基于支持向量机理论的中小企业信用风险预测研究*

沈沛龙　周　浩

内容提要：商业银行实施巴塞尔《新资本协议》内部评级体系的重要任务之一就是估计债务人违约概率，而中小企业是公司风险暴露中的重要一类。因此，研究中小企业违约概率的估计并据此确定信用级别具有重要的现实意义。本文以200家中小企业作为样本，利用具有出色分类能力的支持向量机原理，找出了作为支持向量的财务指标，并引入新的违约概率计算方法，提出了企业个体与分类面的相对距离这一概念，利用该相对距离对企业的违约概率进行估计，进而通过违约概率来确定企业的信用级别。同时，对所给模型与现有模型进行了违约概率的一致性检验，得出了理想的结果。在此基础上，获得了相对距离与KMV模型中违约距离之间的关系。

关键词：中小企业　信用风险　支持向量机　违约概率　信用评级

中图分类号：F831　　文献标识码：A

引　言

为了加强商业银行风险管理，国际银行业从2007年开始实施《新资本协议》（巴塞尔委员会，2006），我国银行业也已经有了自己的时间表，计划从2010~2013年逐步推进。《新资本协议》涉及商业银行面临的三大主要风险，即信用风险、市场风险和操作风险，尤以信用风险为重。对信用风险的识别方法主要以定量为主，且以违约概率刻画企业的信用风险。该方法从上个世纪30年代（Ramser & Foster，1931）提出开始，直至现在，才有了一个长足的发展。目前，西方较常用的评估信用风险的方法有四种：基于信用评级历史资料的信用等级违约概率（CreditMetricsTM模型）、基于期权定价理论的市场分析法（KMV模型）、基于保险精算的Credit Risk+模型以及基于宏观经济模拟的Credit Portfolio View模型（沈沛龙，2002a）。这四种模型又可以分为两大类，即利用企业财务数据模拟信用风险；采用市场数据来模拟企业的信用风险。但是，对于中国企业来说，由于中小企业参与市场的机制不够完善，其市场价格未知。因此，中国的中小企业一般采用会计截面及混合财务数据来预测企业的违约概率。

最早根据财务会计数据提出单变量分析企业破产风险预测的是Beaver（1966），Altman（1968）将其延伸至多变量，即著名的Z评分模型，这些分析均采用最小二乘法进行估计。20世纪80年代，由于最小二乘法的局限性，以极大似然法为核心的Logit模型和Probit模型成为变量估计的主流，最先使用Logit模型进行估计的是Ohlson（1980）。相对于最小二乘法，极大似然法的优势在于无需干扰项的同方差和正态分布假定，但是上述两种方法依然会受到共线性的制约，即

作者简介：沈沛龙，山西财经大学财政金融学院教授，博士生导师；周浩，山西财经大学财政金融学院硕士研究生。

* 基金项目：国家自然科学基金项目（编号：70473054；70873078）；山西省自然科学基金项目（编号：20041008）。

在线性可分的情况下，它能够把有无偿付能力两种情况的企业进行很好的区分，但是如果信用风险分析中，公司信息若不能够做到线性可分的话，Logit 模型和 Probit 模型就不会那么出色了。

伴随着计算机技术的发展，一些非线性分类方法获得了长足的发展，其中包括遗传算法（GA）、分类树（CT）和人工神经网络（ANN）等。这些方法相对于线性分类方法已经有了很大的改善，可以突破共线性的制约。而近 10 年来发展起来的支持向量机（Support Vector Machine，SVM）则是众多非线性分类方法里能够较出色完成信用风险识别的分类器。Wolfgang 等（2006）证明了支持向量机在预测精度上高于多变量分析，其精度大约改进 5%左右；在国内，陈诗一（2008）通过实证发现 SVM 方法的预测精度比 Logit 模型有明显的改进，改进幅度大约为 25%。

基于上述分析，本文利用支持向量机方法建立中小企业违约概率估计模型，具体方法是利用内嵌了线性分类方法支持向量机的国内最新研发的马克威分析系统为平台，通过对 200 家中小企业（其中包含违约类企业 60 家，非违约类企业 140 家）的财务数据进行学习、训练进而获得预测模型，找出影响其违约的主要财务因素，并选取训练集以外的 250 家样本（其中违约类数据 75 家，非违约类数据 175 家）数据进行检验。在建模过程中，我们引入相对违约距离的概念，并以此来构造新的违约概率估计模型，并考察了该模型与以往模型的拟合情况，获得了令人满意的结果。同时，本文还获得了相对违约距离与 KMV 模型中违约距离之间的关系。

