

基于 SVM 的中小企业 集合债券融资个体信用风险度量研究

曾江洪, 王庄志, 崔晓云

(中南大学商学院, 湖南长沙, 410083)

摘要: 中小企业集合债券融资个体的信用风险度量面临小样本、非线性、高维数等现实问题, 传统的评估方法很难适用。为了弥补传统评估方法的不足, 提高信用风险度量的准确性, 建立了适用性更强的信用风险评估指标体系, 并引入基于统计学习理论的 SVM 模型对融资个体信用风险进行度量。选取径向基核函数作为支持向量机的核函数, 通过数据的转化与缩放、参数的优选, 最终获得了分类效果比较好的中小企业集合债券融资个体信用风险度量模型。经实际数据检验, 模型的预测准确率为 90.77%, 具有较强的适用性。

关键词: 信用风险; 融资个体; SVM 模型; 中小企业集合债券

中图分类号: F276.3

文献标识码: A

文章编号: 1672-3104(2013)02-0008-04

中小企业集合债券是我国对中小企业融资问题探索的成果体现, 在央行调控银行信贷规模, 中小企业融资难的情况下, 中小企业集合债券的发行可以帮助中小企业的直接融资, 为中小企业的发展提供了有力的支持。但由于中小企业自身信用水平较低, 参与集合债券的中小企业行业来源多样化, 且目前发行企业数量有限, 研究样本小, 传统的信用风险度量模型很难适用。信用风险的难以计量性一定程度上制约了中小企业集合债券的发展, 基于统计学习理论的支持向量机(SVM)模型提供了一种适宜的度量方法, 能很好的解决此问题。

统计学习理论系统地研究了经验风险最小化原理成立的条件, 是目前针对小样本统计估计和预测学习的最佳理论^[1]。支持向量机方法是 Vapnik^[2]等提出的建立在统计学习理论的 VC 维理论和结构风险最小化原理基础上的新型学习方法, 具有严格的理论基础。它以有限的样本信息为依据, 寻求模型的复杂性与学习能力间的最佳折衷, 以期获得最好的推广能力。在小样本学习方面具有的优越性, 可以较好地解决传统评估方法不能解决的非线性、高维及局部极小点等问题, 并具有较好的泛化性能, 对于未知类别样本具有相当卓越的分类与预测能力。

本文基于支持向量机模型对中小企业集合债券融

资个体的信用风险进行度量, 以期提供更为适用的度量模型, 进一步完善中小企业集合债券的风险管理与控制。

一、支持向量机模型

SVM 由线性可分情况下的最优分类超平面发展而来的。最优超平面就是要求分类线将两种类别进行正确分类, 并且使得分类的间隔最大^[3], 见图 1 所示。

给定训练集 (x_i, y_i) , $i=1, 2, \dots, m$, 其中 $\chi=\{x_1, x_2, \dots, x_m\}$ 为模式样本集, $y=\{y_1, y_2, \dots, y_m\} \in \{\pm 1\}$, 定

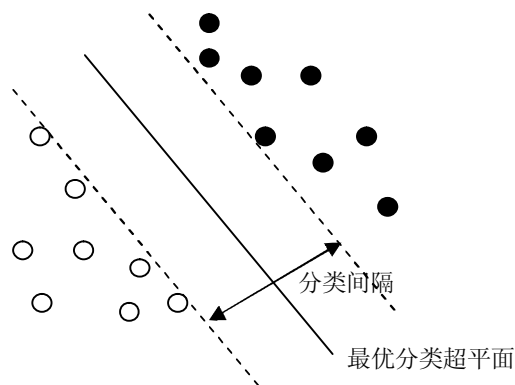


图1 最大间隔分类超平面

收稿日期: 2012-11-26; 修回日期: 2013-03-20

基金项目: 教育部人文社科基金项目“中小企业集合债券信用风险度量及控制研究”(10YJA790011)

作者简介: 曾江洪(1968-), 男, 湖南茶陵人, 中南大学商学院教授, 主要研究方向: 中小企业融资, 公司治理; 王庄志(1990-), 男, 浙江台州人, 中南大学硕士研究生, 主要研究方向: 公司治理, 风险控制。

义核函数 k 如下：

$$\begin{aligned} k: \mathcal{X} \times \mathcal{X} &\rightarrow R \\ (x, x') &\mapsto k(x, x') \end{aligned}$$

