025).

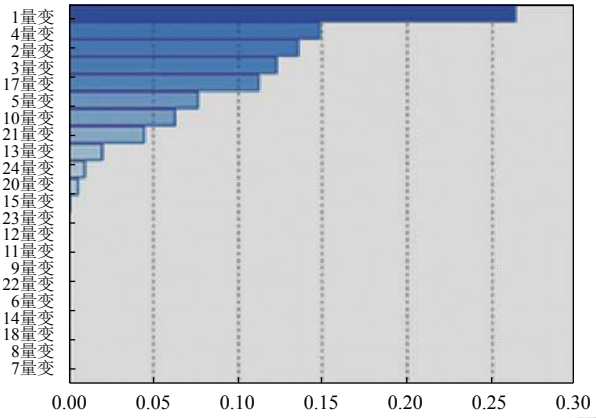

图 2 德国信贷数据特征贡献度

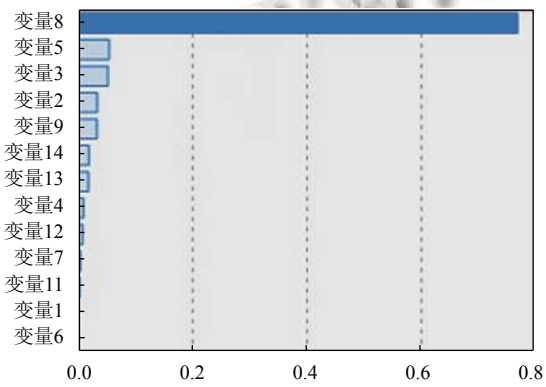

图 3 澳大利亚信贷数据特征贡献度

3.3 基于 C4.5 算法优化 SVM 的个人信用评估

经筛选后, 德国信贷数据从 25 维降低到 13 维, 澳大利亚信贷数据从 15 维降低到 10 维. 将降维后的数据作为 SVM 的输入训练分类器. 实验依托 Matlab2016a 平台, 使用 Libsvm 工具包, 采用 5 折交叉验证减少随机抽样对 SVM 分类结果的影响, 通过网格搜索法与交叉验证的方法, 确定高斯径向基核函数最优参数 γ 和支持向量机惩罚函数 C, 两个参数的网格搜索范围都是$[2^{-5},2^{5}]$,步长均为 0.2. 实验比较了 C4.5 优化 SVM 的模型 (DT+SVM) 与 C4.5 决策树 (DT), SVM 单独模型 (SVM) 以及 BP 神经网络 (BPNN)、logistic 回归, 模糊支持向量机 (B-FSVM) 这些常用于信用评估的模型分类性能. 在两组数据集上进行实验, 测试集实验结果如表 5 表 6 所示, 表格中 *F* 表示 *F-score* 指标值, *A* 代表平均准确率.

表 5 德国信贷数据实证结果

	模型	LG	SVM	DT	DT +SVM	B-F SVM	BPNN
F	1-k	0.74	0.88	0.87	0.96	0.87	0.811
	2-k	0.70	0.87	0.88	1.02	0.84	0.89
	3-k	0.73	0.78	0.85	1.02	0.84	0.89
	4-k	0.71	0.89	0.83	1.00	0.90	0.85
	5-k	0.72	0.85	0.88	1.19	0.90	0.80
	均值	0.72	0.85	0.86	1.04	0.87	0.85
	Rank	6	3	4	1	2	5
A	1-k	73.4	82.3	82.3	87.6	82.9	79.8
	2-k	82.7	83.3	82.1	88.6	83.6	79.7
	3-k	74.3	83.2	81.7	88.1	84.2	80.8
	4-k	76.2	83.7	79.6	87.9	84.6	80.1
	5-k	78.0	83.5	80.5	88.4	84.2	80.2
	均值	76.9	83.2	81.2	88.1	83.9	80.1
	Rank	6	3	4	1	2	5

表 5 给出了各个常用于信用评估的模型在德国数据集上的实验结果, 从实验结果可以看出: (1) 在每次实验中, 不论是从 F-score 还是平均准确率来看, 本文提出的模型的效果都是最优的, 证明了这种方法用于信用评估是有效的. (2) 两个不同的组合模型的效果普遍比单一模型的分类效果性能好. (3) 特征筛选后 SVM 分类效果, 比直接用于 SVM 分类 F-score 提高了 19%, 平均准确率提高了 4.9%, 可以明显得知利用 C4.5 决策树特征筛选可以弥补 SVM 的不足.

表 6 给出了各个信用评估模型用于澳大利亚信贷数据集上的结果, 可以得出如下结论: (1) 基于 C4.5 算法优化 SVM 的个人信用评估模型综合效果最好. (2) 组合模型的效果要优于单一模型. (3) 特征筛选前后, 支持向量机模型的 F-score 提升了 19%, 平均准确率提升了 5%, 说明非重要属性的减少不会降低模型效果, 反而会提升.