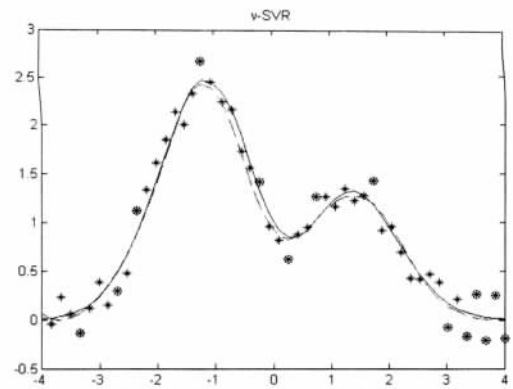
一、支持向量机分类方法

支持向量机是一种基于统计学习理论的分类方法，由 Vapnik（1995）首先提出。该理论已经在很多领域得到了广泛的应用，例如光感信号的识别、早期的医疗诊断和文本的识别与归类等。支持向量机由输入层、特征层以及输出层构成。支持向量机方法是建立在统计学习理论的 VC 维理论[①]和结构风险最小原理基础上的，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以期获得最好的泛化能力。因此支持向量机在解决小样本、非线性及高维模式识别中表现出许多特有的优势，并能够推广应用到函数拟合等其他机器学习问题中。

假定训练集为 $\{(x_i, y_i)\}_{i=1}^{n}$，其中 $x_i \in R^m$ 为 m 维的输入向量，在本文中 m=25；$y_i \in \{+1, -1\} \subset R^1$ 为输出值，这里–1 代表公司有支付能力（不违约），+1 代表公司无支付能力（违约）。我们的目的在于寻找一个潜在的分类函数 f（x），在支持向量机中，该函数的形式如下：

$$f(x)=\sum_{i=1}^{m} w_i\varphi_i(x)+b=w^T\varphi(x)+b \qquad (1)$$

其中 $\varphi(x)=(\varphi_1(x), \cdots\varphi_m(x))^T$, $w=(w_1,\cdots,w_m)^T$。$\varphi(x)$为从输入层到特征层的非线性转换；系数 w 为连接特征层到输出层的权重；b 代表阈值；同时，f（x）由于其特殊的性质，在文中亦称为因子得分。由 f（x）给出的分类面的方程为 f（x）=0。正是由于特征层的设计（该特征层一般由一个核函数通过映射来实现），把输入数据通过非线性转换函数映射到更高维数的特征空间，即 $x\rightarrow\varphi$（x）。虽然数据集在输入空间上是不可区分的，但是经这个映射转换后，特征空间却变得可分了，正因如此，支持向量机具备了出色的分类能力（参见图 1）。同时，支持向量机的优点在于，不需要把所有的变量用来作为研究数据的特征，而只需要变量的一个子集即可，这也加快了支持向量机的运算速度。



（注：圈米字点代表的是样本的支持向量，纯米字点代表的是样本内的非支持向量。由此图可见，支持向量机可以用一个非线性函数将样本进行比较完美的分类。）

资料来源：本图来自文献 Wolfgang（2006）。

图 1　引入核函数后的分类

[①]VC 维是对函数类的一种度量，是体现函数复杂程度的指标，一个问题的 VC 维越高，其计算就越复杂。

支持向量机的出色分类技术来源于其原问题设计的科学性，其原问题可以表达为：

$$\min\frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}\xi_i \quad (2)$$

$$s.t. \quad y_i\cdot\{w^T\varphi(x_i)+b\}+\xi_i\geq 1 \quad (3)$$

$$\xi_i\geq 0, \quad i=1,2,\cdots,n$$

其中，ξ 为松弛变量，在这里主要起软化功能，如果上式中的 $\xi_i=0$，则上式即退化为一个线性判别式；C 为惩罚因子，代表对起软化作用的 ξ 的惩罚程度，以此来控制最后支持向量分类的精度；可见，SVM 不仅像其他线性模型那样可以最小化系统风险，而且还可以最大化分隔距离。

二、数据和变量

研究所需的数据全部来自于某银行类金融机构，其针对中小企业的贷款占比居多，对中小企业财务信息的收集具有得天独厚的优势。研究所用样本数据均由该机构工作人员在数据库中随机抽取，且样本数据来自于各行各业，涵盖了汽贸、批发零售、建筑业、煤炭、能源、旅游以及钢铁等行业。

样本数据包含了企业近两年的资产负债表、现金流量表、损益表以及投资级企业的外部评级数据，利用这些原始数据，我们可以计算很多财务比率，根据以往文献的研究经验以及相关财务分析的常例，我们将重点采用 25 个财务指标进行分析，总体来说，一个企业是否破产将取决于这 25 个财务指标的好坏。表 1 详细描述了财务比率的分类、编号、定义以及计算方法。根据已有的研究，获利能力是一项很强的预测企业具有偿付能力与否的重要因素，因此在选取时，其所占的比重亦较大，指标数额较多。没有获利能力的企业是不具有偿付能力的；然而获利能力过高，但波动性及财务杠杆较大的企业，其偿付能力亦难保证，例如财务杠杆中的利息保障比率这一指标，从数据的反应来看，动辄超过 1000%，因此其将不入选。周转能力与流动性是获利能力的保证，均需考虑；反应公司规模的为总资产的对数；负债的变化率将作为参考指标进行考量，因为其亦包含诸多的预测信息。公司的违约或不违约的既成事实由该金融机构提供。

