第 34 卷 第 7 期 计 算 机 工 程 2008 年 4 月
Vol.34 No.7 Computer Engineering April 2008
·人工智能及识别技术· 文章编号：1000—3428(2008)07—0223—03 文献标识码：A 中图分类号：TP18

基于 GA-SVM 的企业财务困境预测

岑 涌，钟 萍，罗林开
（厦门大学信息科学与技术学院，厦门 361005）

摘 要：通过遗传算法结合支持向量机算法中期望风险边界，对我国上市公司财务数据进行特征提取，并优化构造广义最优分类超平面，从而获得具有较好整体预测性能的联合模型。数值实验表明，该方法可以降低特征空间维数，具有较好的分类准确率。实证结果表明，GA-SVM 联合预测模型具有可靠的预测财务困境能力，有着良好的应用前景。

关键词：遗传算法；支持向量机；财务困境；特征提取

Prediction Financial Distress of Firms Based on GA-SVM

CEN Yong, ZHONG Ping, LUO Lin-kai
（School of Information Science and Technology, Xiamen University, Xiamen 361005）

【Abstract】 This paper uses genetic algorithm and support vector machine to set up a hybrid model of financial distress prediction in Chinese listed firms. Numerical simulation shows that the proposed method can reduce the dimension of the feature space, and has higher correct classification rate. As the result, the proposed GA-SVM hybrid model has reliable financial distress prediction ability, and it has a good application prospect in this area.

【Key words】 genetic algorithm; support vector machine; financial distress; feature selection

财务困境又称财务危机，最严重的财务困境是企业破产。企业因财务困境导致破产实际上是一种违约行为，所以财务困境又可称为违约风险或信用风险。有效的财务困境预测，对于保护投资者和债权人的利益和经营者防范财务危机，及银行评估企业贷款信用风险，都具有十分重要的意义。因此，财务困境预测一直是学术界和金融实业界的一个研究热点。目前，财务危机预测的主流方法是分类，即根据企业的财务历史状况（主要是财务比率），将其分为正常和陷入困境两类。支持向量机(Support Vector Machine, SVM)是在统计学习理论的基础上发展起来的一种新的机器学习方法。它基于结构风险最小化原则，尽量提高学习机的泛化能力，具有良好的推广性能和较好的分类精确性。另外，支持向量机算法是一个凸优化问题，局部最优解一定是全局最优解，这些特点都是包括神经元网络在内的其他算法所不具备的。近来在金融领域得到了应用[1]，并取得不错的效果。本文利用结合遗传算法的支持向量机算法[2]，尝试对中国上市公司中被特别处理(ST)界定陷入财务困境公司，采用其相关的财务报表数据，进行特征提取，并运用于财务困境的预测，从而获得精度更高的预测模型。

1 支持向量机

支持向量机产生的二元分类器，称为最佳分类超平面，通过非线性映射将输入向量映射到高维特征空间中。

定义带有标签的训练样本 $[\boldsymbol{x}_i, y_i]$，输入向量 $\boldsymbol{x}_i \in R^n$，类的值 $y_i \in \{-1,1\}$，$i=1,2,\cdots,l$。考虑线性可分的情形，决策规则的最优超平面分离二元决策类，可以由一些支持向量给出：

$$Y = sign(\sum_{i=1}^{N} y_i \alpha_i (\boldsymbol{x} \bullet \boldsymbol{x}_i) + b) \quad (1)$$

其中，Y是结果；y_i 是训练样本 $\boldsymbol{x}_i$ 得到的类值；（•）表示为内积；向量 $\boldsymbol{x} = (x_1, x_2, \cdots, x_l)$ 为输入，向量 $\boldsymbol{x}_i$，$i=1,2,\cdots,l$ 为支持向量。在式(1)中，b和a_i为决定超平面的参数。

