支持向量机在个人信用评估中的应用

叶小娇，李汪根，黄尧颖
（安徽师范大学 数学计算机科学学院，安徽 芜湖 241003）

摘 要：个人信用评估在银行信贷业务中有着举足轻重的作用。为了提高银行对个人信用评估的准确率，将支持向量机应用到个人信用评估中，以德国信贷数据为数据集，采用网格-5折交叉验证方法获取核函数最优参数，然后选择不同的核函数及其最优参数进行训练建模，实验得出RBF核函数更适合该数据集。针对样本中数据不平衡的问题，通过改变权重的方式对不同类别设置不同的惩罚参数。实验结果表明，该方法在保证总的预测准确率较好的前提下，有效地平衡了第一类和第二类错误率，可以作为银行信贷决策的参考依据。

关键词：信用评估；支持向量机；不平衡数据；分类
中图分类号：TP181 **文献标识码**：A **文章编号**：1673-629X(2011)03-0213-04

Application of Support Vector Machines in Personal Credit Rating

YE Xiao-jiao, LI Wang-gen, HUANG Yao-ying
(College of Mathematics and Computer Science, An hui Normal University, Wuhu 241003, China)

Abstract: Personal credit rating plays a vital role in bank credit business. In order to increase personal credit rating accuracy, support vector machines (SVM) is used to solve the problem of personal credit rating prediction in this paper. With the data set about credit from Germany, optimal parameters are obtained using 5-cross validation via parallel grid search, then four different kernel functions are selected to train the date set. The experimental results demonstrate that RBF kernel function is more suitable for the data set. As the data set is unbalanced, the rates of the first class error and the second class error are efficiently balanced by setting different punishment for different datasets under the premise of better overall prediction accuracy. It can be used as reference for the bank credit decisions.

Key words: credit rating; support vector machines; unbalanced data; classification

0 引 言

随着我国经济的快速发展，银行在获得丰厚利润的同时也要承担更多的风险。这个风险主要来自于有些贷款申请者可能不偿还贷款，因此银行必须在用户提交申请后根据已有的个人信用信息去预测该申请者到底是好客户还是坏客户，预测的准确率高低直接关系到银行承担风险的大小。可见构建一个适当的个人信用评估模型非常重要，它已成为经济活动研究的重要内容。

信用评估问题可看做是判别问题或分类问题[1~3]，已经有学者提出把统计方法、非参数统计方法、人工智能等方法用于信用评估中[4]。统计方法应用较广泛，如判别分析和Logistic回归等，但它们的缺点是线性判别分析法需要数据满足正态和等协方差的前提条件[5]，当样本点存在完全分离时，Logistic回归法可能不存在模型参数的最大似然估计。神经网络法能够有效地解决非正态分布和非线性的信用评估问题，效果比判别分析和Logistic回归方法好[6]，如陈艳的基于神经网络的覆盖算法[7]，葛继科的基于决策树-神经网络模型的近邻聚类算法[8]，但神经网络法容易陷入局部极小点，易出现过学习现象[9]。支持向量机(SVM)是一种比较适合解决小样本，非线性及高维识别问题的方法[10~12]。Nakaya等已成功地将SVM应用到个人信用评估中(采用德国数据集)，分类准确率为76%左右[13]，沈翠华等人提出了一种改进的支持向量分类方法，准确率达到了84.22%[14]，可见支持向量机方法在个人信用评估中体现出优越性[15]。

虽然沈翠华等人的方法总体准确率较高，但并未考虑具体的第一类和第二类数据准确率高低。事实上第一类预测准确率高低能反映利息收入的多少，第二类预测准确率高低能反映承担风险的大小，关心两类准确率高低对银行来说更具实际意义。文中引入支持向量机(SVM)方法来对信用评估建模。针对德国信贷数据，首先通过实验获取较好的核函数以及参数，考

收稿日期：2010-07-20；**修回日期**：2010-10-04
基金项目：安徽省自然科学研究重点项目(KJ2010A140)
作者简介：叶小娇(1983-)，女，浙江台州人，硕士研究生，研究方向为智能算法及其应用；李汪根，硕士生导师，研究方向为生物计算和智能计算。

