第 26 卷第 12 期
2009 年 12 月

控 制 理 论 与 应 用
Control Theory & Applications

Vol. 26 No. 12
Dec. 2009

文章编号: 1000−8152(2009)12−1365−06

基于粗糙集与支持向量机的企业短期贷款违约判别

柯孔林

(浙江工商大学 金融学院, 浙江 杭州 310018)

摘要: 建立了粗糙集和支持向量机集成的企业贷款违约判别模型, 该模型首先利用自组织映射(SOM)神经网络对具有连续属性值的财务数据进行离散处理, 并应用遗传算法约简评价指标, 然后将约简得到的最小条件属性集及相应的原始数据送入支持向量机进行训练, 最后对企业短期贷款检验样本进行违约判别. 采用贷款企业数据库558家制造业样本企业和522家房地产业样本企业进行交叉验证的实证研究, 结果表明, 与BP神经网络、多元判别分析、Logistic 等违约判别模型相比, 粗糙集和支持向量机集成的违约判别模型有更好的预测效果.

关键词: 粗糙集; 支持向量机; 神经网络; 违约判别

中图分类号: TP273　　**文献标识码**: A

Default prediction of short-term loan based on integration of rough sets and support-vector-machines

KE Kong-lin

(Finance School, Zhejiang Gongshang University, Hangzhou Zhejiang 310018, China)

Abstract: An integrated model of rough sets and support-vector-machines for the default prediction of short-term loan is proposed. The financial data is discretized by using self-organizing mapping neural network; and the evaluation indices are reduced with no information loss through genetic algorithm. The reduced indices together with relevant data are used to train support-vector-machines and discriminate between healthy and default testing samples. 558 manufacturing industry's loan firms and 522 real estate industry's loan firms are selected as test samples, The prediction accuracy of the integrated model combining rough sets and support-vector-machines is better than that of other methods such as BP neural network, multiple discriminant analysis and logistic regression.

Key words: rough set; support-vector-machines; neural network; default prediction

1　引言(Introduction)

在2004年6月发布的巴塞尔新资本协议中, 要求商业银行将内部评级法作为银行设置资本充足率、降低信用风险的标准, 而在内部评级法实施过程中, 违约可能性测度是一个基础的、必然的问题. 目前, 关于贷款企业违约可能性估算主要从以下两个方面展开: 一是利用商业银行内部评级的各个信用等级历史数据长期估算, 如JP Morgan 银行的Credit Metrics模型、Mckinsey公司的Credit Portfolio View 模型、CSFB 的Credit Risk+模型等, 由于我国商业银行违约数据库建设还不尽完善, 所以该类模型在我国的应用受到了限制. 二是利用比例分析、统计分析和人工智能技术构建违约判别模型, 但判别分析法、Logistic 法等统计分析模型和神经网络等人工智能技术都存在固有的缺陷[1]. 近年来, 支持向量机(support vector machine, SVM)受到了广泛关注, 它是以统计学习理论为基础的基于结构风险最小化原理的新的机器学习方法, 在企业贷款违约判别领域的应用越来越得到重视[2].

然而支持向量机具有不能确定数据中哪些评价指标是冗余的, 哪些是关键的, 当评价指标较多时, 会大大增加系统的处理时间, 从而影响违约判别效率这一不足, 本文将粗糙集理论作为支持向量机违约判别模型的前置系统, 提出一种粗糙集和支持向量机集成的企业贷款违约判别模型, 通过粗糙集方法减少评估系统的评价指标数量, 提高了支持向量机模型违约判别效率, 并且在基于粗糙集理论约简评价指标的具体过程中, 引入自组织映射(self-organizing mapping, SOM)神经网络对财务数据进行离散化, 引入遗传算法处理粗糙集的属性约简问题, 同时考虑到行业因素对企业违约可能性会产生影响, 本文分别采用制造业和房地产行业的大样本数

收稿日期: 2008−06−27; 收修改稿日期: 2009−03−12.
基金项目: 国家教育部人文社会科学研究项目(08JC790096); 浙江省高校人文社会科学浙江工商大学金融学重点研究基地项目.

据进行交叉验证的实证研究, 说明该方法的可行性和有效性, 为内部评级法在国内的顺利实施做铺垫.