其中, x 和 x' 为 $\mathcal{X}$ 中任意两类模式, k 满足对称性。对任意两个 n 维实向量 $x, x' \in R^n$, 定义点积：

$$\langle x, x' \rangle = \sum_{i=1}^n x_i x'_i$$

其中, x_i 和 x'_i 分别为 x 和 x' 的第 i 个分量。令 $\hat{H}$ 为点积空间, 作映射：

$$\begin{aligned} \phi: x &\rightarrow \hat{H} \\ x &\mapsto X = \phi(x) \end{aligned}$$

于是, $k(x, x') = \langle X, X' \rangle = \langle \phi(x), \phi(x') \rangle$ 。

若两类样本 A、B 是线性可分的, 则存在超平面：

$$\langle w, x \rangle + b = 0$$

使得

$$\begin{cases} \text{当 } \langle w, x_i \rangle + b \geq 1 \text{ 时, } & x_i \in A \\ \text{当 } \langle w, x_i \rangle + b \leq -1 \text{ 时, } & x_i \in B \end{cases}$$

判别函数可表示为：

$$f(X) = \text{sgn}(\langle w, x \rangle + b)$$

依据统计学相关理论可知, 当训练样本集没有被超平面错误的分开, 且与超平面距离最近的样本数据和超平面之间的距离最大时, 该超平面就是最优超平面, 可表示为：

$$\max_{w \in \hat{H}, b \in R} \min \{ \|x - x_i\| \mid x \in \hat{H}, \langle w, x \rangle + b = 0, i = 1, 2, \dots, n \}$$

而距离超平面最近的样本数据与超平面之间的距离成为间隔。为了求得最优超平面, 就必须使间隔 $2/\|w\|$ 最大, 因此需要最小化 $\|w\|^2/2$ 。于是, 最优超平面的问题就可以转化为求解如下的二次规划问题^[4]：

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

$$\text{s.t. } y_i(\langle w, x_i \rangle + b) \geq 1, i = 1, 2, \dots, n$$

其 Lagrange 函数为

$$L(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^n \alpha_i (y_i(\langle w, x_i \rangle + b) - 1)$$

其中, $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_n)$, $\alpha_i \geq 0$ 为 Lagrange 系数。令

$$\begin{cases} \frac{\partial}{\partial b} L(w, b, \alpha) = 0 \\ \frac{\partial}{\partial w} L(w, b, \alpha) = 0 \end{cases}$$

可得到最优分类超平面所对应的系数 α_i^* 满足：

$$\sum_{i=1}^n \alpha_i^* y_i = 0$$

最优超平面对应的法向量 w^* 为：

$$w^* = \sum_{i=1}^n \alpha_i^* y_i x_i$$

支持向量机是在对上式的展开中那些具有非 0 的系数 α_i^* 所对应的 x_i 。可知

$$w^* = \sum_{\text{支持向量机}} \alpha_i^* y_i x_i$$

由 Kuhn-Tucker 条件可知, 当分类超平面满足：

$$\alpha_i (y_i(\langle w, x_i \rangle + b) - 1) = 0 \quad i = 1, 2, \dots, n$$

时, 超平面为最优平面。

若两类样本是线性不可分的, 则分类超平面可能不存在, 此时需要引入松弛变量 ξ_i ($i = 1, 2, \dots, n$), 于是分类超平面的最优化问题可描述为如下的二次规划问题^[5]：

$$\text{s.t. } \begin{cases} y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i \\ \xi_i \geq 0, \quad i = 1, 2, \dots, n \end{cases}$$

其中, C 表示惩罚函数, C 越大则意味着对错误分类的惩罚就越大。

Lagrange 函数为

$$\begin{aligned} L(w, b, \alpha, \beta) = & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i - \sum_{i=1}^n \alpha_i (y_i(\langle w, x_i \rangle + b) - 1 + \xi_i) - \\ & \sum_{i=1}^n \beta_i \xi_i \end{aligned}$$

其对偶问题为：

$$\max w(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle$$

判别函数可表示为：

$$\text{s.t. } \begin{cases} \sum_{i=1}^n \alpha_i y_i = 0 \\ 0 \leq \alpha \leq C, \quad i = 1, 2, \dots, n \end{cases}$$

$$f(x) = \text{sgn}(\sum_{i=1}^n \alpha_i y_i < x, x_i > + b)$$

当训练集是非线性时, 可以通过一个非线性函数 $\phi(\cdot)$, 将训练集的数据 χ 映射到一个高维线性特征空间, 映射后的样本点可以被线性分类。

最优分类超平面可以描述为:

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

$$s.t. \begin{cases} y_i (< w, \phi(x_i) > + b) \geq 1 - \xi_i \\ \xi_i \geq 0, \quad i = 1, 2, \dots, n \end{cases}$$

其对偶问题为:

$$\max w(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j < \phi(x_i), \phi(x_j) >$$

$$s.t. \begin{cases} \sum_{i=1}^n \alpha_i y_i = 0 \\ 0 \leq \alpha_i \leq C, \quad i = 1, 2, \dots, n \end{cases}$$

引入核函数 $k, < \phi(x_i), \phi(x_j) > = k(x_i, x_j)$ 。判别函数为:

$$f(x) = \text{sgn}(\sum_{i=1}^n \alpha_i y_i k(x_i, x_j) + b)$$

二、中小企业集合债券融资个体信用风险度量的实证研究

(一) 信用风险评估指标体系的构建

为了对参与集合债券融资个体的信用风险进行科学、客观的度量, 所建立的信用风险评估指标体系应该能较为全面、系统的反映参与集合债券融资企业的信用风险状况, 如偿债、盈利、运营、发展能力等。由于中小企业集合债券发行企业分散, 对全部发行企业进行调研难度较大, 而其他途径也无法获得全部企业的非财务数据, 因此度量指标的选取受到限制。为了弥补定性指标的缺失, 增强评估指标体系的全面性与适用性, 本文增加了反映参与发行企业的现金状况和资产状况的指标。具体指标如表 1 所示。

(二) 样本选取及数据预处理

截止到 2011 年底, 我国共发行了 8 支中小企业集合债券, 参与发行企业 76 家。对 76 家企业进行数据收集, 剔除数据缺失严重的企业, 最后选定 65 家发行企业作为实证研究样本, 进行三分类研究, 依据信用

表 1 中小企业集合债券融资个体信用风险评估指标体系

一级指标	二级指标	符号
偿债能力指标	流动比率	X_1
	速动比率	X_2
	资产负债率	X_3
盈利能力指标	总资产报酬率	X_4
	股东权益报酬率	X_5
	营业利润率	X_6
	销售净利率	X_7
运营能力指标	总资产周转率	X_8
	存货周转率	X_9
	应收账款周转率	X_{10}
发展能力指标	总资产增长率	X_{11}
	净利润增长率	X_{12}
其它指标	总资产	X_{13}
	销售现金比	X_{14}

风险状况将发行企业分为“信用等级较高”“信用等级一般”“信用等级较低”三类。依据本文建立的信用风险度量指标体系收集数据, 数据来源于各中小企业集合债券总募集说明书、各发行企业的分募集说明书及中国债券信息网公布的相关财务报告。

对数据进行分析前, 需要将原始数据转换为 LIBSVM 要求的数据格式:

<label> <index1>:<value1> <index2>:<value2> ...

<label>是标识类别的整数, 代表训练数据集的目标值; <index>是表示特征的序号, 是以 1 开始的整数; <value>是自变量, 为实数。本文使用 sas 软件编程对数据进行转化, 生成训练数据集和测试数据集。

此外, 缩放对 SVM 是非常重要的。缩放可以简化计算过程中数值的复杂程度, 此外还可以避免大数值区间属性对小数值区间属性的过支配。本文利用 scale 工具将数据缩放到[-1, 1]之间。

(三) 核函数选择与参数优化

1. 核函数的选择

在支持向量机模型中, 对实验结果起着至关重要作用的是核函数的选择, 核函数的选择不恰当会导致欠学习或过学习^[6]。目前常用的核函数主要有以下 4 种。

(1) 线性核函数(linear): $K(x_i, x_j) = x_i \cdot x_j$;

(2) 多项式核函数(polyominal): $K(x_i, x_j) = [x_i \cdot x_j + 1]^q$;

(3) 径向基核函数(RBF):

$$K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \gamma > 0;$$

表 2 支持向量机模型实证结果

分类名称	训练集 (45)			测试集 (20)			总数(65)
	信用等级较高 (8)	信用等级一般 (23)	信用等级较低 (14)	信用等级较高 (4)	信用等级一般 (10)	信用等级较低 (6)	
误判数	0	2	1	0	2	1	6
分类 准确率(%)	100	91.30	92.86	100	80	83.33	90.77

