文章编号: 1003- 207(2010) 05- 0028- 06

信用评估中的鲁棒赋权自适应 L_p 最小二乘支持向量机方法

刘京礼[1, 2], 李建平[2], 徐伟宣[2], 石 勇[3]

(1. 中国科学技术大学管理学院, 安徽 合肥 230026;
2 中国科学院科技政策与管理科学研究所, 北京 100190;
3 中国科学院虚拟经济与数据科学研究中心, 北京 100190)

摘 要: 消费者信用评估是金融风险管理和信用产业竞争的一个重要方面。信用评估数据中常带有噪声点, 并且其类别是不均衡的。最小二乘支持向量机是一个被广泛应用的分类模型, 其模型简单, 求解速度快, 但鲁棒性差。本文提出了一个鲁棒赋权自适应 L_p 最小二乘支持向量机模型, 能够适应信用评估样本数据库类别不均衡的特点, 可以有效处理信用评估数据中带有噪声点的问题。在仿真数据和三个信用数据库上的实证分析表明, 本文所提出的模型具有较好的鲁棒性和分类能力。

关键词: 信用评估; 鲁棒; 自适应; 最小二乘支持向量机
中图分类号: C93; T P3 文献标识码: A

1 引言

消费者信用风险评估, 即评估与个人借贷相关的风险, 是许多商业银行所面临的主要风险之一, 也是统计、人工智能和概率模型应用的一个重要领域。消费者信用风险评估问题实质上是一个分类问题, 针对这个问题, 研究人员提出了许多消费者信用评估的模型[1, 2], 包括金融专家方式、统计、数学规划、人工智能等方法。Suykens 等(1999) [3] 提出最小二乘支持向量机(LS- SVM) 模型之后, 该模型被广泛应用于消费者信用风险的评估[4, 5]。

在使用 LS- SVM 模型进行消费者信用风险评估时需要解决两个问题: 鲁棒性和样本类别的不均衡性。鲁棒性原是统计理论中分析数据时所使用的一个词, 主要是研究数据的较小摄动对统计模型的影响, 后来被广泛应用到控制理论中。本文中的鲁棒性是指使用样本数据所建立的模型的解对噪声点不敏感。在消费者信用风险评估数据中, 样本类别是非常不均衡的。假定样本中存在两个类别, 一种是能够及时归还贷款的好人, 一种是违约的不归还贷款的坏人。坏人这个类别的样本数量在总样本中所占的比例很小, 因而在使用 LS- SVM 建立分类模型时需要考虑这种不均衡性。

消费者信用风险评估的样本数据是通过各种渠道进行采集的, 由于测量或者人为的因素造成信用数据库中的数据带有噪声点。当样本数据中存在较多的噪声点或者误差不服从高斯正态分布的时候, 使用 LS- SVM 构造分类模型会导致模型的鲁棒性变差。为了解决 LS- SVM 模型鲁棒性差的问题, 文献[6, 7] 分别提出了两个采用不同方法赋权的最小二乘支持向量机模型。这两个模型都是给误差变量 ξ_i 增加一个系数因子 μ_i, 当样本中存在噪声点使得误差 ξ_i 的取值较大时, 相应的 μ_i 会较小, 从而使得噪声点所对应的误差项对目标函数的影响较小。这两种不同的赋权方式都可以增强 LS- SVM 模型解的鲁棒性, 但这两个赋权模型并没有考虑样本数据中类别的不均衡性问题。为了使所建立的分类模型能够适应不均衡的数据结构, 研究人员提出了许多自适应范数的 SVM 和 LS- SVM 模型[8- 13]。这些自适应范数的 SVM 和 LS- SVM 模型考虑了模型对样本数据结构的适应性问题, 但是并没有考虑当样本数据中存在噪声点或奇异值时模型解的鲁棒性问题。

收稿日期: 2009- 11- 7, 修改日期: 2010- 7- 12
基金项目: 国家自然科学基金资助项目(70531040, 70621001, 70921061)
作者简介: 刘京礼(1975-), 男(汉族), 山东胶州人, 中国科学技术大学管理学院, 中国科学院科技政策与管理科学研究所博士生, 研究方向: 数据挖掘.