2 基于粗糙集和支持向量机集成的贷款违约判别系统(Loan default prediction system based on integration of rough sets and support vector machine)

贷款企业的大量数据样本以二维数据表格表示, 每一行表示一个样本企业, 每一列表示一个属性, 属性可分为条件属性(即评价指标)和决策属性(即正常企业或违约企业), 以信息系统表示为$S = (U, C \cup D, V, f)$, 其中U表示对象的论域, C表示条件属性集, D表示决策属性集, $V_{\rm A}$表示$a \in C \cup D$ 的值域, $f : U \times (C \cup D) \to V$是一个信息函数. 令$R$为一等价关系族, 且$r \in R$, 如果$\mathrm{ind}(R) = \mathrm{ind}(R - \{r\})$, 则称r为R中可约去的, 否则r为R中不可约去的. 如果$p = R - \{r\}$是独立的, 则P是R中的一个约简, R 中所有不可约去的关系称为核[3].

基于粗糙集理论和支持向量机的企业贷款违约判别系统主要分为两大部分: 第1部分是模型训练, 具体步骤如下: 将粗糙集理论作为前置系统, 首先利用自组织映射(SOM) 神经网络对具有连续属性值的财务数据进行离散处理, 并利用遗传算法对属性约简去掉冗余评价指标, 得到决策表的最小条件属性集; 然后根据约简得到的最小条件属性集及相应的原始数据重新形成新的训练样本集, 送入支持向量机进行训练. 第2部分是模型检验, 将检验样本集中最小条件属性集及相应的原始数据输入到SVM 判别系统中进行违约判别, 输出判别结果.

2.1 基于SOM网络的连续属性离散化(Discretization for continuous attributes based on SOM network)

粗糙集算法仅能处理离散化数据, 所以需对连续性质的财务指标值进行离散化处理. 本文应用SOM神经网络进行离散化处理, 离散过程中只需要指定聚类数目, 离散结果能够比较客观地反映数据分布情况[4]. SOM 网络中归一化权值向量$\overline{W}_j = W_j/\|W_j\|$与$\bar{P}_k$之间的欧氏距离为

$$d_j = [\sum_{i=1}^{N} (\bar{p}_i^k - \bar{w}_{ji})^2]^{1/2}. \tag{1}$$

对邻域内节点的连接权值进行更新的表达式为

$$w_{ji}(t+1) = w_{ji}(t) + \eta(t)[p_i^k - w_{ji}(t)]. \tag{2}$$

其中$j = 1, 2, \cdots, M$, $\eta(t)$为学习率.

2.2 条件属性的约简(Reduction of attributes)

条件属性的约简是粗糙集理论的核心内容之一, 就是在保持企业贷款违约判别能力不变的条件下, 删除其中不相关或不重要的属性, 选出能明显区分正常和违约企业的关键变量. 目前有多种属性约简的计算方法, 如扩展法则、动态约简法及基于属性重要性的启发式算法等. 本文将属性约简视为多目标优化问题, 采用遗传算法计算约简, 其基本原理为: 染色体每个位串代表一个评价指标, 某位为1时表示选择其对应的评价指标, 否则去除其对应的评价指标, 这样每个位串是一个约简的侯选[5]. 定义适值函数为:

$$F(V) = \frac{n - L_{\rm V}}{n} + \frac{2C_{\rm V}}{m^2 - m}. \tag{3}$$

式中: n是评价指标的个数, $L_{\rm V}$是候选约简V所涉及到的评价指标数, $C_{\rm V}$是染色体V能区分的训练样本个数, m是训练样本总数. 该函数的前一部分的目的是希望$L_{\rm V}$的长度尽可能小, 后一部分的目的是希望区分的训练样本尽可能多.

2.3 支持向量机的训练和测试(Traning and testing of support vector machine)

采用约简得到的最小条件属性集及相应的原始数据重新形成新的训练样本集, 对支持向量机进行学习和训练的具体算法如下[6,7]: 对于给定的训练样本集, $\{x_i, y_i\}_{i=1}^n$, $x_i = (x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(m)}) \in \mathbb{R}^m$为输入向量, 即基于粗糙集理论约简后得到的最小条件属性集; $y_i \in \{-1, 1\}$为对应的期望输出, 即新的训练样本中决策属性集. 定义最优化问题为:

$$\begin{aligned} &\min_{w,b,\xi} \frac{1}{2} w^{\rm T} w + \gamma \sum_{i=1}^{n} \xi_i, \\ &\text{s.t. } y_i[\sum_{i=1}^{n} w_i \varphi_i(x_i) + b] \geqslant 1 - \xi_i,\ \xi_i \geqslant 0. \end{aligned} \tag{4}$$