表 1　各财务比率的定义、分类及计算

分类	编号	计算方法	定义
获利能力	X_1	净收入/总资产	资产利润率
	X_2	净收入/销售收入	净利润幅度
	X_3	主营业务收入/总资产	营运利润率
	X_4	主营业务收入/销售收入	营运利润幅度
	X_5	息税前利润/总资产	息税前利润率
	X_6	（息税前利润+摊销折旧）/总资产	息税折旧摊销前利润率
	X_7	息税前利润/销售收入	息税前利润与销售收入比
财务杠杆	X_8	所有者权益/总资产	本金比率（简单）
	X_9	（所有者权益-无形资产）/（总资产-无形资产-现金-土地和建筑）	本金比率（复杂）
	X_{10}	流动负债/总资产	流动负债率
	X_{11}	（流动负债-现金）/总资产	净流动负债额
	X_{12}	总负债/总资产	总负债率
	X_{13}	长期负债/总资产	长期债务比率
流动性	X_{14}	现金/总资产	资产现金比率
	X_{15}	现金/流动负债	现金比率
	X_{16}	速动资产/流动负债	速动比率
	X_{17}	流动资产/流动负债	流动比率
	X_{18}	周转资金/总资产	资金周转率
	X_{19}	流动负债/总负债	负债周转率
周转能力	X_{20}	总资产/销售收入	资产周转率
	X_{21}	库存资产/销售收入	库存周转率
	X_{22}	应收账款/销售收入	应收账款周转率
	X_{23}	应付账款/销售收入	应付账款周转率
规模	X_{24}	LOG（总资产）	规模
负债变化率	X_{25}	负债变动/总负债	负债变化率

资料来源：本表选用的相关财务指标根据以往的相关做法和有关资料归纳整理。

根据国内外已有的研究，公司的流动比率、营运比率、留存收益 / 总资产和资产利润率都是企业是否遭遇财务困境的主要标志（Altman，1968；沈沛龙，2006），而这些财务比率均已包含在上述的 25 个财务指标中。

三、模型预测结果及违约概率的计算

分析软件采用马克威分析系统 4.0 版本，

数据类型采用浮点型，决策变量为整型，保留两位小数，同时不设置缺失值和标签值，变量尺度为有序型。训练过程中，惩罚因子设置为1，该值主要用于控制运算速度及运算精度。

（一）模型的建立

根据马克威分析系统4.0版本，我们可得支持向量机分类超平面函数式如下：

$$F(x)=-0.071X_1+0.22X_2-0.51X_3-0.23X_4-0.40X_5+0.41X_6-0.23X_7-0.25X_8+0.06X_9-0.28X_{10}-0.27X_{11}+0.43X_{12}+1.06X_{13}+0.04X_{14}-0.10X_{15}-0.60X_{16}-0.07X_{17}-0.24X_{18}+0.21X_{19}+0.24X_{20}+0.55X_{21}-0.49X_{22}+0.28X_{23}-0.63X_{24}-0.49X_{25} \quad (4)$$

经训练后发现，X_2和X_6与现实结果不符，即若两个比率增大，则F（x）的值等于1的概率增加，公司发生违约的概率增加。究其原因，是数据之间存在多重共线性。

为去除多重共线性，采用计量经济学中的相关系数矩阵方法，该过程在Excel中的数据分析中完成。通过观察25维相关矩阵，我们发现除X_3，X_8，X_{10}，X_{11}，X_{12}，X_{13}，X_{14}指标之间的相关系数小于0.5外，其他指标之间的相关系数值均存在大于0.5的情况，因此数据最后遴选的结果为X_3，X_8，X_{10}，X_{11}，X_{12}，X_{13}，X_{14}。再次进行支持向量机的学习，学习结果即线性支持向量机分类超平面函数为：

$$F(x)=-0.716X_3-0.596X_8-0.236X_{10}-0.266X_{11}+0.001X_{12}+1.084X_{13}-0.024X_{14} \quad (5)$$

据此，我们可以根据F（x）进行违约判定，且遵照的判断准则是

$$F(x)=\begin{cases}>0，\text{发生违约}\\<0，\text{不发生违约}\end{cases} \quad (6)$$

（二）模型的正确率检验

1. BS检验

检验方式将采用未纳入训练集的250个样本外公司数据，该数据来源于同一个金融机构，这些样本外数据无论从地域还是产业结构上都非常接近于训练集内的样本数据，因此可以作为检验用数据。

Brier Score是一种典型的用于检验判别模型精确度的评分指标（沈中华，林功勋，2006）。该指标的实际意义是求解预期违约概率和实际违约概率的最小变异，其定义如下：

$$B=\frac{1}{n}\sum_{i=1}^{n}(\hat{p}_i-I_i)^2 \quad (7)$$

式中$\hat{p}_i$为模型的预测值，I_i为实际的违约情况，在这里其违约情况假设为违约为1，不违约为0。从定义可知，B越小表示模型的预测值与实际值的差异越小，模型的检验程度越高。