为了使得在消费者信用数据库上所建立的模型适应样本类别不均衡的特性、消除噪声点的影响、简化模型的计算。我们以 LS- SVM 模型为基础, 建立一个鲁棒赋权自适应 L_p 最小二乘支持向量机模型(Robust Weighted Adaptive L_p - LS- SVM)。

2　模型描述

对消费者的信用风险进行评估可以帮助借贷者判别是否给予申请者贷款或者信用。这个问题可以应用分类技术来寻找合适的决策规则, 根据这个决策规则对样本进行分类。样本数据的特征主要有申请者的学历, 年龄, 住房情况, 职业等。本文我们考虑样本数据中存在两个类别的情况。对于给定的消费者样本 $\{x_i\}_{i=1}^N$, 其对应的类别标签是 y_i, $y_i \in \{+1, -1\}$, $i = 1, 2, \cdots, N$, 分别代表相应的消费者是好人和坏人。消费者信用评估就是根据给定的样本数据建立分类模型, 得到决策函数 $f(x)$, 然后根据 $f(x_i)$ 的值来预测一个新的消费者 x_i 对应的标签 y_i。

2.1　最小二乘支持向量机模型

给定训练数据集 $\{x_i, y_i\}_{i=1}^N$, 其中 $x_i \in \mathbf{R}^m$ 是第 i 个输入模式, $y_i \in \mathbf{R}$ 是对应的第 i 个输出模式, 样本数据的分布未知。LS- SVM 模型是如下形式的优化问题[14]:

$$\min_{\omega, b, \xi} J(\omega, \xi) = \frac{1}{2}\omega^T\omega + \frac{\gamma}{2}\xi^T\xi$$
$$s.t.\quad D(\omega^T\varphi(A) + eb) + \xi = e \qquad (1)$$

由对应的 KKT 最优条件可以得到一个线性方程组

$$\begin{pmatrix} 0 & e^T \\ e & \Omega + \frac{1}{\gamma}I \end{pmatrix}\begin{pmatrix} b \\ \alpha \end{pmatrix} = \begin{pmatrix} 0 \\ y \end{pmatrix} \qquad (2)$$

其中 $y = (y_1, y_2, \cdots, y_N)^T$ 为标签向量, $\Omega_{ij} = \varphi(x_i)^T\varphi(x_j)$, $i, j = 1, 2, \cdots, N$, 是核矩阵, A 是样本特征值对应的矩阵, D 为样本所对应的标签矩阵, 其对角元素为 + 1 或者 - 1。根据 Mercer 定理, 存在一个核函数 $\psi(\cdot)$, $\psi(x_i, x_j) = \varphi(x_i)^T\varphi(x_j)$, $i, j = 1, 2, \cdots, N$, 那么最后定义的分类决策函数就是

$$f(x) = sign\left[\sum_{i=1}^{N}\alpha_i y_i \psi(x, x_i) + b\right] \qquad (3)$$

2.2　鲁棒赋权自适应 L_p- LS- SVM 模型(RWAL_p - LS- SVM)

LS- SVM 模型的鲁棒性较差, 为了改进其鲁棒性, 我们采用赋权的方法, 对于那些在分类超平面的构造中起重要作用的点赋予较大的权重, 给予作用较小的点较小的权重。通过使用权重系数 μ_i, 我们期望可以降低样本数据中噪声和奇异值对模型分类能力的影响, 使得我们所得到的分类超平面具有鲁棒的特征。那么, 对于每一个输入数据的误差项, 都存在一个权重 μ_i。因为样本数据的分布是未知的, 我们无法使用概率值来计算样本点对应的权重, 因而采用启发式的方法来计算权重, 其计算公式是:

$$\mu_i = \frac{s_i - \min\limits_{i=1,\cdots,N} s_i}{\max\limits_{i=1,\cdots,N} s_i - \min\limits_{i=1,\cdots,N} s_i} \qquad (4)$$

其中, s_i 是通过已知的方法例如标准的最小二乘支持向量机模型所计算出的每一个样本数据所对应的分数值。即

$$s_i = \sum_{k=1}^{N}\alpha_k y_k \varphi(x_i)^T\varphi(x_k) + b \qquad (5)$$

为了更好的适应不同样本的数据结构, 降低噪声数据对模型的影响, 使模型有更好的推广能力, 我们采用一种数据驱动的建模方法, 也就是根据数据的结构来建立模型, 这种建模方法使得模型能够自适应的选择目标函数, 即采用 L_p 范数作为 LS- SVM 的目标函数。我们所提出的鲁棒赋权自适应 L_p - LS- SVM 模型是:

$$\min \left\| \begin{pmatrix} b \\ \alpha \end{pmatrix} \right\|_p$$
$$s.t. \begin{pmatrix} 0 & -Y^T \\ Y & Q + \gamma^{-1}I\mu \end{pmatrix}\begin{pmatrix} b \\ \alpha \end{pmatrix} = \begin{pmatrix} 0 \\ e \end{pmatrix} \qquad (6)$$