其中: $\varphi_i(x)$为非线性映射, w是把特征空间连接到输出空间的权重向量, b是阀值, ξ_i为允许错分类的松弛变量, γ为一指定常数, 控制对错分样本惩罚的程度, 实现错分样本的比例与算法复杂度之间的折衷. 利用对偶原理、拉格朗日乘子法, 并采用满足Mercer条件的内积核函数$K(x_i, x_j) = \varphi(x_i)\varphi(x_j)$代替高维空间的内积$\varphi(x_i)\varphi(x_j)$, 可得非线性变换后优化问题对偶形式的线性最大化函数为:

$$\begin{aligned} &\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j), \\ &\text{s.t. } \sum_{i=1}^{n} \alpha_i y_i = 0, 0 \leqslant \alpha_i \leqslant \gamma. \end{aligned} \tag{5}$$

这是一个凸二次规划问题, 由于目标函数和约束条件都是凸的, 根据最优化理论, 这一问题存在唯一全局最优解. 在式(5)中只有少数$\alpha_i > 0$, 对应的样本点

称为支持向量; 由KKT最优条件可得阀值b, 其对应的违约判别决策函数为:

$$f(x)=\mathrm{sgn}[\sum_{i=1}^{n}\alpha_i y_i K(x_i,x)+b]. \tag{6}$$

常用的核函数有: 1) 径向基核函数(RBF): $K(x_i,x_j)=\exp(-\|x_i-x_j\|^2/2\sigma^2)$; 2) Sigmoid核函数: $K(x_i,x_j)=\tanh(kx_i\cdot x_j-\theta)$; 3) 多项式核函数: $K(x_i,x_j)=(x_i\cdot x_j+1)^d, d=1,2,\cdots$. RBF核函数将训练样本非线性地映射到更高维的空间中, 从而能够解决类标签和属性间非线性的关系问题. Sigmoid 核函数只有取一些特定参数时性能和RBF相同, 在取某些参数值时则可能无效. 多项式核函数参数数目比RBF核函数多, 因此其模型选择更为复杂且训练阶段耗时更多; 此外, 多项式核函数值在自由度大时可能会趋于无穷或零. 因此, 本文采用RBF函数作为支持向量机核函数.

3 实证分析(Experiment)

3.1 数据来源及指标变量(Data and variables)

本文选取东部地区贷款发放时间为 2002 年7月~12月, 贷款到期日为2003年12月之前的1年内大、中型制造业短期贷款企业为样本, 删除不合格和有异常值的样本, 最终得到违约企业279家(包括次级、可疑和损失贷款), 然后再随机抽样相同数量的正常企业(包括正常、关注贷款)与其配对, 共558家. 为了增加建模的样本量, 使模型能有比较好的泛化能力和鲁棒性, 本文采用3–折交叉验证的方法, 其操作方法是: 将558家样本企业随机划分成3个互不相交的子集S_1,S_2,S_3, 每个子集都包含93家正常企业和93家违约企业, 训练和检验进行3次, 在第1次迭代时, 将S_2,S_3作为训练样本, S_1作为检验样本; 在第2次迭代时, 将S_1,S_3作为训练样本, S_2作为检验样本; 在第3次迭代时, 将S_1,S_2作为训练样本, S_3作为检验样本; 训练样本违约判别精度估计时用3次迭代的训练样本正确判别数除以1116, 检验样本违约判别精度估计时用3次迭代的检验样本正确判别数除以558. 按照以上相同的方法, 最终选取东部地区房地产业违约企业261家和正常企业261家, 共522家, 3–折交叉验证时每个子集都包含87家正常企业和87家违约企业.