由于前文将违约情况设置为1，不违约设置为-1，因此在此模型中，须经如下过程将违约检验结果进行转换：将250个公司的数据再进行一次支持向量机的预测，而所选取的预测模型为训练获得的模型，得出相关预测结果之后，取之与既成的违约事实进行比较。

$$N_i=\begin{cases}1，\text{若预测结果与即成违约事实相同}\\0，\text{若预测结果与即成违约事实相反}\end{cases}$$

则模型预测的Brier Score为：

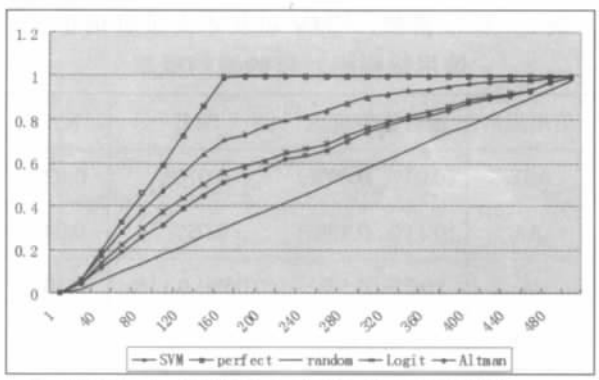
$$B=\frac{1}{n}\sum_{i=1}^{n}(\hat{p}_i-I_i)^2=\frac{1}{250}\times\sum_{i=1}^{250}N_i=221/250=0.884 \quad (8)$$

所以，在预测该银行样本外数据的中小企业违约风险中，模型的正确率为88.4%。

2. AR检验

累积精确度（CAP，Cumulative Accuracy Profiles）和精确比率（AR，Accuracy Ratios）两个统计指标一般用来进行模型之间预测性能的优劣比较。累积精度是利用数学中的Bootstrap原理，来计算各个模型中变量数据对因变量的解释能力的累积。计算过程如下，将检验所用的250个样本按违约概率从大到小的顺序排列平均分为50组，每组的样本为5个，然后查看每组中所包含的违约企业个数，例如该组中违约企业的个数为4，则该组对CAP值的贡献即为4/75=0.0533；第二组违约企业个数为5，则该组的累积精确度为（4+5）/75=0.12；依此类推，第五十组的累积精确度为1。CAP曲线如图2所示。图2中，位于最低端的光滑曲线为随机模型曲线，位于其上方的是Z模型，位于Z模型上方的为Logit模型，其上方为本文中的SVM模型，图2中最上方的折线为完美模型曲线。

本文模型的AR值即为各个模型CAP曲线



资料来源：作者根据文中模型和数据模拟获得。

图 2　SVM、Logit 和 Altman Z 模型的 CAP 曲线图示

与随机模型 CAP 曲线之间的面积与完美模型与随机模型 CAP 曲线之间面积的比值。经过测算，本文模型的 AR 值为 0.704，相同变量与数据条件下，Altman 的 Z 模型的 AR 值为 0.316，Logit 模型的 AR 值为 0.401。因此，与 Logit 模型和 Z 模型相比，本文所建立的模型具有较强的预测能力。

（三）违约概率以及信用风险级别的确定

1. 违约概率确定

根据分类器分类的数学原则，在样本个体位于违约侧时，如果距离分类平面越远，那么企业个体违约的可能性越大，即该企业的违约概率越大；如果样本个体位于非违约侧，情况则正好相反，即距离分类超平面的距离越远，则企业个体违约的可能性越小，其违约概率也越小。即

$$PD_i=\begin{cases} f(D_i), \text{当企业位于违约侧时} \\ f(-D_i), \text{当企业位于非违约侧时} \end{cases} \quad (9)$$

其中 f(x)为一个单调增函数。具体计算步骤如下：

(1)计算所有违约点（即企业个体）与分类超平面之间的距离；

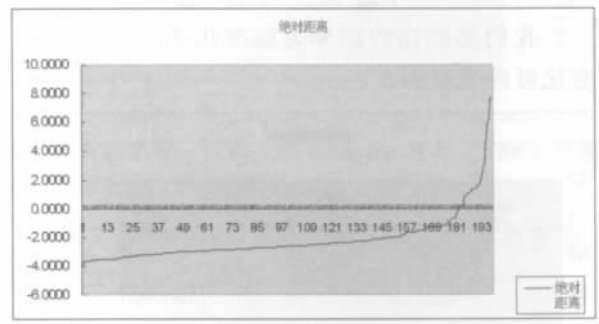
(2)量化距离与违约概率之间的对应关系，即确定级别的定义原则；

(3)根据距离以及级别确定原则，确定个体的违约概率。

根据泛函分析理论，点与超平面之间的距离为：

$$D=\frac{1}{\|w\|}\times F_i(a_1,\cdots,a_i,\cdots,a_n) \quad (10)$$

其中 D 代表点与超平面之间的距离（为表述方便，非违约侧的点与平面的距离符号取为负）；$(a_1,\cdots,a_i,\cdots,a_n)$代表点的坐标，为企业的七个财务指标，n=7；$F_i(a_1,\cdots,a_i,\cdots,a_n)$为前文中所提到的企业的因子得分；$\|w\|=\left(\sum_{i=1}^{n} w_i^2\right)^{1/2}$为向量的欧几里得范数；$w_1,\cdots,w_i,\cdots,w_n$ 代表通过支持向量机得出的分类面方程的权重。经模拟，我们得到图 3 所示的点与超平面之间的距离图：