其中, $\Omega_{ij} = y_i y_j \varphi(x_i)^T\varphi(x_j)$, $Y = (y_1, y_2, \cdots, y_N)^T$, $\mu = (\mu_1, \mu_2, \cdots, \mu_N)$。

2.3　算法设计

首先, 我们来讨论一下本文目标函数中所使用的 L_p 范数中参数 p 的取值范围。

Mangasarian(1999)[15] 讨论了任意 p - 范数($p \in [1, \infty]$)的分离超平面, 其规划中的目标函数是极小化误分点到分类超平面的距离之和。其中, 采用 1- 范数的目标函数可以在多项式时间内通过求解 $2n$ 个线性规划得到解。对于任意的 p - 范数, 则可以通过转换变成极小化凸集上的一个凸函数与一个双线性函数的和。这实际上是一种近似求解算法, 在求解时需要增加参数, 设置合适的初始值, 这给规划的求解带来了一定的困难。对于任意两个范数 L_p 和 L_q, $p, q \in [1, \infty]$, $\frac{1}{p} + \frac{1}{q} = 1$, L_p 和 L_q 被称为是相互对偶的范数 Mangasarian(1999)[14]。因此, 如果在目标函数中所出现的范数 p 的取值较

大, 那么可以通过考虑其对偶范数替换相应的目标函数。Liu 等 (2007)[9] 在其自适应的 L_q 范数 SVM 模型中对范数 q 的取值做了分析。在 $q = 0$ 时, SVM 模型中的目标函数在原点是非凸、不连续的, 是 NP- 难问题, 不容易计算。而当 $q > 1$ 时, 假如所估计系数的值较大, 那么目标函数中的 ω 收缩到零的数量会增多。为了避免较大的参数值造成过大的偏差, p 的取值被限定在 $(0, 2]$。对于 $p \in [1, \infty]$, L_p 范数是单调不增的, 即对于任意的 $x \in \mathbf{R}^n$, $\|x\|_\infty \leqslant \|x\|_2 \leqslant \|x\|_1$。因此, 为了使得规划模型易于计算, 提高预测模型的准确率, 改进其推广能力, 我们所考虑的范数值 p 的取值范围是 $p \in [1, 2]$。

对于自适应的 L_p - LS- SVM 模型来说, 它包含三个参数, 核参数 σ^2, 惩罚参数 γ 和范数参数 p。为了更快的寻找最优参数的取值, 我们采用了演化策略算法[16]。演化算法采用个体选择, 进化, 交叉结合和适应度函数来模拟生物进化, 其算法包括遗传算法, 遗传规划, 演化策略, 演化规划等算法。演化算法的优势在于该算法对于适应度不需要过多的假设, 而且对于各种不同类型的问题都能够取得较好的效果。其具体执行的步骤可参见图 1[17]。

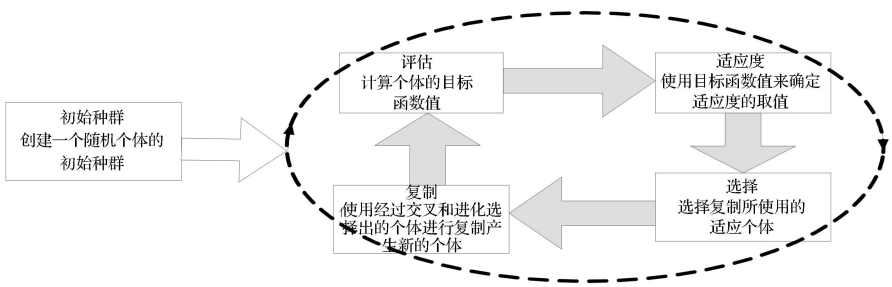

图 1 演化算法流程图

鲁棒赋权自适应 L_p - LS- SVM 模型中的权重系数 μ 可采用公式来计算, 需要首先求解标准的 LS- SVM 模型。在实验中我们使用 LS- SVMlab 1.5[18] 软件包来进行 LS- SVM 模型的求解, 分类准确率是采用 5- 折交叉验证的平均值。对于其余三个参数值, 则是采用演化算法来寻找最优参数值。给定权重系数 μ, 范数 p, $p \in [1, 2]$ 和 γ 以及 σ^2 的值, 规划模型的求解则可以使用 IRMS 算法[19] 来求解。