表 6 澳大利亚信贷数据实证结果

	模型	LG	SVM	DT	DT+SVM	B-F SVM	BPNN
F	1-k	0.73	0.88	0.88	0.93	0.88	0.83
	2-k	0.74	0.9	0.87	0.97	0.89	0.83
	3-k	0.74	0.89	0.84	1.32	0.89	0.85
	4-k	0.72	0.91	0.85	1.20	0.9	0.87
	5-k	0.75	0.87	0.85	1.00	0.89	0.86
	均值	0.74	0.89	0.86	1.08	0.89	0.85
	Rank	6	2	4	1	2	5
A	1-k	79.9	84.5	83.3	89.5	84.5	84.3
	2-k	80.0	84.1	82.7	89.0	84.1	83.9
	3-k	80.4	84.0	83.1	89.3	84.0	83.8
	4-k	80.3	84.4	81.2	89.4	84.6	84.5
	5-k	80.4	84.2	82.4	89.0	84.3	84.2
	均值	80.2	84.2	82.5	89.2	84.3	84.1
	Rank	6	3	5	1	2	4

综上所述, 本文提出的基于 C4.5 算法优化 SVM 的个人信用评估模型可以取得更加优异的性能, 具有实用性; 部分含有信息量多且对分类结果影响较大的属性, 可以代表全部的属性变量来作为建模的数据集, 并且这样训练出来的模型效果优于全部数据用于建模所取得的模型的效果, 银行或者金融机构可以参考本文方法进行信用评估.

4 结语

在信贷消费逐渐普及的高速信息化社会, 个人信用评估的研究意义越来越重要, 信用评估模型的好坏直接影响了信贷消费的走向健康和银行等金融机构的坏账率, 分类器效能哪怕很小的 1% 的提升都会挽回金融机构数以万计的损失. 考虑到支持向量机处理多数据性能下降的缺点, 本文提出基于 C4.5 算法优化支持向量机的个人信用评估方法. 该方法将 C4.5 决策树和支持向量机这两种高效的信息处理方法组合, 优势互补用于个人信用评估领域. 在 UCI 两组公开数据集上, 用 *F-score* 与平均准确率两个指标对模型测试. 实验可得, 该组合模型可以取得很好的分类效果, 有效且实用性较高, 可以为科学决策提供支持.

未来进一步研究的方向: (1)C4.5 算法由于使用了熵模型, 里面有大量的复杂的对数运算, 会导致算法复杂度高, 如何全方面考虑到信息增益又降低算法复杂度有待进一步研究. (2) 文中仅进行 SVM 二分类, 多分类问题有待研究. (3) 由于数据保密原因, 本文仅在两个公开数据集上进行了测试, 模型在其他的数据集上是否有效有待进一步的验证.

参考文献

1 Wang G, Hao JX, 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]

2 陈启伟, 王伟, 马迪, 等. 基于 Ext-GBDT 集成的类别不平衡信用评分模型. 计算机应用研究, 2018, 35(2): 421–427. [doi: 10.3969/j.issn.1001-3695.2018.02.022]

3 向小东, 宋芳. 基于核主成分与加权支持向量机的福建省城镇登记失业率预测. 系统工程理论与实践, 2009, 29(1): 73–80. [doi: 10.3321/j.issn:1000-6788.2009.01.010]

4 Guo W, Cao MY, Zheng JF. Study on Chinese banks of credit risk evaluation models of real-estate based on the BP-neural network model. Proceedings of 2009 WRI World Congress on Computer Science and Information Engineering. Los Angeles, CA, USA. 2009. 288–292.

5 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]

6 向晖, 唐剑琴. 基于 bagging 的决策树集成消费者信用评估模型. 消费经济, 2015, 31(3): 72–74.

7 谢娟英, 王春霞, 蒋帅, 等. 基于改进的 F-score 与支持向量机的特征选择方法. 计算机应用, 2010, 30(4): 993–996.

8 吴冲, 夏晗. 基于支持向量机集成的电子商务环境下客户信用评估模型研究. 中国管理科学, 2008, 16(S1): 362–367.

9 肖智, 王明恺, 谢林林. 基于支持向量机的大学生助学贷款个人信用评价. 清华大学学报 (自然科学版), 2006, 46(S1): 1120–1124.

10 姚尚锋. 基于主分量分析和 BP 神经网络的个人信用评估模型. 数学的实践与认识, 2007, 37(21): 21–24. [doi: 10.3969/j.issn.1000-0984.2007.21.005]

11 Nanni L, Lumini A. An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. Expert Systems with Applications, 2009, 36(2): 3028–3033. [doi: 10.1016/j.eswa.2008.01.018]

12 Sain SR. The nature of statistical learning theory. Technometrics, 1996, 38(4): 409.

13 Han JW. Data mining: Concepts and techniques. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc, 2005.

14 Safavian SR, Landgrebe D. A survey of decision tree classifier methodology. IEEE Transactions on Systems, Man, and Cybernetics, 1991, 23(3): 660–674.

15 Yang Y, Chen WG. Taiga: Performance optimization of the C4.5 decision tree construction algorithm. Tsinghua Science and Technology, 2016, 21(4): 415–425. [doi: 10.1109/TST.2016.7536719]