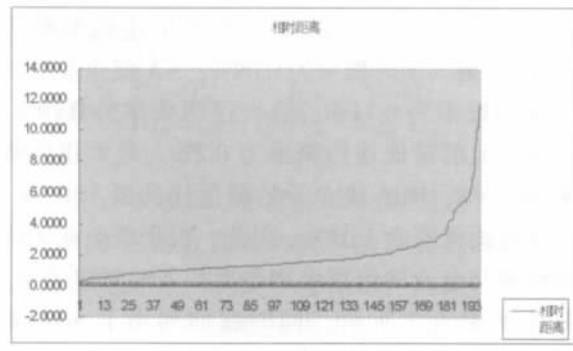
资料来源：作者根据文中模型和数据模拟获得。

图 3　点与超平面之间的距离 D 的图(绝对距离)

通过距离计算可知，两类样本中，距离最大的点位于违约侧，其值为 7.74；距离最小的点位于非违约侧，其值为–4.07。将得出的距离重新进行排序，位于违约侧的最大值记为 $D_{max}=D_{200}$；位于非违约侧的最小值记为 $D_{min}=D_1$。为计算违约概率的方便，我们引入相对距离的概念，即以距离的最小值–4.07 为起点 0，经平移变换，重新定义相对距离如下：

$$D_{相对}=D-D_{min} \quad (11)$$

由上式可知，相对距离中，最大值为 $D_{max}-D_{min}=D_{200}-D_1=11.81$，最小值为 0。



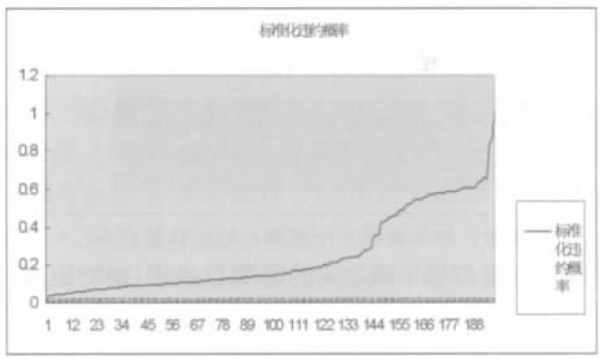
资料来源：作者根据文中模型和数据模拟获得。

图 4　修正后的相对距离

由违约距离来确定违约概率的思路是（该思路完全建立在分类器的数学原理之上，即公式9）：假定最大相对距离为1，则从违约侧的最远违约点到非违约侧的最远非违约点之间的距离可以认定其违约概率的跨度为从0到100%，即位于违约侧距离分类超平面最远的点的违约概率为100%；而位于非违约侧距离分类超平面最远的点违约概率为0。相应地，样本内数据的违约概率为：

$$P_i=D_{\text{相对}}\times\frac{1}{D_{max}-D_{min}}=\frac{D-D_{min}}{D_{max}-D_{min}} \qquad (12)$$

我们称此违约概率为标准化违约概率，其变化可以通过图5表示。



资料来源：作者根据文中模型和数据模拟获得。

图5 标准化违约概率变化图

2. 违约概率与级别的对应关系

在该金融机构提供的部分企业的外部评级数据（提供的数据包含了投资级企业的外部评级）中，AAA级企业4家，AA级企业9家，A级企业21家，BBB级企业26家。下面对该金融机构外部评级的投资级企业的违约概率进行界定（即外部评级达到BBB以上）。根据式(12)分别计算外部评级为AAA、AA、A以及BBB级企业的违约概率，AAA级企业的最小违约概率为0.03%，最大违约概率为0.09%；AA级企业的最低违约概率为0.11%，最大违约概率为0.19%；A级企业的最低违约概率为0.2%，最大违约概率为0.5%；BBB级企业的最低违约概率0.5%，最高违约概率为1.03%。因此，投资级企业的违约概率与企业的信用级别是有如下的对应关系，其关系如表2所示，同时我们列出了S&P和KMV的投资级企业的相应级别的违约概率，便于对比。

表2 标准普尔、KMV以及本文投资级企业信用级别相对应的违约概率

信用级别	违约概率范围	S&P	KMV
AAA	(0.03%, 0.09%)	0.01%	0.02%
AA	(0.11%, 0.19%)	0.03%	0.04%
A	(0.2%, 0.5%)	(0.05%, 0.11%)	0.10%
BBB	(0.5%, 1.03%)	0.28%	0.26%

通过信用级别与违约概率之间的关系，可以看出，与标准普尔以及KMV公司估计出来的级别的一般水平相比，该金融机构样本数据对应的违约概率偏高。

3. 样本内数据违约概率计算方法的检验

对于样本内数据来说，利用相对距离的方法可以计算出各个级别对应的相对违约概率，但是在实证方面却缺乏数据的相关支持，为了检验结果的合理性，我们通过Matlab数学软件，利用Wolfgang等人（2006）提出的核函数法对样本内数据的违约概率进行计算，在计算完成后，利用相关的统计指标对本文提出的基于相对距离的违约概率计算方法式（10）进行检验，根据检验结果来判定该方法的可行性。