考虑线性不可分的情形，在高维情形下的式(1)的表达式为

$$Y = sign(\sum_{i=1}^{N} y_i \alpha_i K(\boldsymbol{x}, \boldsymbol{x}_i) + b) \quad (2)$$

支持向量分类机应用是通过执行线性约束的二次规划找到支持向量和决策参数 b 和 α_i。针对可分的情形，在式(1)中的 α_i 为下边界 0。在不可分的情形下，SVM 能利用放置上边界 C 到系数 α_i 添加到下边界来泛化。

2 GA-SVM 联合模型

2.1 遗传算法

特征提取是机器学习领域里一项复杂的复合性工作，同时又与实际问题具有很高的相关性。遗传算法(Genetic Algorithm, GA) 是一类具有较强鲁棒性的优化算法，借鉴生物的自然选择和遗传进化机制，是一种全局自适应概率搜索算法。它使用群体搜索技术，通过对当前群体施加选择、交叉、变异等一系列遗传操作，从而产生出新一代群体，并逐步使群体进化到包含或接近最优解的状态。隐含并行性和全局搜索性是遗传算法的两大显著特征。

2.2 GA-SVM 联合模型

对于每一个分类器 f 的目标就是最小化所有可能模式在不清楚分布函数 $P(x,y)$情况下的期望风险：

$$R[f] = \int_{X \times Y} \ell(\boldsymbol{x}, y, f(\boldsymbol{x})) \mathrm{d}P(\boldsymbol{x}, y) \quad (3)$$

其中，ℓ 为损失函数。但是在不清楚 $P(\boldsymbol{x}, y)$的情况下，无法

基金项目：厦门大学“985”计划基金资助项目“海量数据挖掘方法及应用”

作者简介：岑 涌(1981－)，男，硕士研究生，主研方向：智能计算，数据挖掘和机器学习；钟 萍，硕士研究生；罗林开，副教授

收稿日期：2007-04-30 **E-mail：**caesaryong@yahoo.com.cn

计算期望风险式(3)，因此，需要通过估计泛化表现。除此之外有统计方法和针对 SVM 的留一法(leave-one-out)边界估计。其中比较常见的有文献[3]提出的 R^2W^2 边界，简要说明如下：

设 ρ 为最大间隔值，$\phi(x_1),\phi(x_2),\cdots,\phi(x_n)$ 代表位于半径为 R 的球域状特征空间中的训练模式，α^* 是决定超平面参数的拉格朗日乘子。若 n 个训练数据在半径为 R 的球域状特征空间中能被间隔 ρ 分隔，则期望风险概率即边界为

$$EP_{err}^{n-1} \le \frac{1}{n} E\left\{\frac{R^2}{\rho^2}\right\} = \frac{1}{n} E\left\{R^2 W^2(\alpha^*)\right\} \quad (4)$$

其中，P^{n-1} 为测试错误概率；n-1 表示评估期望风险的样本数；R^2通过表达式 $\frac{1}{n}\sum_{i=1}^{n} k(\boldsymbol{x}_i,\boldsymbol{x}_i) - \frac{1}{n^2}\sum_{i,j=1}^{n} k(\boldsymbol{x}_i,x_j)$ 进行计算，通常这个边界被称为 R^2W^2 边界。

Holger 等人利用 R^2W^2 边界作为遗传算法中的适应度函数评估标准，联合 SVM 算法在建立模型的过程中进行特征选取，并达到优化整体模型的目的。本文同样使用该方法，对我国上市公司建立财务困境预警模型进行研究。

3 样本选取与预处理

3.1 样本选取

关于何为财务困境，文献[4]综合了学术界描述财务困境的 4 种情形，分别是经营失败(failure)、无偿付能力(insolvency)、违约(default)、破产(bankruptcy)。总体而言，财务困境的定性描述多集中在破产清算或无偿付能力等。国内学者则将特别处理(ST)的上市公司作为财务困境公司。