(4) Sigmoid 核函数:

$$K(x_i, x_j) = \tanh(\gamma(x_i \cdot x_j) + r), \gamma > 0。$$

一般情况下，首先考虑径向基函数(RBF)，主要原因如下：首先，径向基函数可以将样本映射到一个高维的空间，当样本类别标签和特征之间不是线性关系时尤为适用。线性核函数只是径向基核函数的一个特例，因此如果已经考虑选用径向基核函数，就没有必要再尝试线性核函数了^[7]。其次，核函数参数的多少会对模型的复杂程度产生影响，而与多项式核函数相比，径向基核函数需要确定的参数相对较少。当多项式核函数的阶数较大时，核矩阵元素值则会趋近于无穷小或者无穷大。相比之下，若使用径向基核函数作为支持向量机的核函数，则矩阵的元素就会落在(0, 1]区间内，有助于降低计算的困难程度。鉴于此，本文选取径向基核函数(RBF)为支持向量机的核函数，则决策函数为：

$$f(x) = \text{sgn}[\sum_{i=1}^n \alpha_i y_i \exp(-\gamma \|x_i - x_j\|^2) + b]$$

2. 参数的优化

关于支持向量机参数的优化选取，并没有公认的最优方法。本文采用交叉验证选择最佳的参数。交叉验证对与训练数据相独立的数据集有较好的泛化能力，可以避免过度拟合，是一种机器学习算法。常用的交叉验证方法包括 2-CV、k-CV、LOO-CVDENG 等^[8]。选取 k-CV 交叉验证法进行交叉验证，结果为 Best c=32.0, γ=0.007 812 5，交叉验证精度 CV rate=93.33%。

(四) 实证结果

选用 c=32.0, γ=0.007 812 5，利用 45 个训练样本数据建立中小企业集合债券融资个体信用风险评估模型，对 20 个测试样本进行分类，按照信用等级较高、一般、较低分为三类，所得结果整理如表 2 所示。

从上表可以看出，支持向量机模型的预测结果比较准确。应用模型对样本信用风险状况进行判断，具体结果见表 3。

表 3 分类结果

	样本数	分类误差	准确率(%)
训练集	45	3	93.33
测试集	20	3	85
总样本	65	6	90.77

三、结语

风险管理与控制对中小企业集合债券的发展至关重要，而风险管控的前提是对信用风险的准确度量。基于径向基核函数的支持向量机模型一定程度上解决了中小企业集合债券融资个体信用风险度量面临的小样本、非线性、高维数等现实问题。通过信用风险评估指标体系的建立、数据的转化与缩放、参数的优化，最终获得了分类效果比较好的中小企业集合债券融资个体信用风险度量模型，经实际数据检验，模型的预测准确率为 90.77%。实证结果表明，本文构建的评估指标体系及模型对中小企业集合债券信用风险度量具有较强的适用性。

参考文献:

[1] 谭东宁, 谭东汉. 小样本机器学习理论: 统计学习理论[J]. 南京理工大学学报, 2001, 25(1): 108-112.

[2] Vapnik V N. The Nature of Statistical Learning Theory [M]. Berlin: Springer, 1995.

[3] 杨斌, 路游. 基于统计学习理论的支持向量机的分类方法[J]. 计算机技术与发展, 2006, 16(11): 56-58.

[4] B. Scholkopf, A. J. Smola. Learning with kernels. Cambridge, Massachusetts [M]. London, England: The MIT Press, 2002.

[5] 王涛, 程良伦. 基于快速 SVM 的大规模网络流量分类方法[J]. 计算机应用研究, 2012, 29(6): 2301-2304.

[6] 王敬宇. 支持向量机方法及其在企业信用风险评级中的应用 [D]. 长春: 吉林大学, 2010: 4.

[7] Keerthi S S, Lin C J. Asymptotic behaviors of support vector machine with gaussian kernel [J]. Neural Computation, 2003, 15(7): 1667-1689.

[8] Hsu C, Lin C J. A comparison of methods for multi-class support vector machines [J]. IEEE Transactions on Neural Networks, 2002, 13(2): 415-425.

(下转第 19 页)