对于样本内数据来说，Wolfgang等人（2006）提出了通过核函数计算个体违约概率的方法，即

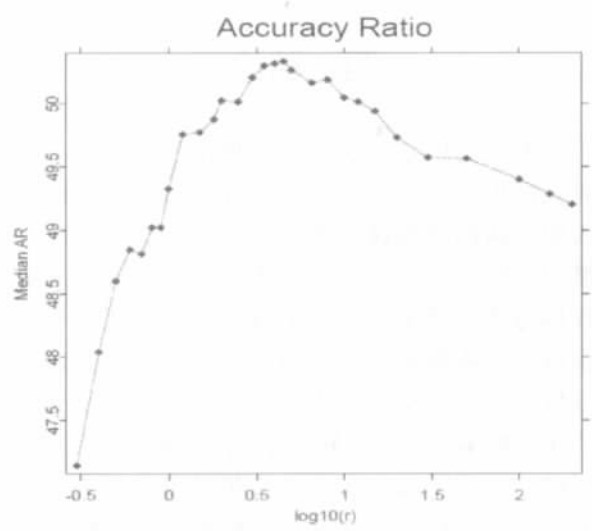
高斯核函数：

$$K(x_i,x_j)=\exp\{-(x_i-x_j)r^{-2}\Omega^{-1}(x_i-x_j)^T/2\} \qquad (13)$$

违约概率定义：

$$PD(x_i)=\frac{\sum_{j=1}^{n}K(x_i,x_j)I_{\{yj=1\}}}{\sum_{j=1}^{n}K(x_i,x_j)} \qquad (14)$$

其中$PD(x_i)$为待求个体违约概率；$K(x_i,x_j)$为支持向量机中所使用的高斯核函数；Ω为财务数据矩阵的协方差矩阵；r代表的是分类函数的复杂程度，r值越大，则其复杂度越低；r与预测精度AR之间亦存在关系（参见图6）。在该方法中，r的取值为4。$I_{y_i\in\{0,1\}}$为关于违约事件的示性函数，$I_{\{y_i=1\}}=1$表示违约，$I_{\{y_i=0\}}=0$表示不违约。



资料来源：来自文献 Wolfgang 等人（2006）。

图 6　r 与 AR 之间的变化关系

经 Matlab 编程计算得知，违约概率的范围为[4.82%,69.77%]，其范围未达到[0,100%]，但是其大部分概率值的拟合表现很好。为了对两种方法的差异进行量化分析，下面应用线性回归分析中的拟合优度、改进拟合优度以及残差平方和来对拟合程度进行评价。经计算可知：

$$R^2=\frac{SSR}{SST}=91.29\%$$

$$\text{Adjust}-R^2=\bar{R}^2=1-\frac{SSR/(T-K)}{SST/(T-1)}$$

$$=1-\frac{T-1}{T-K}(1-R^2)=91.56\%$$

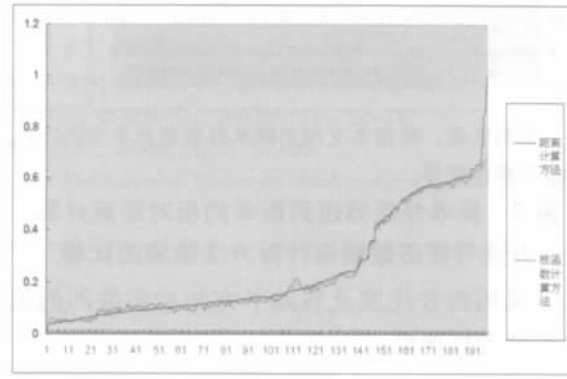
拟合优度为两种计算方法给出的违约概率的平方和的比值；改进的拟合优度则同时考虑了样本个数、变量个数等因素。其中，SSR 在此为求出的各违约概率的平方和；SST 为所提出相对距离概念计算出概率的平方和。根据经验，若拟合优度和改进的拟合优度高于 90%，则其拟合程度为优。显然，上述结果满足该条件。另外，数据进行拟合时，残差平方和也是一个较为优秀的指标。

$$\text{S.E.of Regression}=s=[u'u/(T-K)]^{1/2}=0.0324$$

其中，u 为残差的列向量；T 为样本容量；K 为变量的个数，根据前文剔除变量的结果，K=7。

若该指标小于 0.05，则拟合程度较好。显然，上述结果也满足该条件。

综上可见，上述统计检验指标及计算结果表明，利用相对距离的方法，能够很好地与 Wolfgang 等人（2006）提出的核函数法进行拟合，所以利用相对距离来计算违约概率的方法是可行的。