由于收集数据的限制，本文中对财务困境的界定以是否被ST 为准。本文将我国沪深股市的上市公司作为样本选取对象，选择了工业板块中 290 家、商业板块中 56 家以及综合板块中 80 家被ST处理的公司，根据配对原则，分别从 3 个板块中随机选择同样数目的非ST公司共同组成样本集。所选择数据均为被特殊处理前一年的数据。本文中的数据均来源于《CSMAR数据库》。早先有研究[5]表明，在设计分类器时，如果训练集中两类样本数据的数量相当，则所建模型具有较强的鲁棒性。所以本文中 3 个板块内的数据集中的ST和非ST的公司都选取其中的 70%组合作为训练样本，剩余的 30%则作为测试样本。

3.2 财务比率选择

财务比率名称及其定义如表 1 所示。

表 1 财务比率名称和计算公式

ID	比率名称(*)	比率计算公式	ID	比率名称(*)	比率计算公式
R1	净资产收益率	主营业务利润/股东权益合计	R9	资产周转率	主营业务收入净额/资产总计
R2	流动比率	流动资产合计/流动负债合计	R10	固定资产周转率	主营业务收入净额/固定资产合计
R3	速动比率	（流动资产合计－存货净额）/流动负债合计	R11	资本充足率	股东权益合计/资产总计
R4	存货流动负债比率	存货净额部/流动负债合计	R12	债务资本比率	负债合计/股东权益合计
R5	现金流动负债比率	货币资金/流动负债合计	R13	债务资产比率	负债合计/资产总计
R6	现金负债比率	货币资金/负债合计	R14	净利润率	净利润/主营业务 收入净额
R7	存货周转率	主营业务成本/存货净额	R15	资产收益率	净利润/资产总计
R8	应收账款周转率	主营业务收入净额/应收账款			

财务报表中的财务比率反映着企业不同的财务侧面，如盈利能力、偿债能力、营运能力和现金能力等方面，因此，本文总共选取了反映企业这些方面的其中 15 个财务比率，这些财务比率名称及其定义如表 1 中所示(*表示依据 CCER 中国证券市场财务数据库的数据字典)。

3.3 数据预处理

由于核函数依赖于输入样本向量的内积，而且大的属性值容易导致计算复杂，训练时间较长，因此每一维特征下所有数据都将依据式(5)归一化到区间[-1,1]：

$$x' = a + \frac{x - \min_x}{\max_x - \min_x} \times (b - a) \quad (5)$$

其中，$\min_x$和$\max_x$分别代表了特征中的最小和最大值；a和b则分别表示为归一化后的最小和最大值，在本文的实例验证中，a=-1, b=1。

4 GA-SVM 模型实证分析

4.1 核函数及相关参数确定

建立支持向量机最优模型的关键就是选择核函数以及如何确定相关参数的最优值。本文选用使用最多且比其它类核函数获得更好效果的径向基核函数作为核函数，表达式如：

$$K(\boldsymbol{x}_i,x_j) = \exp(-\gamma\left\|\boldsymbol{x}_i - x_j\right\|^2) \quad \gamma \succ 0 \quad (6)$$

在构建 SVM 预测模型时需要确定 2 个重要参数，分别是调谐参数 C 和核参数 γ。当参数值选择不适当时，将导致训练中过拟合或者欠拟合的问题。为解决这个问题，本文使用网格搜索和交叉验证结合的方法来确定最优的参数对(C, γ)，主要过程为：

(1)找到合适的正则化参数集和核参数集。按照指数增长方式生成两种参数集，比如$\{C = 2^{-5},2^{-3},2^{-1},\cdots,2^{1},\cdots,2^{5}\}$，$\{\gamma = 2^{-8},2^{-6},\cdots,2^{-1}\}$。网格搜索简单直接，因为每一个参数对($C$, γ)是独立的，可以并行地进行网格搜索。