借鉴国内外有关文献成果和商业银行企业资信评估指标体系, 考虑数据的可得性, 笔者分别从偿债能力(包括资产负债率C_1, 负债/有形净资产C_2, 利息保障倍数C_3, 流动比率C_4, 速动比率C_5, 现金比率C_6), 财务效益(包括销售利润率C_7, 净利润率C_8, 成本费用利润率C_9, 总资产报酬率C_{10}), 资金营运(包括总资产周转率C_{11}, 固定资产周转率C_{12}, 存货周转率C_{13}, 应收账款周转率C_{14})和发展能力(包括销售收入增长率C_{15}, 利润增长率C_{16}, 总资产增长率C_{17})4个方面选取了17个定量指标作为初始财务指标变量, 见表1. 此外, 考虑到实际操作中几家企业同属于同一个集团下时, 其中一家企业违约可能会造成剩余的牵连子公司企业的违约可能性提高; 保证贷款、抵押贷款、信用贷款和质押贷款等不同担保方式对企业违约可能性的影响也不同; 当一家企业向多家银行进行贷款时, 多家银行会催促企业还款, 当企业还了一家银行贷款后, 另几家的信贷风险可能会提高, 本文引入定性指标来表示这些因素, 如贷款企业为集团企业用“1”表示, 为非集团企业用“2”表示, 见表2.

3.2 连续属性离散结果(Continuous attributes discretised values)

本文首先应用SOM网络对558家制造业样本企业财务指标变量进行离散化处理, 其中正确选取聚类数目非常关键, 聚类数目少, 可能会得到不相容的决策系统, 聚类数目多, 会出现过度离散情况, 笔者经过不断试算, 利用SOM神经网络将每个财务指标变量离散为4类, 结果见表1.

3.3 指标变量约简结果(Result of attributes reduction)

应用遗传算法求解制造业企业财务指标和定性指标最小约简时, 交叉概率 $P_{\mathrm{c}}=0.8$, 变异概率 $P_{\mathrm{m}}=0.04$, 共得到67个约简结果, 其中3个最小约简为$\{C_2, C_5, C_8, C_{12}, C_{16}\}$, $\{C_3, C_5, C_9, C_{12}, C_{15}\}$, $\{C_3, C_5, C_7, C_{11}, C_{15}\}$, 在进行企业短期贷款违约判别时, 希望约简后决策系统中的规则数目越少越好, 而每条规则的适用范围越大越好, 即评价指标的每一聚类中能够包含的对象较多, 并且产生较少的规则, 该指标可以通过聚类比来描述: $R_{\mathrm{c}}=(N_{\mathrm{o}}-N_{\mathrm{r}})/(N_{\mathrm{o}}-1)$, 式中$N_{\mathrm{o}}$为约简前的训练样本数; N_{r}为约简后的判别规则数. 根据最大聚类比的原则, 从最小约简中选取约简$\{C_3, C_5, C_9, C_{12}, C_{15}\}$对支持向量机进行训练. 本文也分别采用扩展法则法、动态约简法和基于属性重要性启发算法对制造业检验样本进行3–折交叉验证, 得出遗传算法约简后的评价指标数和检验样本的平均误判率最小, 效果最好.

3.4 SVM 模型的训练和检验结果分析(Training and testing results of SVM model)

径向基核函数(RBF)有两个参数: 误差惩罚因子γ和宽度δ^2, 这两个参数的选择对支持向量机模型预测精度起着至关重要的作用, 不适当的参数选择可能导致过度拟合或拟合不够, 但是最优参数的选择没有通常的规律, 本文采用交叉验证法来

寻找最优的参数[8]. 设置误差惩罚因子γ的范围是从1到200变化, 宽度δ^2的范围是从1到200变化, 不同参数下支持向量机3–折交叉验证的违约判别精度如表3所示.

