公司信用风险评估新模型：基于Isomap的SVM模型

蒲晓辉

（成都农业科技职业学院 成都 611130）

【摘要】本文针对样本数据较少的特点，将基于小样本的支持向量机（SVM）方法用于我国上市公司信用风险评价中。由于考虑到财务数据特征的非线性和高维性，本文采用等距特征映射（Isomap）算法对财务指标进行特征提取，以减少数据的冗余，再针对人为选择SVM参数的盲目性，应用遗传算法优化其参数。最后通过以我国上市公司财务数据为基础的实证分析表明：基于Isomap的SVM模型比BP神经网络、PCA-SVM模型具有更强的信用风险评估能力，小样本评估准确率达到91%。

【关键词】信用风险评估 等距特征映射 支持向量机

一、信用风险评估方法总结

信用风险是指借款方由于种种原因，无力或不愿偿还贷款本息导致放贷方损失的可能性。以企业财务数据及相关资料为基础，建立数学模型对其进行科学的分析和度量，是建立风险管理系统和管理流程的有效途径。

国内外不少学者对信用风险评估问题进行了探索和研究，韩东平等（2006）以2003～2006年ST上市公司为研究对象，选取14个现金流指标建立了一个多元判别财务预警模型，该模型在财务危机发生前一年和前两年判别精度分别为93.3%和83.7%。姜秀华、孙铮（2001）以2000年11月20日为基准点，选取了沪、深证券交易所的84家上市公司（ST和非ST公司各占一半），筛选出四个财务指标建立了Logistic判别模型，财务危机发生前1年对ST公司与非ST公司判定准确率分别为88.10%和80.95%。上述多元判别分析、Logistic回归等传统评估方法局限于在假设条件下，用线性决策函数来描述信用风险与财务数据之间的非线性映射关系，存在明显缺陷。

神经网络能较好拟合二者之间的非线性关系，且无严格的假设限制，已成为信用风险评估的重要方法。杨淑娥、黄礼（2005）采用BP人工神经网络（BPNN）工具，以120家上市公司的截面财务指标作为训练集，并使用同期的60家公司作为测试集，建立了财务危机预警模型，取得了训练样本90.8%和测试样本90%的判正率。但神经网络也存在收敛速度慢，易陷入局部极小等缺点，加之该方法没有统计理论基础，解释性不强，应用受到很大限制。

财务数据特征提取是信用风险评估的重要前提，但财务指标种类繁多、差异迥然，包含了盈利能力、偿债能力、成长性、结构性等若干大类，简单地将所有指标简化或合并，会造成大量有用信息的丢失或重叠，从而影响评估精度。李杏（2009）用逐步判别法、刘淑莲等（2008）用因子分析和聚类分析法、刘彦文等（2007）用粗糙集理论等方法对上市公司的财务指标进行了属性约简。上述方法是线性提取，简单、易于实现，然而财务数据存在非线性结构，应用这些方法往往不能取得满意结果。

针对上述现状，本文采用Isomap和SVM相结合的思路建立数学模型。由于它同时具有非线性降维、支持向量机分类识别的特点，所以更适合对上市公司信用风险进行实时评估。

二、基于Isomap的SVM模型构建

1. SVM算法。SVM算法是由V.Vapnik等人于1995年在统计学理论基础上提出的一种新的学习方法，可以有效地实现对基于小样本的高维非线性系统精确拟合，并且采用结构风险最小化原则，具有很好的泛化能力。

设输入样本集为$\{(x_i,y_i)|i=1,2,\cdots,n\}$，$x\in R^d$，类别标号$y\in\{+1,-1\}$，在高维特征空间建立的分类超平面为：

$$w\cdot x+b=0 \quad (1)$$

其中：w为权向量；b为分类阈值。

若得到上式的最优分类超平面，就可以用其来对测试集进行预测了。最优超平面可以通过解下面的二次优化问题来获得：

$$\min\phi(w)=\frac{1}{2}\|w\|^2 \quad (2)$$

约束条件：$y_i(w\cdot x_i+b)\geqslant 1$

根据泛函的有关理论，只要一种核函数满足Mercer条件，它就对应某一变换空间中的内积，因此结合Lagrange方法在线性不可分的情况下，把上述二次规划问题转换为对偶问题：

$$\max W(a)=\sum_{i=1}^{n}a_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}y_iy_ja_ia_jK(x_i\cdot x_j) \quad (3)$$