虑到数据集的不平衡性,针对选取的核函数和参数,选择不同的惩罚参数对数据集训练建模并预测两个类别数据准确率。

1 支持向量机原理

支持向量机是 Vapnik 于 1995 年提出的[11],它在解决小样本、非线性及高维模式识别中表现出许多特有的优势。

对于线性可分问题,SVM 算法的目的是得到最优超平面,将两类样本无错误地分开且使两类的分类间隔 $\frac{2}{\|w\|}$ 最大,相当于 $\frac{2}{\|w\|^2}$ 最大,也就等同于 $\frac{\|w\|^2}{2}$ 最小化,分类超平面方程为:$(w\cdot x)+b=0$,对所有是 $y_i=1$ 的下标 i,有 $(w\cdot x_i)+b\geq 1$,对所有 $y_i=-1$ 的下标 i,有 $(w\cdot x_i)+b\leq -1$。进行归一化后,可表示成下面的形式[12]:

$$\min_{w,b}\frac{1}{2}\|w\|^2$$

$$y_i((w\cdot x_i)+b)\geq 1, i=1,\cdots,l$$

原问题相对比较难求,可以通过 Lagrang 乘子法求出它的对偶形式,从而通过求解相对简单的对偶问题来求解原分类问题的算法。其对偶问题为:

$$\min_{\alpha}\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}y_iy_j\alpha_i\alpha_j(x_i\cdot x_j)-\sum_{j=1}^{l}\alpha_j$$

$$\text{s.t.}\quad \sum_{i=1}^{l}y_i\alpha_i=0$$

$$\alpha_i\geq 0, i=1,\cdots,l$$

可求得最优解:$\alpha^*=(\sigma_1^*,\cdots,\sigma_l^*)^T$,再计算出 $w^*=\sum_{i=1}^{l}y_i\alpha_i^*x_i$,然后选择 α^* 的一个正分量 α_j^*,就可求出 $b^*=y_i-\sum_{i=1}^{l}y_i\alpha_i^*(x_i\cdot x_j)$。构造分类超平面 $(w^*\cdot x)+b=0$,求得决策函数 $f(x)=\text{sgn}(\sum_{i=1}^{l}y_i\alpha_i^*(x_i\cdot x)+b^*)$。

对于近似线性可分问题,算法中引入松弛变量来软化对间隔的要求,用 $\sum_{i=1}^{l}\varepsilon_i$ 来描述训练集被错划的程度。引进一个惩罚参数 C 作为调整间隔最大化和误差最小化这两个目标的平衡点。问题可以描述成以下的形式:

$$\min_{w,b,\varepsilon}\frac{1}{2}\|w\|^2+C\sum_{i=1}^{l}\varepsilon_i \qquad (1)$$

$$\text{s.t. } y_i((w\cdot x_i)+b)+\varepsilon_i\geq 1, i=1,\cdots,l$$

$$\varepsilon_i\geq 0, i=1,\cdots,l$$

对于线性不可分问题,通过引入核函数把输入数据映射到高维空间再进行线性分类。其决策函数是:$f(x)=\text{sgn}(\sum_{i=1}^{l}y_i\alpha_i^*K(x,x_i)+b^*)$,该算法把一个复杂的最优化问题求解转化成对样本数据进行内积运算,只需选择适当的核函数及其参数、惩罚因子就可以了。

2 基于 SVM 的个人信用评估模型

2.1 数据集

文中采用德国信用数据集 german. data-numeric。该文件是由 Strathclyde 大学在 german. data 的基础上加入 4 个属性而生成的更适合研究的数字向量矩阵,可从网站 http://archive. ics. uci. edu/ml/machine-learning-databases/statlog/german/中下载。国内外有些学者使用该数据集[13-15]。该数据集有 1000 个样本,每个样本有 24 个属性,分为好客户和坏客户两类,其中好客户样本(用标号 1 表示)有 700 个,坏客户样本(用标号 2 表示)有 300 个。在本实验中随机选择其中 660 个样本作为训练集,剩下的 340 个样本作为测试集,并且使训练集和测试集中好客户和坏客户的比都是 7:3。

2.2 核函数及参数的选择

核函数 K 的选择需要满足 Mercer 条件。目前常用的核函数主要有四种,分别是线性核函数、多项式核函数、RBF 核函数以及 Sigmoid 核函数。不同的核函数可以产生不同的支持向量机,因此核函数的选取对实验结果是有影响的,针对不同的数据集选择哪个核函数较合适,目前没有统一的定论[16]。文章将利用实验的方法来确定适合本数据集的核函数。