表 1 财务指标离散化区间

Table 1 Condition attributes intervals for discretization

评价指标	符号	1	2	3	4
资产负债率	C_1	[0, 0.4127)	[0.4127, 0.6435)	[0.6435, 0.8072)	[0.8072, 1]
负债/有形净资产	C_2	[0, 1.15)	[1.15, 2.24)	[2.24, 3.18)	[3.18, +∞]
利息保障倍数	C_3	[−∞, −0.11)	[−0.11, 0.98)	[0.98, 3.61)	[3.61, +∞]
流动比率	C_4	[0, 0.68)	[0.68, 1.79)	[1.79, 2.83)	[2.83, +∞]
速动比率	C_5	[0, 0.33)	[0.33, 0.91)	[0.91, 1.47)	[1.47, +∞]
现金比率	C_6	[0, 0.04)	[0.04, 0.12)	[0.12, 0.20)	[0.20, 1]
销售利润率	C_7	[−∞, 0.0199)	[0.0199, 0.0785)	[0.0785, 0.1917)	[0.1917, +∞]
净利润率	C_8	[−∞, −0.0093)	[−0.0093, 0.057)	[0.057, 0.095)	[0.095, +∞]
成本费用利润率	C_9	[−∞, −0.0312)	[−0.0312, 0.0459)	[0.0459, 0.0936)	[0.0936, +∞]
总资产报酬率	C_{10}	[−∞, 0.0022)	[0.0022, 0.0154)	[0.0154, 0.057)	[0.057, +∞]
总资产周转率	C_{11}	[0, 0.0228)	[0.0228, 0.0733)	[0.0733, 0.1944)	[0.1944, +∞]
固定资产周转率	C_{12}	[0, 0.0173)	[0.0173, 0.1825)	[0.1825, 1.076)	[1.076, +∞]
存货周转率	C_{13}	[0, 0.437)	[0.437, 1.962)	[1.962, 3.138)	[3.138, +∞]
应收账款周转率	C_{14}	[0, 0.551)	[0.551, 2.177)	[2.177, 3.746)	[3.746, +∞]
销售收入增长率	C_{15}	[−∞, −0.1243)	[−0.1243, 0.0791)	[0.0791, 0.2148)	[0.2148, +∞]
利润增长率	C_{16}	[−∞, −0.114)	[−0.114, 0.0911)	[0.0911, 0.3676)	[0.3676, +∞]
总资产增长率	C_{17}	[−∞, −0.056)	[−0.056, 0.0721)	[0.0721, 0.148)	[0.148, +∞]

表 2 定性指标变量

Table 2 Qualitative attributes

评价指标	符号	1	2	3	4
企业规模	C_{18}	集团企业	非集团企业		
担保方式	C_{19}	保证贷款	抵押贷款	信用贷款	质押贷款
借款状况	C_{20}	仅向一家银行借款	向多家银行借款		

表 3 不同参数下支持向量机3–折交叉验证的违约判别精度(%)

Table 3 Classification accuracy of various parameters in SVM using 3-fold cross-validation

γ	$\delta^2=1$		$\delta^2=25$		$\delta^2=50$		$\delta^2=75$		$\delta^2=100$		$\delta^2=200$	
	训练	检验	训练	检验	训练	检验	训练	检验	训练	检验	训练	检验
1	86.92	73.12	81.36	76.34	79.84	75.45	78.49	74.55	77.24	72.94	76.79	72.58
10	92.20	70.97	82.97	78.67	82.71	78.85	81.00	77.24	80.20	75.45	80.02	74.73
50	94.80	68.10	84.05	79.21	85.66	81.00	83.87	79.39	83.87	78.49	82.26	78.32
75	96.33	68.64	85.39	77.96	86.29	82.80	84.77	81.72	84.41	79.93	82.26	78.67
100	97.67	67.74	86.47	77.24	86.11	81.54	85.39	82.08	84.68	80.47	83.24	79.75
150	98.03	67.56	87.28	77.42	86.83	80.29	85.66	80.11	84.59	78.67	84.14	80.11
200	98.66	67.20	87.90	76.70	86.92	79.03	85.84	79.21	84.50	77.42	84.59	79.03

图1给出了δ^2=50时, 违约判别精度与误差惩罚因子γ关系的曲线变化图, 从表3和图1可看出, 当γ值单调增加时, 制造业训练样本的违约判别精度增加, 制造业检验样本的违约判别精度开始时随γ值增加而增加, 当达到一定程度后又随γ值增加而轻微下降(δ^2=1情况例外, 其违约判别精度随γ值增加而下降), 这表明小的γ值会使训练样本权重太小而拟合不够, 导致训练样本和检验样本违约判别精度不高. 图2给出了γ=75时, 违约判别精度与宽度δ^2关系的曲线变化图, 从表3和图2可看

出, 大多数训练样本的违约判别精度随δ^2增加而下降, 大多数检验样本的违约判别精度开始时随δ^2值增加而增加, 当达到一定程度后又随δ^2值增加而下降并趋于稳定, 这表明小的δ^2值会导致训练样本过度拟合, 适当的δ^2值能得到最佳的违约判别效果. 从表3可以看出, 当γ=75和δ^2=50时, 支持向量机得到最好的检验样本违约判别精度82.8%.