(2)应用网格搜索法选择一个参数对(C,γ)，用该参数对进行交叉验证。

(3)循环选择参数对进行交叉验证，计算每个参数对的均方误差直到网格搜索停止，使得均方误差最小的参数对(C, γ)是最佳的。

4.2 特征选择及模型建立

本文中支持向量机算法工具使用Holger 提供的算法[2]和LIBSVM软件包。首先根据网格搜索和交叉验证得到SVM模型中最优(C,γ)参数对后，结合常用广义边界概念的遗传算法来进行特征选择和模型的建立。对表 1 中 15 个财务比率编码为 15 位二进制串(串中位数据只有 0 与 1)，编码串中的“1”表示相应位置上被选择的特征变量，结合支持向量机算法建立预测财务困境的模型。每一个被选择的特征变量子集都将采用R^2W^2边界作为适应度函数进行运算评估的标准，而不同于以前使用交叉验证方法进行评估。本文中将最大迭代次数限定为 1 000 次，即若在 1 000 次迭代过程中不能以种群最大适应度收敛则将在达到指定迭代最大次数为结束标准。

4.3 实验结果

对各板块原始数据样本通过网格搜索和交叉验证得到的各个最优的参数对分别为：工业板块数据集为(2^8,2^{-1})、商业板块数据集为(2^4,2^{-3})、综合板块数据集为(2^0,2^{-5})。然后在确定最优的参数对(C,γ)后，固定最优参数对值，针对 3 个板块的训练样本数据分别建立GA-SVM联合预测模型，并分别在各自的测试样本上进行评估。为了考察本文方法的有效性，本文在同样的训练和测试数据集上运用文献[6]提出的

RF+RM-SVM(Random Forest and RM- bound SVM)方法。表 2 列出了RF + RM-SVM和GA-SVM联合预测模型 2 种方法分别在 3 个板块上获得最佳预测结果时所选择的特征情况。

表 2　RM-SVM 和 GA-SVM 方法特征提取结果

板块集合	RF + RM-SVM 方法选取的特征	GA–SVM 方法选取的特征
工业板块	R4,R1,R15,R13,R9,R10,R8,R5	R1,R15,R9,R10,R13, R4,R5,R7,R8,R12,R14
商业板块	R15,R1,R9,R10,R3,R13,R5,R4,R8, R12	R1,R15,R9,R10,R5, R13,R5,R4,R3,R8
综合板块	R1,R9,R10,R15,R4,R3, R13,R8	R1,R15,R9,R10,R13, R8

通过特征提取后，根据表 2 中的特征，通过 RM-SVM 和 GA-SVM 建立的模型与 SVM 建立的模型的预测结果如表 3 所示。

表 3　GASVM 与 RMSVM 模型在不同板块集上的预测准确率 (%)

板块	SVM			RF + RM-SVM 模型			GA -SVM 模型		
	1 类错误	2 类错误	总体准确率	1 类错误	2 类错误	总体准确率	1 类错误	2 类错误	总体准确率
工业板块	19.54	20.69	79.89	18.39	20.69	80.46	14.94	16.09	84.48
商业板块	17.65	23.53	79.41	11.76	23.53	82.35	11.76	17.65	85.29
综合板块	29.17	20.83	75.00	25.00	20.83	77.08	20.83	16.67	81.25

其中第 1 类错误是指非 ST 企业判断为即将陷入困境企业，第 2 类错误是指财务陷入困境企业被判别为正常企业。从表 3 中可知，在工业、综合以及商业 3 个板块上，GASVM 联合模型的总体准确率都比通过 RM-SVM 模型和 SVM 模型高，分别为 84.48%、81.25%和 85.29%，而且在工业和商业板块上的第 1 类错误比第 2 类错误要低。在综合板块上，3 个模型的第 1 类准确率都要高于第 2 类错误。