参数的选取也是影响模型好坏的一个重要方面[18]。目前寻找最佳参数的方法较多,如双线性网格寻参、网格寻参、基于遗传算法寻参等。文中采用网格-5 折交叉验证搜索法来获取不同核函数的最佳参数对。所谓 5 折交叉验证就是随机地将数据剖分为 5 等分,依次将其中的一份作为测试集,其余 4 份作为训练集进行试验,得出相应的正确率并求其平均值作为最终的准确率[18]。

2.3 样本非均衡问题

本实验中好客户与坏客户的数据比是 7:3,很明显该数据是不平衡的。传统的 SVM 算法在解决这类数据时它的分类器具有很大的偏向性,对较少类不利。而实际中把好客户预测成坏客户比把坏客户预测成好客户的损失要少[19],相对来说,更关注坏客户的预测准确率。解决不平衡分类问题可以从两方面进行考虑:一是改变训练样本,降低不平衡程度;二是适当地修改算法使之适应不平衡分类问题。文中通过设置不

同的惩罚因子来解决不平衡问题。问题可描述成以下形式[16]:

$$\min_{\omega,b,\varepsilon_i,\varepsilon_i^*} \frac{1}{2}\|\omega\|^2 + C_1\sum_{i=1}^{l}\varepsilon_i + C_2\sum_{i=2}^{l}\varepsilon_i^* \qquad (2)$$

好客户惩罚因子为 C_1,坏客户惩罚因子为 C_2,可以通过设置不同的 C_1 和 C_2 来影响 C-SVM 对好客户和坏客户的分类准确度。不同的 C_1,C_2 值是通过权重 W 来设置的。第一类的惩罚参数 C_1 的值用 $W_1 * C$ 表示,第二类的惩罚参数 C_2 的值用 $W_2 * C$ 来表示。公式(1)其实是公式(2)的特殊情况,此时 $W_1 = W_2 = 1$。

3 实验及分析

3.1 实验步骤

本实验主要是利用 libsvm 工具箱来完成,主要步骤如下:

(1)对数据进行归一化处理。

文中我们对数据进行了[0,1]规范化,这样做的目的是以免某个特征值过大,在算距离时主导了结果,同时也避免了训练时为了计算核函数而计算内积的时候引起数值计算的困难。训练集和测试集用相同的归一化方式。

(2)分别对四种核函数进行网格-5 折交叉验证法搜索,找出不同核函数的最佳参数。

(3)选择较好的核函数及参数,设置不同的惩罚参数权重,对训练集进行训练建模。

(4)利用建立的模型分别对训练集和测试集进行预测。

3.2 实验结果

为了寻找一个比较适合该数据的核函数,分别对四种核函数进行网格-5 折交叉验证法实验,实验结果见表1。

表1　不同核函数的最佳参数及5折交叉验证准确率

Function	c	g	d	r	5折交叉验证准确率
Linear	2.000	-	-	-	75.15%
Polynominal	0.125	0.2500	3.0	2.0	74.70%
RBF	16.000	0.0625	-	-	76.06%
Sigmoid	8.000	0.0625	-	0	75.00%

从表1中可以看出四种核函数的准确率都在75%左右,RBF 核函数准确率达到了76.06%,具有较好的准确率,因此选择 RBF 核函数及其相应的参数进行下面的实验。