算法步骤具体如下:

该算法包含两部分, 一部分是样本权重的计算, 另一部分则是采用演化算法来计算最优参数值, 即对规划模型 进行求解。对于演化算法计算中参数值的设置, 我们参考了魏利伟(2008)[11] 使用演化算法计算自适应 L_q 范数最小二乘支持向量机模型的参数取值。

(1) 样本权重的计算:

1) 参数初值的设定: 对模型中惩罚参数 γ 以及核参数 σ^2 设置初值, 其取值范围是 $[-15, 15]$。范数 p 的取值设定为区间 $[1, 2]$。

2) 权重系数的计算: 首先使用 LS- SVMlab 1.5 软件包求解 LS- SVM 模型, 计算样本分数值, 然后根据公式计算出权重系数 μ。

(2) 规划模型的求解:

令 I 为个体空间, 假定每一次进化产生 μ 个个体, $\alpha \in I$ 为一个个体。α 中元素包括核参数 σ^2, 惩罚参数 γ, 范数参数 p 及适应度函数 $F: I \rightarrow R$, 本文中我们把分类的总准确率作为适应度函数。在第 t 代, 通过上一代个体的重组、突变及选择形成的个体为 $P(t) = \{\alpha_1(t), \alpha_2(t), \cdots, \alpha_\mu(t)\}$, 进化策略的选择采用 (μ, λ)- ES 方法。

那么, 求解规划模型的演化算法如下:

1) 初始化参数值: 初始代数 $t = 0$, 个体数目为 $\mu = 3$, 下一代个体的数量为 $\lambda = 7$。

初始个体为 $P(0) = \{\alpha_1(0), \alpha_2(0), \cdots, \alpha_n(0)\}$;

2) 循环: 如果进化的代数 $< l$ 或者新个体的分类总准确率小于给定的准确率, 则进行如下操作

个体重组: $P'(t) = recombination(P(t))$。

个体进化: $P''(t) = mutation(P(t))$。

计算个体的适应度: $F(P''(t))$, 即计算个体的分类总准确率。如果范数参数 $1 \leqslant p < 2$, 则采用 IRMS 算法来求解得到分类总准确率; 如果 $p = 2$, 则直接采用伪逆算法求解。

进行个体选择: 使用 (μ, λ)- ES 选择适应度大于上一代个体的新个体。

$t = t + 1$

循环终止

对于演化算法计算中参数值的设置, 我们参考了魏利伟(2008)[11] 使用演化算法计算自适应 L_q 范数最小二乘支持向量机模型的参数取值。其中, 初始自适应进化速率 $\tau = 0.577$, 适应度函数的初值设为 0.5。

3　实证结果

3.1　仿真数据集上的测试结果

仿真数据集为一个超立方体[9] 所产生的数据集, 其每一个点 x 是从超立方体 $[0, 1]^{20}$ 均匀产生的, 其对应的类别标签为 $sign(f(x))$, 其中 $f(x) = 2x_1 + 4x_2 + 4x_3 - 4.8$。为了检验模型去除噪声特征影响的能力, 我们假定每一个点向量 x 中的前三个特征向量是重要的, 其余 17 个是噪声特征向量。训练和测试数据的样本都是 400 个。其计算结果参见表 1。

表 1　在仿真数据集上进行分类的结果

模型	测试误差	特征数量
贝叶斯规则	0.2216	3
L_1 - SVM	0.2578	11.79
L_2 - SVM	0.2672	19.97
L_p - SVM	0.2415	5.28
RWA L_p - LS- SVM	0.2475	9.8

从表 1 的结果来看, 测试误差最小的是采用贝叶斯规则的分类模型, 其所选择的样本特征数量也是最少的。当噪声数量增加时, L_p - SVM 模型所选择的范数 p 值会逐渐变小。从实验结果看, RWA L_p - LS- SVM 模型所选择的特征数量比 L_p - SVM 模型增加不多的情况下, 其测试准确率并没有降低太少, 这说明我们所提出的模型还是具有鲁棒性的特征的。