5　结束语

本文以中国上市公司作为研究对象，以因财务状况异常而被特别处理作为界定上市公司陷入财务困境的标志，构造了一个预测财务困境的联合模型(GA-SVM)。该模型通过遗传算法利用R^2W^2风险边界作为适应度函数，对预测财务困境的特征属性(即输入变量)进行了选择优化，然后利用SVM在此基础上构建的最优决策面进行预测，通过实际数据样本的验证，获得了可靠的结果，证明了利用公开披露的信息研究公司财务危机的可预测性。

考虑实际情况，尚有问题值得进一步研究：在构建预测模型时考虑到两类预测错误风险的不同，在实际应用中，第一类错误成本远高于第 2 类错误成本。因此，若在模型建立时能够考虑这些因素，使得模型不仅总体错误风险最小，而且尽量降低第 1 类错误，则将使模型具有更高的实用价值。

致谢　本文课题研究过程中得到了 Holger Frohlich 和 Olivier Chapelle 教授的帮助和支持，在此表示衷心感谢。

参考文献

[1] 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.

[2] Holger F, Olivier C, Bernhard S. Feature Selection for Support Vector Machines by Means of Genetic Algorithms[EB/OL]. (2003-03-30). http://ieeexplore.ieee.org/iel5/8840/27974/ 01250182.pdf.

[3] Vapnik V, Chapelle O. Bounds on Error Expectation for Support Vector Machines[J]. Neural Computation, 2000, 12(9): 2013-2016.

[4] Altman E, Haldeman R, Narayanan P. Zeta Analysis: A New Model to Identify Bankruptcy Risk of Corporations[J]. Journal of Banking & Finance, 1977, 1(1): 29-54.

[5] Wilson R L, Ramesh S. Bankruptcy Prediction Using Neural Networks[J]. Decision Support Systems, 1994, 11(5): 545-557.

[6] Chen Yiwei, Lin Chihjen. Combining SVMs with Various Feature Selection Strategies[EB/OL]. (2007-03-13). http://www.csie.ntu.edu.tw/~cjlin/papers/features.pdf.

（上接第 222 页）

识别率明显高于全脸。对于半边人脸识别，用局域标准化方法(LN)补偿后的正确识别率高于用直方图均衡化补偿后的正确识别率，而用直方图均衡化补偿后的正确识别率高于没有做补偿的原图像的正确识别率。

表 1　半边脸和全脸正确识别率比较 (%)

人脸子集	无补偿		直方图补偿		LN 补偿	
	半脸	全脸	半脸	全脸	半脸	全脸
Subset2	73.3	70.0	100.0	98.4	100.0	98.4
Subset3	52.1	29.2	73.5	54.2	98.6	98.4
Subset4	34.1	15.0	40.0	33.3	92.6	92.5

实验还表明，本方法识别速率达到每秒 24.02 张人脸，因此，该方法可用于实时的人脸识别；半脸识别因维数比全脸识别小一半，识别速度也大于全脸识别(每秒 20.7 张)。

参考文献

[1] Adini Y, Moses Y, Ullman S. Face Recognition: The Problem of Compensating for Changes in Illumination Direction[J]. IEEE Trans. on Pattern Anal. Machine Intell., 1997, 19(7): 721-732.

[2] Xie Xudong. An Efficient Illumination Normalization Method for Face Recognition[J]. Pattern Recognition Letters, 2006, 27(6): 609-617.

[3] Son T T, Mita S. Face Recognition Under Variable Lighting Using the Mean-field Method and the Gray-level Pyramid[C]//Proc. of IEEE International Conference on Systems, Man and Cybernetics. [S. l.]: IEEE Press, 2005: 2107- 2113.

[4] Athinodoros S G. From Few to Many: Illumination Cone Models for Face Recognition Under Variable Lighting and Pose[J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2001, 23(6): 643-660.

[5] Belhumeur P N, Hespanha J P, Kriegman D J. Eigenfaces Versus Fisherfaces: Recognition Using Class Specific Linear Projection[J]. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1997, 19(7): 711-720.