为了了解不同惩罚参数对第一类准确率(正确预测好客户)第二类准确率(正确预测坏客户)的影响,设置了不同的权重,选择 RBF 核函数及参数对数据集

表2　不同权重的训练集与测试集预测准确率

权重	训练集			测试集		
	第一类准确率(%)	第二类准确率(%)	总的准确率(%)	第一类准确率(%)	第二类准确率(%)	总的准确率(%)
W1=2.0,W2=1.0	99.35	29.80	78.48	97.90	22.55	75.29
W1=1.8,W2=1.0	98.92	33.84	79.39	97.06	30.39	77.06
W1=1.6,W2=1.0	98.70	44.44	82.42	95.80	36.27	77.94
W1=1.4,W2=1.0	96.75	53.03	83.64	92.44	45.10	78.24
W1=1.2,W2=1.0	95.89	59.60	85.00	92.02	49.02	79.12
W1=1.0,W2=1.0	94.16	64.65	85.30	89.08	51.96	77.94
W1=1.0,W2=1.2	90.26	70.71	84.39	85.71	58.82	77.65
W1=1.0,W2=1.4	87.88	75.76	84.24	83.19	61.76	76.76
W1=1.0,W2=1.6	86.36	77.27	83.64	82.35	63.73	76.76
W1=1.0,W2=1.8	83.12	80.81	82.42	78.15	65.69	74.41
W1=1.0,W2=2.0	80.52	83.33	81.36	75.21	67.65	72.94
W1=1.0,W2=2.2	79.65	85.35	81.36	72.27	69.61	71.47
W1=1.0,W2=2.4	78.79	87.88	81.52	71.85	71.57	71.76
W1=1.0,W2=2.6	77.27	90.40	81.21	70.17	72.55	70.88
W1=1.0,W2=2.8	75.32	92.93	80.61	67.23	75.49	69.71
W1=1.0,W2=3.0	74.46	93.94	80.30	65.97	76.47	69.12
W1=1.0,W2=3.2	72.73	93.94	79.09	65.13	75.49	68.24
W1=1.0,W2=3.4	71.65	93.94	78.33	64.29	75.49	67.65
W1=1.0,W2=3.6	71.21	93.94	78.03	63.45	74.51	66.76
W1=1.0,W2=3.8	73.35	94.95	77.73	62.18	75.49	66.18
W1=1.0,W2=4.0	70.13	95.45	77.73	60.92	75.49	65.29

进行训练建模预测。实验中的 $c = 16.000, g = 0.0625$，所有的准确率采用四舍五入的形式保留两位小数。

3.3 实验分析

(1)从表1可见，选择不同的核函数，其5折交叉验证准确率是不同的，可见核函数的选择对预测准确率是有影响的。

(2)图1是表2中测试集准确率的折线图，从图中可以看出，随着第一类与第二类权重比值的减少，第一类准确率呈下降趋势，而第二类准确率先上升后趋向稳定。总的预测准确率起先是慢慢的上升，直到在权重是(1.2,1)的时候达到了最大，为79.12%，随后准确率又开始慢慢呈下降趋势。由此可见，设置不同的惩罚参数对预测第一类和第二类的准确率是有影响的。

(3)从图1中可知，当权重 $W_1:W_2$ 为1:2.4的时候，第一类准确率，第二类准确率和总体准确率非常接近，而数据集中好客户和坏客户的比是7:3，约为2.33:1，2.33与2.4这两个数字比较接近，它们之间是一种巧合还是有一定的联系需要进一步的研究证实。

(4)银行是个盈利机构，它希望利润最大化，风险最小化。理想的信用评估模型应该要尽可能地使第一类准确率越高越好(把好客户预测成好客户)而第二类错误率越低越好(把坏客户预测成好客户)。一方面，从图1中可以看到，随着第二类权重的增加，第一类准确率和第二类准确率大小变化基本上呈相反的趋势。另一方面，第二类误判的损失远远高于第一类误判的损失。银行必须权衡这两方面的得与失，选择一个比较合适的权重。

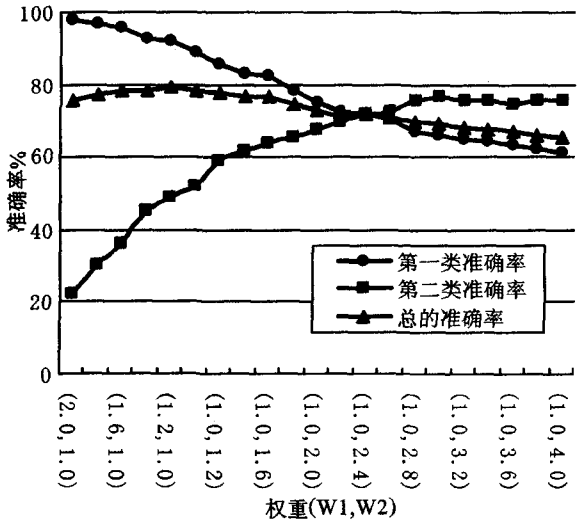

图1 不同权重值的测试集准确率折线图

4 结束语

文中运用SVM对个人信用评估进行了分析。分别采用了不同的核函数以及相应的最佳参数对训练集和数据集进行预测实验。对于样本中数据不平衡的问题，通过改变权重的方式设置不同类别惩罚参数的方法，在保证总的预测准确率较好的前提下，有效的平衡了第一类和第二类错误率。从实验结果分析，文中的方法对于不平衡数据集的分类是有效的。

下一步的研究方向主要是通过改进SVM方法及进一步处理原始数据这两个方面来提高个人信用评估预测准确率。再者，考虑把信用评估问题从二分类问题扩展成多分类问题，有利于银行掌握较详细的客户信用信息，针对不同的信用类别设置不同的贷款政策。

参考文献：

[1] Anderson T W. An introduction to multivariate statistical analysis[M]. New York,NY: Wiley,1984.