约束条件：$0\leqslant a_i\leqslant C$，$\sum_{i=1}^{n}a_iy_i=0$

其中：C为误差惩罚系数；a_i为Lagrange乘子；$K(x_i\cdot x_j)$为核函数。

采用适当的核函数可实现某一非线性变换后的线性分类，而计算的复杂程度没有增加。SVM常用的核函数包括线性、多项式、径向基和多层感知器等。本文采用高斯核函数，描述为：

$$K(x_i \cdot x_j)=\exp(-\| x-x_i \|^2/2\sigma^2) \quad (4)$$

根据分类函数的正负可判断样本所属的类别，相应的分类决策函数为：

$$f(x)=\mathrm{sgn}\left(\sum_{i=1}^{n} a_i^* y_i K(x_i,x_j)+b^*\right) \quad (5)$$

其中：sgn()为符号函数；a_i^*为最优Lagrange乘子；b^*为最优分类阈值。

根据以上分析，模型将通过求解一个二次规划问题得到的决策函数来对信用风险进行评估。

2. Isomap*算法*。Isomap是Tenenbaum在2000年提出的一种基于距离保持的非线性特征提取方法，核心思想是保持点对之间的测地距离，算法如下：

设输入样本集$X=\{x_1,x_2,\cdots,x_N \mid x_i \in R^D\}$

(1)构造近邻赋权图G。

将X中每一个点与所有点进行比较，当满足公式(6)就认为它们是相邻的。

$$d(x_i,x_j)=\| x_i-x_j \|_{L^2}<\varepsilon(\text{或}i\text{是}j\text{的邻域}) \quad (6)$$

其中：$d(x_i,x_j)$为两点之间的欧氏距离；ε为固定半径。

将边长为$d(x_i,x_j)$的点连接起来即得到近邻赋权图G。

(2)计算测地距离矩阵D_G。

在近邻赋权图G上，利用Dijkstra算法计算两点之间的测地距离。

$$d_G(x_i,x_j)=\begin{cases} d(x_i,x_j) & \text{图有边} \\ \infty & \text{其他} \end{cases} \quad (7)$$

对于所有的k=1,2,…,N,利用下列公式最小化测地距离。

$$d_G(x_i,x_j)=\min\{d_G(x_i,x_j),d_G(x_i,x_k)+d_G(x_k,x_j)\} \quad (8)$$

由近邻赋权图中所有点对的最短路径组成距离矩阵$D_G=\{d_G(x_i,x_j)\}$。

(3)应用传统的MDS算法计算距离变换矩阵。

$$\tau(D_G)=-\frac{1}{2}[H(D_G)H] \quad (9)$$

其中：H是与D_G同阶的单位矩阵。

设$\tau(D_G)$的最大q个特征值$\lambda_1,\lambda_2,\cdots,\lambda_q$所对应的特征向量为$u_1,u_2,\cdots,u_q$，所构成的矩阵为$U=[u_1,u_2,\cdots,u_q]$，则得到q维低维嵌入数据Y。

$$Y=\mathrm{diag}(\sqrt{\lambda_1},\sqrt{\lambda_2},\cdots,\sqrt{\lambda_q})U^T \quad (10)$$

(4)定义残差E来衡量降维的误差，以确定降维的维数d，定义残差如下式：

$$R=1-R^2(D_G,D_Y) \quad (11)$$

其中：R^2为线性相关系数；D_Y是d维空间中的欧式距离矩阵。一般来说，降维的维数d越大，残差越小。确定d有2种情况，一是残差曲线出现拐点，二是残差小于一定的阈值。