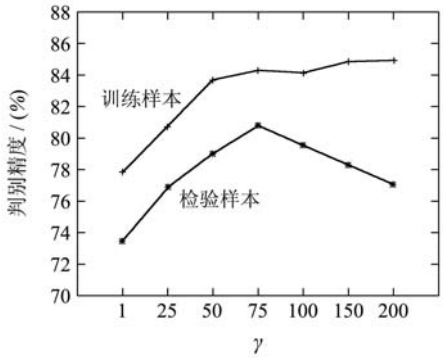

图 1　判别精度随惩罚参数γ变化折线图(δ^2=50)

Fig. 1　Classification accuracy of various γ in which δ^2=50

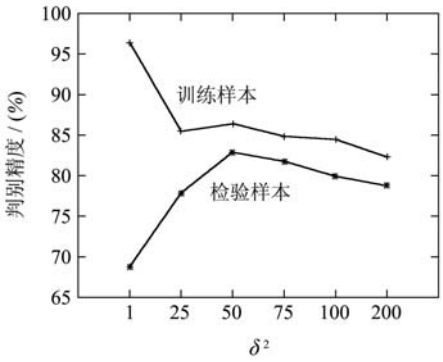

图 2　判别精度随参数δ^2变化折线图(γ=75)

Fig. 2　Classification accuracy of various δ^2 in which γ=75

本文也分别采用多项式核函数和Sigmoid核函数进行3-折交叉验证, 得出径向基核函数(RBF)对检验样本的违约判别精度略高于其他核函数, 同时考虑到径向基核函数(RBF)相对于其他核函数的优势[9], 本文最终将径向基核函数(RBF)作为支持向量机模型的核函数.

3.5　制造业检验样本各模型违约判别精度比较 (Comparing classification accuracy of manufacturing industry testing sample in different models)

为了更好地检验粗糙集与支持向量机集成模型在短期贷款违约判别中的优越性, 本文将RS-SVM模型对制造业检验样本的最好判别精度与BP神经网络、多元判别分析(MDA)和Logistic模型等最好判别精度进行比较, 通常用两类错误来度量模型的预测精度, 第1类错误定义为将违约企业判为正常企业, 第2类错误定义为将正常企业判为违约企业, 3组检验样本的违约判别精度同列于表4.

由表4得知, 第S_1, S_2, S_3组检验样本以及3组平均值显示, 粗糙集和支持向量机集成的违约判别模型发生第1类错误和第2类错误概率都明显低于BP神经网络模型、MDA模型和Logistic模型, 从3组平均误判率来看, 粗糙集和支持向量机集成的违约判别模型的误判率分别比BP神经网络模型、MDA模型和Logistic模型低4.48%、10.04%、8.96%, 说明粗糙集和支持向量机集成的模型违约判别能力最强.

表 4　制造业检验样本3-折交叉验证的违约判别结果比较

Table 4　Classification accuracy of manufacturing industry testing sample using 3-fold cross-validation

模型	S_1组			S_2组		
	第1类错误/(%)	第2类错误/(%)	总误判率/(%)	第1类错误/(%)	第2类错误/(%)	总误判率/(%)
RS-SVM	15.05	17.20	16.13	18.28	17.20	17.74
BPN	19.35	21.51	20.43	20.43	23.66	22.04
MDA	24.73	26.88	25.81	30.11	26.88	28.49
Logistic	25.81	24.73	25.27	29.03	25.81	27.42
模型	S_3组			3组平均值		
	第1类错误/(%)	第2类错误/(%)	总误判率/(%)	第1类错误/(%)	第2类错误/(%)	总误判率/(%)
RS-SVM	16.13	19.35	17.74	16.49	17.92	17.20
BPN	23.66	21.51	22.58	21.15	22.23	21.68
MDA	29.03	25.81	27.42	27.96	26.52	27.24
Logistic	26.88	24.73	25.81	27.24	25.09	26.16

3.6 房地产业检验样本各模型违约判别精度比较(Comparing classification accuracy of real estate industry testing sample in different models)

应用制造业样本同样的训练和检验方法, 得到房地产业最小约简$\{C_5, C_7, C_{12}, C_{15}, C_{19}\}$对支持向量机进行训练, 当$\gamma$=75和$\delta^2$=50时, 粗糙集和支持向量机集成模型有最好的违约判别精度, RS--SVM模型对房地产业检验样本的最好判别精度与BP神经网络、多元判别分析(MDA)和Logistic模型的最好判别精度比较结果如表5所示. 从3组平均误判率来看, 基于粗糙集和支持向量机集成的违约判别模型总误判率明显低于其它3种违约判别模型, 由低到高依次是(粗糙集和支持向量机集成模型)⊂(BP神经网络模型)⊂(logistic模型)⊂(MDA模型).