[2] Anderson J A, Rosenfeld E. Neurocomputing: Foundations of research[M]. Cambridge: MIT Press,1988.

[3] Arminger G, Enache D, Bonne T. Analyzing credit risk data: A comparison of logistic discrimination classification tree analysis and feed forward networks[J]. Computational Statistics, 1997,12(2):293-310.

[4] Thomas L C. A survey of credit and behavioral scoring: Forecasting financial risks of lending to customers[J]. International Journal of Forecasting, 2000,16(2):149 - 172.

[5] Reichert A K, Cho C C, Wagner G M. An examination of the conceptual issues involved in developing credit-scoring models [J]. Journal of Business and Economic Statistics, 1983(1): 101 - 114.

[6] Lee Tian-Shyug, Chiu Chih-Chou. Credit scoring using the hybrid neural discriminant technique[J]. Expert Systems with Applications,2002, 23(3): 245 - 254.

[7] 陈 艳，张燕平. 数据挖掘技术在保险客户信用评估的应用[J]. 计算机技术与发展,2008,18(5):179-181.

[8] 葛继科，赵永进. 数据挖掘技术在个人信用评估模型中的应用[J]. 计算机技术与发展,2006,16(12):172-174.

[9] Craven M W, Shavlik J W. Using neural networks for data mining[J]. Future Generation Computer Systems, 1997,13(2): 221-229.

[10] 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报,2000,26(1):32-42.

[11] Cortes, Vapnik V. Support - vector networks[J]. Machine Learning, 1995,20(3):273 - 297.

[12] 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,2004:164-223.

[13] Nakaya A, Furuukawa H, Morishita S. Weighted Majority Decision Among Several Region Rules for Scientific Discovery [J]. Discovery Science, 1999,1721(3): 17-29.

[14] 沈翠华，刘广利，邓乃扬. 一种改进的支持向量机分类方法及其应用[J]. 计算机工程, 2005, 31(8): 153-154.

[15] 肖文兵，费 奇. 基于支持向量机的个人信用评估模型及最

(下转第220页)

控电压、电流、转速的变化，以及电压、电流、转速运行对比，方便实验者观察电机运转状况。本文给出了 LabVIEW 与 1#DSP 控制主板通信时电机转速的图形变化情况，其中电机运动时转速波形如图 5 所示。