3. *算法步骤*。传统构建线性函数来评估信用风险的方法费时且不一定最优，利用Isomap-SVM算法对其进行评估是一种有效的手段，具体过程如下：步骤1：采集上市公司数据样本集，并选择财务指标；步骤2：利用Isomap算法对数据样本集进行降维，得到低维嵌入数据；步骤3：选择SVM核函数，并用遗传算法优化参数σ和C；步骤4：将低维嵌入数据中的训练样本输入SVM模型进行训练，得到支持向量；步骤5：用低维嵌入数据中的测试样本对训练后的模型进行检验、评价。

三、实证分析

1. *样本选取*。由于我国证券市场的退市制度不完善，退市的企业不多，所以选择因“财务状况异常”而被“特别处理”(Special Treatment，ST)作为企业陷入财务危机的标志。根据上市公司披露制度，上市公司在t年是否被特别处理是由其t-1年财务报告的公布所决定的，这两个时间几乎是同时发生，用t-1年的财务报告数据来预测上市公司t年的状态应用价值不大。另外，如果一个公司在t-1年有利润，那么该公司即使在t-2年亏损，它在t年也肯定不会被ST；而如果一个公司在t-1年亏损，基于这一年数据对t年ST的预测将变成简单地对t年亏损还是盈利的预测。使用ST之前第三年的数据，则不存在这些问题。此外，何沛俐、章早立(2002)通过利用时序样本实证研究发现，在t-4年时，财务危机企业与正常企业之间的差异是不明显的。

因此，本文选取上市公司被ST之前的第t-3年截面数据为样本的时间范围，非ST公司财务数据按照其所对应的年份选取。样本的数量范围，选取沪、深证券交易所2009年100家上市公司作为训练样本，2010年100家上市公司作为测试样本，样本分布如表1所示。原始财务数据略(数据来源于CCER经济研究中心色诺芬数据库)。

表1　　样本数量分布情况

	ST公司(个)	正常公司(个)	合计	备注
训练样本	20	80	100	2009年
测试样本	20	80	100	2010年
总样本	40	160	200	

2. *指标选取*。笔者在借鉴了国内外研究成果和穆迪、标准普尔等公司资信评级指标体系后，从偿债能力、获利能力等6个方面选取了20个财务指标作为研究变量，如表2所示。这些指标的选取既考虑了公司的资产与负债能力，同时兼顾到公司的盈利与成长能力，能够充分体现上市公司的信用状况。

表2　　模型使用的财务指标

类　型	变量	财务指标名称
偿债能力	X_1	流动比率
	X_2	速动比率
	X_3	资产负债率
	X_4	营运资本资产比率
获利能力	X_5	资产净利率
	X_6	主营业务利润率
	X_7	留存收益比率
	X_8	市盈率
	X_9	每股净资产

继表

类 型	变量	财务指标名称
营运能力	X_{10}	应收账款周转率
	X_{11}	存货周转率
	X_{12}	总资产周转率
	X_{13}	流动资产周转率
成长能力	X_{14}	主营业务增长率
	X_{15}	主营利润增长率
	X_{16}	净利润增长率
现金流量	X_{17}	现金流量利息保障倍数
	X_{18}	销售收入现金实现率
资产构成	X_{19}	股东权益比率
	X_{20}	固定资产比率

3. 数据降维。Isomap算法中邻域个数k要求大于流形的维数以提高算法的稳健性。过大会导致在高度扭曲或折叠的流形上产生短路现象，过小会导致嵌入结果的偏差过大甚至近邻赋权图不连通。本文采用"留一法"的交叉验证方式，以评估诊断器对校验采样的误判个数Ne最低来优选参数k。图1为k与Ne的关系趋势曲线(d=4)，随着k的增多，误判个数先减少后增多，结合算法的稳健性最后选定k=7。图2是邻域k=7的情况下构造近邻图后获得的残差曲线，可见在维数为4时有拐点出现，并且残差的绝对值小于0.05。