3.2　信用数据库上的测试结果

我们使用了 3 个信用数据库, 包括 UCI 的两个信用数据库[20], 一个是澳大利亚的信用数据库(AUC), 一个是德国的信用数据库(GC), 还有一个是美国某商业银行的信用数据库(AMC)。其具体信息参见表 2。

表 2　模型测试所使用的信用数据库信息

数据库	类别数目	- 1 类别	+ 1 类别	样本数量	特征数量	样本属性
AUC	2	307	383	690	14	离散值
GC	2	300	700	1000	24	离散值
AMC	2	815	4185	5000	65	离散值

从样本的信息来看, 这三个数据库的样本类别数目都是 2。在 AUC 数据库中+ 1 类别和- 1 类别的样本数量相差不大, 而在 GC 数据库中, + 1 类别是- 1 类别的 2.3 倍。AMC 数据库中两个类别的样本数量差别最大, + 1 类别的样本数量是- 1 类别样本数量的 5.13 倍。在对模型的计算结果进行比较时, 我们采用了三类准确率作为对比的指标。这三类准确率分别是平均准确率 T(总分类精度)、敏感性 T^2(第二类准确率) 以及特异性 T^1(第一类准确率)。这三个准确率分别表示的是分类模型 5- 折交叉验证的平均分类准确率、模型在样本数量较少的类别- 1 上分类的准确率以及模型在样本数量较多的类别+ 1 上分类的准确率。

表 3　RWAL_p LS- SVM 在三个信用数据库上的结果

数据库	AUC	GC	AMC
T^2%	92.27	78.65	76.46
T^1%	86.61	74.56	72.73
T%	88.9	76.73	73.94
特征数量	7	12.4	24
支持向量数	245	350	568
p 范数值	1.3245	1.4152	1.4563

表 4 给出的是 RWAL_p - LS- SVM 模型与其他几个模型在 AUC 和 GC 两个数据库上的三类准确率。其他几个模型的计算结果取自[21, 22]。在 AUC 数据库上, RWAL_p - LS- SVM 的平均准确率在几个模型中是最高的。对于信用数据库来说, 其类别是不均衡的, 样本中的坏人的数量占少数, 因此, 在信用数据库上的分类模型的第二类分类准确率 T^2 是模型分类能力的一个重要指标。从 T^2 的值来看, RWAL_p - LS- SVM 在几个模型中是最高的。对于 GC 数据库, RWAL_p - LS- SVM 的 T^2 值也是最高的, 但 T^1 的值低于 See5 和基于遗传算法的 SVM 模型, 其平均准确率仅次于最高的 GA- SVM。因此, 与传统的几个分类模型相比, 我们所提出的模型还是有优势的。

表 4　RWA L_p - LS- SVM 与其他模型在 AUC 和 GC 数据库上的比较

数据集 / 准确率 / 模型	AUC 690× 14			GC 1000× 20		
	T^1%	T^2%	T%	T^1%	T^2%	T%
LDA	80.68	92.19	85.8	72.57	71.33	72.2
See5	87.99	84.69	86.52	84	44.67	72.2
SVM^{light}	18.03	90.65	44.83	77	42	66.5
MCCQP	87	85.52	86.38	74.38	72	73.5
GA- based SVM	84.72	92.18	88.1	89.6	76.62	85.6
RWA L_p - LS- SVM	86.61	92.27	88.9	74.56	78.65	76.73

表5 是本文提出的三个模型与魏利伟(2008)[11]所提出的 2 个自适应模型的分类准确率比较结果。RWAL_p – LS– SVM 在 AUC 数据库上的敏感度值要高于另外两个模型, 所选择的特征数量在三个模型中是最少的, 其平均分类准确率则与其他两个模型相差不大。在 AMC 数据库上, 三类准确率最高的是 L_p – LS– SVM。在 GC 数据库上, RWAL_p – LS– SVM 的敏感度要高于 L_p – LS– SVM, 选择的特征数量最少, 但其平均准确率要低于另外两个模型。从模型所选择的特征数量来看, RWAL_p – LS– SVM 选择的特这数量较少, 具有稀疏性的特征。