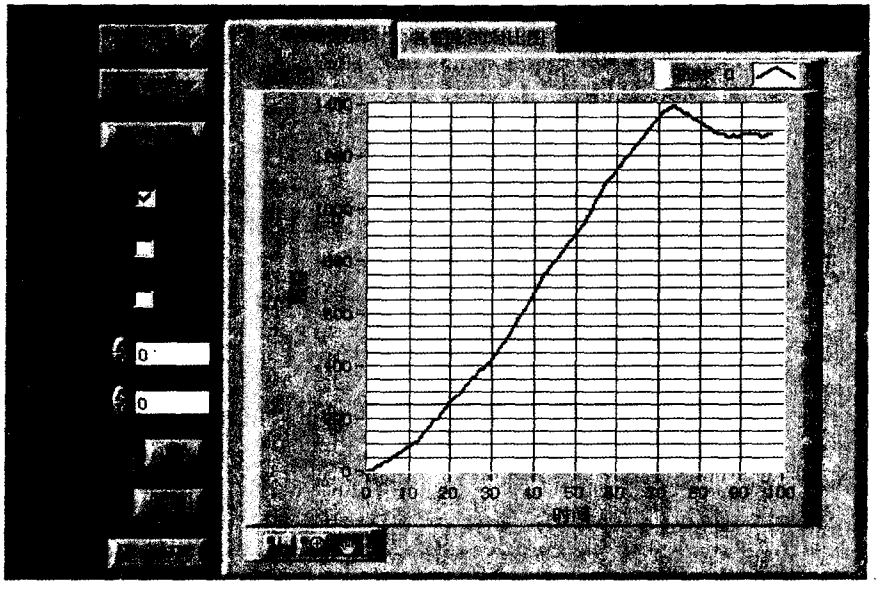

图 5 显示转速时的控制界面图

4 结束语

这一电机集群控制系统被应用在无刷直流电机实验系统中，并取得了想要的结果。通过 RS-232 总线实现了 LabVIEW 与集内控制主板 DSP 的串行通信。实践证明，该平台具有以下特点：操作简单，界面友好，使用者直观的观察电机转动的参数，同时可以调节下位机控制参数；可移植性强，可以应用在无刷直流电机集群控制系统中，简单的改变一些参数就可以应用到异步电机控制系统中，单片机控制的步进电机控制系统中等；使用起来灵活，编译生成 exe 文件，可以方便的安装在没有安装 LabVIEW 的机器上。

但由于受 RS-232C 串行通信的限制，只能用于短距离的数据发送与接收，在今后改进方面上，可对此试验平台做进一步的扩展，如硬件系统性能的提高，成本的进一步降低，其他总线接口的扩展，用户控制界面的改进，无线网络化的控制等，使此试验平台不断完善。

参考文献：

[1] 栾美艳. 采用虚拟测控软件 LabVIEW 实现控制系统的监控功能[D]. 大连：大连交通大学，2004：9-16.

[2] 王晋杰. 基于 LabVIEW 的 PC 与 PLC 实时监控的实现[J]. 武汉理工大学学报，2006，28(11)：53-55.

[3] 王 葵，董 罡，邢在奎. 基于 LabVIEW 虚拟仪器的数据采集和故障录波[J]. 电子测量与仪器学报，2004，18(4)：83-88.

[4] 吴异卉，王启志. 基于 LabVIEW 的模型参考自适应控制的实现[J]. 计算机技术与发展，2008，18(11)：180-182.

[5] 刘小刚. 基于 DSP 的无刷直流电机运动控制实验平台的研究与设计[D]. 西安：陕西科技大学，2008：40-43.

[6] 韩丰田. TMS320F2812x DSP 原理及应用技术[M]. 北京：清华大学出版社，2009.

[7] Kehtarnavaz N，Gope C. DSP System Design Using LabVIEW and Simulink：A Comparative Evaluation[C]// IEEE，ICASSP2006.[s.l.]：[s.n.]，2006：985-988.

[8] 曹军军，陈小勤，吴 超. TMS320F2812 型数字信号处理器与 PC 的串行通信[J]. 国外电子元器件，2005(8)：38-40.

[9] 戴 鹏，刘 剑，符 晓，等. 基于 TMS320F2812 与 LabVIEW 的串口通信[J]. 计算机工程，2009，35(4)：94-96.

[10] 韩丰田. TMS320F2812x DSP 原理及应用技术[M]. 北京：清华大学出版社，2009：71-97.

[11] 蒋本兵. 基于 LabVIEW 的 AMT 数据采集与分析系统的开发[D]. 合肥：合肥工业大学，2007：20-22.

[12] 王克峰，吴 森，曹永欣. NI FP-2010 与 DSP 串口通讯的研究[J]. 计算机技术与发展，2006，16(7)：227-229.

[13] 张 凯，周 陬，郭 栋. LabVIEW 虚拟仪器工程设计与开发[M]. 北京：国防工业出版社，2004：244-253.

(上接第 216 页)

优参数选择研究[J]. 统工程理论与实践，2006(10)：73-79.

[16] 甄 彤，范艳峰. 基于支持向量机的企业信用风险评估研究[J]. 微电子学与计算机，2006，23(5)：136-139.

[17] Chung Chih-Chung，Lin Chih-Jen. LIBSVM：a library for support vector machines [EB/OL]. [2010-06-01]. http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html.

[18] Hsu Chih-Wei，Chang Chih-Chung. A Practical Guide to Support Vector Classification. [EB/OL]. [2010-06-01]. http://www.csie.ntu.edu.tw/cjlin/~cjlin/papers/guide/guide.pdf.

[19] 王 波，郝艳友. 基于 SVM 的房贷信用评估的应用研究[J]. 计算机工程与设计，2008，29(19)：5110-5113.