资料来源：根据文献 Wolfgang 等人（2006）的模型及本文提出的模型模拟获得。

图 7　相对距离计算方法与核函数计算方法所得出违约概率的比较

4. 样本外数据违约概率的计算

根据 Wolfgang 等人（2006）提出的对于样本外数据利用样本内插值计算违约概率的方法，即

$$PD(x)=PD(x_i)+\frac{f(x)-f(x_{i-1})}{f(x_i)-f(x_{i-1})}\{PD(x_i)-PD(x_{i-1})\} \quad (15)$$

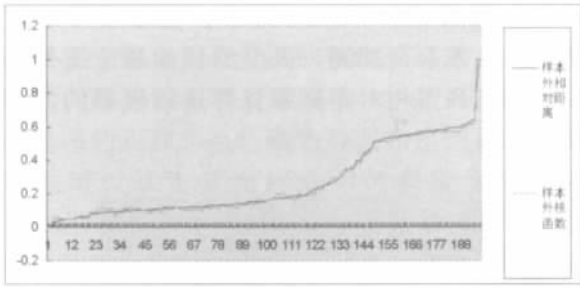
对因子得分 $f(x)$ 位于 $f(x_{i-1})$ 与 $f(x_i)$ 之间的个体 x，既可以采用上述的方法来计算。由上文式（10）知 $D=\|w\|^{-1}\cdot f(x)$，则 $f(x)=\|w\|\cdot D$，代入式（15），得

$$PD(x)=PD(x_{i-1})+\frac{D-D_{i-1}}{D_i-D_{i-1}}\{PD(x_i)-PD(x_{i-1})\} \quad (16)$$

而 $PD(x_i)=\frac{D_i-D_{min}}{D_{max}-D_{min}}$，代入式（16）得

$$PD(x)=\frac{D_{i-1}-D_{min}}{D_{max}-D_{min}}+\frac{D-D_{i-1}}{D_i-D_{i-1}}\times\frac{D_i-D_{i-1}}{D_{max}-D_{min}}=\frac{D-D_{min}}{D_{max}-D_{min}} \quad (17)$$

由上可见，对于样本外的数据进行插值推导和样本内的数据违约概率的计算在表达式上是没有区别的。下面针对样本外数据，运用相对距离方法和核函数插值方法计算 200 个样本外数据的违约概率，并用统计指标来衡量两种方法所给结果的拟合程度。



资料来源：根据本文相关样本外数据及本文所提出的模型模拟获得。

图 8　样本外数据违约概率的相对距离计算方法与核函数插值计算方法结果的比较

采用拟合优度及残差平方和对两数列的拟合程度进行评价：

$$R^2=\frac{SSR}{SST}=97.43\%$$

$$Adjust-R^2=\bar{R}^2=1-\frac{SSR/(T-K)}{SST/(T-1)}$$

$$=1-\frac{T-1}{T-K}(1-R^2)=97.51\%$$

$$S.E.of\ Regression=s=[u'u/(T-K)]^{1/2}=0.0177$$

显然，拟合优度和修正拟合优度指标值均大于 90%，拟合优度很高，拟合程度很好；残差平方和远远小于 0.05。基于上述模型统计量的表现可见，对于样本外数据来说，运用距离计算方法和运用核函数内插值计算方法在统计上是没有差异的。

5. 违约个体与分类面之间的距离与 KMV 模型中的违约距离之间的关系

在利用 KMV 模型或结构模型所确定的违约概率计算模型中，所涉及的违约距离（DD）与违约概率 PD 的关系式为：

$$PD=EDF=N(-DD) \qquad (18)$$

而结合本文的违约概率计算公式（17），应当有

$$N(-DD)=\frac{D-D_{min}}{D_{max}-D_{min}}$$

所以，以企业个体到分类面之间的距离 D 表示KMV 的违约距离 DD，有如下结果：

$$DD=-G\left(\frac{D-D_{min}}{D_{max}-D_{min}}\right) \qquad (19)$$

其中 DD 代表 KMV 中的距违约点距离；D 代表个体距分类超平面之间的距离；G 为标准正态分布函数 N 的反函数。

四、结论

本文利用基于统计学习理论的支持向量机对我国中小企业的信用状况进行研究，同时引入相对违约距离这一概念，对中小企业的违约概率进行估计，并在此基础上建立了外部评级数据与违约概率的映射关系，获得了我国中小企业信用级别对应的违约概率范围。通过比较分析可知，在相同信用级别下，样本中小企业的违约概率要高于 KMV 与 S&P 所评价的公司。同时，验证了文中所提出的相对违约距离与前人所给出的核函数计算违约概率方法在统计上的一致性。最后，还获得了本文定义的相对违约距离与 KMV 模型中的违约距离之间的关系，这一基于非结构模型给出的违约距离，在信用评级实践中具有重要的应用价值。