表 5 房地产业检验样本违约判别结果均值比较

Table 5 Classification accuracy of real estate industry testing sample using 3-fold cross-validation

模型	第1类错误/(%)	第2类错误/(%)	总误判率/(%)
RS–SVM	18.39	19.54	18.97
BPN	22.99	24.52	23.75
MDA	30.65	27.97	29.31
Logistic	29.12	27.97	28.54

4 结论(Conclusion)

本文提出了一种粗糙集理论与支持向量机集成的企业贷款违约判别模型, 选取我国东部地区制造业和房地产业短期贷款企业的财务数据进行3–折交叉验证的实证研究, 结果表明: 运用粗糙集理论对评价指标进行约简, 大大简化了支持向量机结构, 粗糙集和支持向量机集成的违约判别模型预测精度明显优于BP神经网络模型、MDA模型和Logistic模型, 其中制造业为82.8%, 房地产业为81.03%. 粗糙集和支持向量机集成的违约判别模型预测精度与误差惩罚因子γ和宽度δ^2有关, 对于制造业和房地产业, 当γ=75和δ^2=50时, 粗糙集和支持向量机集成模型都有最好的检验样本违约判别精度.

参考文献(References):

[1] 庞素琳, 黎荣舟, 徐建闽. BP算法在信用风险分析中的应用[J]. 控制理论与应用, 2005, 22(1): 139 – 143.
(PANG Sulin, LI Rongzhou, XU Jianmin. Application of BP algorithm in credit risk analysis[J]. *Control Theory & Applications*, 2005, 22(1): 139 – 143.)

[2] SHIN K S, LEE T S, KIM H J. An application of support vector machines in bankruptcy prediction model[J]. *Expert Systems with Applications*, 2005, 28(1): 127 – 135.

[3] MALCOLM J B, MICHAEL J P. Variable precision rough set theory and data discretisation: an application to corporate failure prediction[J]. *The International Journal of Management Science*, 2001, 29(6): 561 – 576.

[4] 飞思科技产品研发中心. 神经网络理论与MATLAB 7实现[M]. 北京: 电子工业出版社, 2005.
(FEISI R&D CENTER OF SCIENCE AND TECHNOLOGY. *Neural Network Theory and Matlab 7 Realizing*[M]. Beijing: Electronics Industry Press, 2005.)

[5] 陶志, 许宝栋, 汪定伟, 等. 基于遗传算法的粗糙集知识约简方法[J]. 系统工程, 2003, 21(4): 116 – 122.
(TAO Zhi, XU Baodong, Wang Dingwei, et al. Rough set knowledge reduction approach based on GA[J]. *Systems Engineering*, 2003, 21(4): 116 – 122.)

[6] MIN S H, LEE J, HAN I. Hybrid genetic algorithms and support vector machines for bankruptcy prediction[J]. *Expert Systems with Applications*, 2006, 31(3): 652 – 660.

[7] 肖文兵, 费奇. 基于支持向量机的个人信用评估模型及最优参数选择研究[J]. 系统工程理论与实践, 2006, 26(10): 73 – 79.
(XIAO Wenbing, FEI Qi. A study of personal credit scoring models on support vector machine with optimal choice of kernel function parameters[J]. *Systems Engineering Theory & Practice*, 2006, 26(10): 73 – 79.)

[8] MIN J H, LEE Y C. Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters[J]. *Expert Systems with Applications*, 2005, 28(4): 603 – 614.

[9] TAY F E H, CAO L J. Application of support vector machine in financial time series forecasting[J]. *Omega*, 2001, 29(4): 309 – 317.

作者简介:

柯孔林 (1978—), 男, 浙江工商大学金融学院副教授, 西安交通大学金禾经济研究中心博士研究生, 主要研究领域包括银行风险管理, E-mail: kekonglin@163.com.