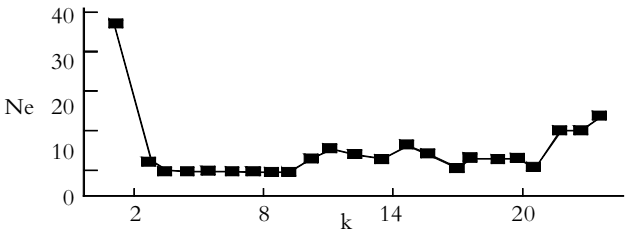

图1 不同近邻个数下的误判个数

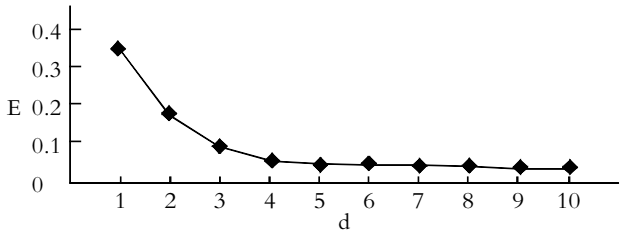

图2 残差曲线

同时，将样本数据归一化后输入SPSS软件，进行PCA降维，当主成分提取到第4个时，它们对信息累积贡献率达到86.328%，涵盖了原始指标变量85%以上的信息(为方便比较，选择与Isomap相同的降维维数)。

4. 实验结果。使用Matlab7工具箱函数，建立BPNN、PCA-SVM、Isomap-SVM三种信用风险评估模型。模型参数设置如下：BPNN输入输出节点数分别为20和1，输出层采用purelin函数，隐含层节点数为10，激活函数用tansig函数。SVM参数σ和C采用遗传算法优化，实数编码，最大进化代数取100，种群最大数量为20，交叉概率0.8，变异概率0.4，变化范围0.1~100，3倍交叉验证。

模型的性能通过三个指标来衡量：①准确率，即评估正确的公司占总样本比例；②ST误判率，即ST公司误判为正常公司的概率；③非ST误判率，即正常公司误判为ST公司的概率。实验结果如表3所示。

表3 三种评估模型对比结果

模 型	训练集			测试集		
	准确率	ST误判率	非ST误判率	准确率	ST误判率	非ST误判率
BPNN	97%	5.00%	2.50%	74.00%	25.00%	26.25%
PCA-SVM	100%	0.00%	0.00%	86.00%	15.00%	13.75%
Isomap-SVM	100%	0.00%	0.00%	91.00%	10.00%	8.75%

对比实验表明，SVM对训练集进行回测评估，准确率达到100%。BPNN对测试样本的评估精度较差，准确率仅为74%，平均误判率达到25%，SVM对测试样本进行评估，准确率均能达到85%以上，这表明SVM比BPNN评估识别精度高，小样本泛化推广能力强。引入非线性Isomap降维与线性PCA降维比较，SVM的评估准确率提高了5%，说明上市公司财务数据中存在非线性结构。

四、结论

本文提出的新型上市公司信用风险评估模型将Isomap理论和SVM技术有机结合起来，既精简了模型输入维数，降低样本间的干扰，又解决了小样本下神经网络识别精度低的问题。根据研究，得出以下结论：

第一，对于存在非线性结构的财务数据，传统的全局线性降维方法不能发现其低维结构。引入非线性降维算法Isomap对财务指标进行属性约简，有效降低了信用风险评估模型的复杂结构，提高了模型的评估精度。

第二，神经网络是研究样本无穷大时的渐进理论，基于经验风险最小化原则，而上市公司的财务数据是有限的，这就不能保证网络的泛化能力。SVM有严格的理论和数学基础，基于结构风险最小化原则，可以有效地解决上市公司有限样本条件下的高维数据模型构建问题。经对比实验表明，SVM的评估精度明显优于BPNN。

第三，SVM核函数σ和误差惩罚参数C共同影响着测试样本的分类效果和推广能力，针对目前人为选择参数的盲目性，利用遗传算法对其进行优化，达到了较好的评估效果。

主要参考文献

1. 姜秀华，孙铮.治理弱化与财务危机：一个预警模型.南开管理评论，2001；5

2. 杨淑娥，黄礼.基于神经网络的上市公司财务预警模型.系统工程理论与实践，2005；1

3. 李杏. 基于基本面的上市公司投资风险——判别分析评价.企业技术开发，2009；5

4. 刘淑莲，王真，赵建卫.基于因子分析的上市公司信用评级应用研究.财经问题研究，2008；7