我国中小企业融资难的问题一直未有实质性的改变，其根源在于商业银行缺乏有效的工具和方法科学地评价企业的信用风险水平和违约状况。如何在信用质量普遍较低的中小企业集群中，准确地找到信用水平较好的企业，是每一个为中小企业服务的金融机构很渴望也是必须要完成的工作之一。本文基于统计学习理论的支持向量机原理所给出的违约距离和违约概率的计算方法，让我们很好地解决了这个问题。具体来讲，根据文中得出的中小企业判别方程，金融机构在为中小企业融资时，不应过分地看重企业的资产，而应将重点放诸于企业的财务杠杆和盈利能力等指标上。在计算中小企业的违约概率时，因为模型中相对违约距离的计算是在前述判别模型的基础上进行的，因此与判别方程直接相关联的财务指标应该重点关注。同时，在对中小企业评级中，不能盲目使用国际上著名评级机构所使用的模型结果，因为通过对比已经知道，相同级别下，国外著名评级机构的违约概率高于相对距离计算方法得出的违约概率。相对距离违约概率计算方法因结合了支持向量机所给出的分类面方程，因此在建立模型的精准程度上，会有一个很大的提高。

（责任编辑：李文杰）

参考文献：

[1] 巴塞尔银行监管委员会，中国银行业监督管理委员会翻译.统一资本计量和资本标准的国际协议——修订框架[M]. 北京：中国金融出版社，2004：55~57.

[2] 陈诗一. 德国公司违约概率预测及其对我国信用风险管理的启示[J]. 金融研究，2008（8）：58~70.

[3] 邓乃扬，田英杰. 数据挖掘中的新方法：支持向量机[M]. 北京：科学出版社，2004：168~198.

[4] 葛哲学. 精通 Matlab[M]. 北京：机械工业出版社，2007：55~101.

[5] 李国正，王猛，曾华军译. 支持向量机导论[M]. 北京：电子工业出版社，2004：68~109.

[6] 沈沛龙，任若恩. 现代信用风险管理模型和方法的比较研究[J]. 经济科学，2002 a（3）：32~41.

[7] 沈沛龙，任若恩. 新巴塞尔协议资本充足率计算方法剖析[J]. 金融研究，2002b（6）：22~31.

[8] 沈沛龙. 上市公司财务风险分析与信用评级[J]. 中国流通经济，2006（12）：57~60.

[9] Altman E. Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy [J]. The Journal of Finance, 1968 (23): 589~622.

[10] Beaver W.Financial Ratios as Predictors of Failures, Empirical Research in Accounting: Selected Studies[J]. Journal of Accounting Research, 1966 (6): 71~111.

[11] Gupten M., Finger C. CreditMetricsTM-Technical Document [R]. New York: J P Morgan Co. Inc., 1997: 109~121.

[12] KMV Corporation. Credit Monitor Overview[R]. San Francisco: KMV Corporation, 1993: 2 ~25.

[13] Krink T., Paterlini S. & Restic A. Using Differential Evolution to Improve the Accuracy of Bank Rating Systems [J] . Computational Statistics & Data Analysis, 2007 (52): 68~87.

[14] Martens D., Baesens B., Van Gestel T. & Vanthienen J. Comprehensible Credit Scoring Models Using Rule Extraction from Support Vector Machines [EB/OL]. http://www.ssrn. com /abstract=878283, 2007.

[15] Merton R. On the Pricing of Corporate Debt: the Risk Structure of Interest Rates [J]. Journal of Finance, 1974 (29): 449~470.

[16] Ohlson J. Financial Ratios and the Probabilistic Prediction of Bankruptcy [J]. Journal of Accounting Research, 1980 (1): 109~131.

[17] Ramser, J., Forster, L. A Demonstration of Ratio Analysis[R]. Urbana, Ill. University of Illinois, Bureau of Business Research, 1931, Bulletin No.40: 40~59.

[18] Ren Z. Modifications on Default Probability Calculation Methods of Commercial Banks in China[J]. China Business Review, 2008 (9): 58~66.

[19] Shen C., Lin K. Prediction of Probability of Default and Outlier--Robust Logistic Regression[J]. Journal of Financial Studies, 2005 (3): 31~32.

[20] Vapnik V. The Nature of Statistical Learning Theory[M]. New York: Springer, 1995: 4~59.

[21] Wolfgang K., Rouslan A., Schafer D. Graphical Data Representation in Bankruptcy Analysis [EB/OL]. http://edoc.hu-berlin.de/series/sfb-649-papers/2006-15/PDF/15.pdf.

Abstract: During the process of carrying out the internal rating system referred in the New Basel Capital Accord, one of the most important tasks for the commercial banks is to estimate the default probability of the debtors´ credit risk. The small and medium-sized enterprises are among the company risk exposure. Therefore, estimation of the small and medium-sized enterprises´ default probability and further definition of the credit level based on this probability is of practical significance. This paper takes 200 small and medium-sized enterprises as samples and uses SVM theory which has extraordinary classifying ability to analyze a lot of financial indices of the samples. The paper reveals the financial indices that are defined as SVM, brings in the new method of computing the default probability, and puts forward the definition of the relative distance between the individual enterprise and the classification surface. It estimates the enterprises´ default probability based on this relative distance and defines the credit level according to this default probability. At the same time, this thesis carries out a consistency test on the present model and the given model, with expected results achieved. In addition, it deduces the relations between the relative distance mentioned in this paper and the default distance in the KMV Model.

Keywords: Small and Medium-sized Enterprises; Credit Risk; SVM; Default Probability; Credit Ratings