表 5 **RWA L_p – LS– SVM** 与其他自适应模型的比较

模型	指标	AMC	AUC	GC
L_p - MKMCP	T 2%	86. 19	90. 27	87. 13
	T 1%	73. 93	88. 39	73. 5
	T%	73. 28	89. 01	78. 92
	特征数量	11. 9	8. 6	18. 7
	p 范数值	1. 5769	1. 3725	1. 3376
L_p - LS- SVM	T 2%	86. 3	91. 9	72. 15
	T 1%	78. 49	90. 46	85. 63
	T%	81. 89	91. 3	77. 15
	特征数量	15. 7	7. 32	20. 2
	p 范数值	1. 6769	1. 2852	1. 4386
RWA L_p - LS- SVM	T 2%	76. 46	92. 27	78. 65
	T 1%	72. 73	86. 61	74. 56
	T%	73. 94	88. 9	76. 73
	特征数量	24	7	12. 4
	p 范数值	1. 4563	1. 3245	1. 4152

4 结语

本文针对信用评估数据的不均衡性和常带有噪声点, 要求所建立的分类模型具有鲁棒性和适应样本数据结构的能力的特点, 设计了一种鲁棒赋权自适应 L_p 最小二乘支持向量机模型, 并给出了具体算法。通过在仿真数据集和三个消费者信用风险评估数据库上的测试表明: 本文所提出的模型具有较好的鲁棒性, 能够适应信用数据库不均衡类别的特点, 可以作为消费者信用风险评价的一个有效的备选模型。后续我们将进一步利用本文的基本思路, 探讨如何用多目标最优化模型替换支持向量机的表达方式, 力求寻找更广泛的基于最优化的高鲁棒性和高准确性算法, 更好解决信用风险评价问题。

参考文献:

[1] Baesens, B , Van Gestel, T., et al Benchmarking state– of– the– art classification algorithms for credit scoring [J]. Journal of the Operational Research Society, 2003, 54(6): 627– 635.

[2] 李建平, 徐伟宣, 刘京礼, 石勇. 消费者信用评估中的支持向量机方法研究[J]. 系统工程, 2004, (10): 35– 39.

[3] Suykens, J. A. K., Vandewalle, J.. Least squares support vector machine classifiers [J]. Neural Processing Letters, 1999, 9(3): 293– 300.

[4] Yu, L A., Wang, S. Y., et al A modified least squares support vector machine classifier with application to credit risk analysis [J]. International Journal of Information Technology and Decision Making, 2009, 8(4): 697– 710.

[5] Li, J., Chen, Z., et al Feature selection via least squares support feature machine [J]. International Journal of Information Technology and Decision Making, 2007, 6(4): 671– 686.

[6] Suykens, J A. K., De Brabanter, J., et al Weighted least squares support vector machines: robustness and sparse approximation [J]. Neurocomputing, 2002, 48(1– 4): 85– 105.

[7] Valyon, J , Horváth, G.. A weighted generalized LS – SVM [J]. Periodica Polytechnica Serial: Electrical Engineering, 2003, 47(3– 4): 229– 251.

[8] Pedroso, J P., Murata, N.. Support vector machines with different norms: Motivation, formulations and results [J]. Pattern Recogniton Lettres, 2001, 22(12): 1263– 1272.

[9] Liu, Y. F., Zhang, H. H., et al Support vector machines with adaptive L– q penalty [J]. Computational Statistics and Data Analysis, 2007, 51(12): 6380– 6394.

[10] Sun, D., Li, J., et al Credit risk evaluation: Support vector machines with adaptive Lq penalty [J]. Journal of Southeast Univeristy (English Edition), 2008, 24: 33– 36.

[11] 魏利伟. 多目标规划数据挖掘分类算法研究及应用[D]. 中国科学院科技政策与管理科学研究所, 2008.

[12] Huang, K., Zheng, D., et al Arbitrary norm support vector machines [J]. Neural Computation, 2009, 21(2): 560– 582.

[13] Liu, Z Q., Lin, S. L., et al Sparse support vector machines with L– p penalty for biomarker identification [J]. IEEE– ACM Transactions on Computational Biology and Bioinformatics, 2010, 7(1): 100– 107.

[14] Mangasaian, O. L., Musicant, D. R.. Lagrangian support vector machines [J]. Journal of Machine

Learning Research, 2001, 1(3): 161-177.

[15] Mangasarian, O. L.. Arbitrary-norm separating plane [J]. Operations Research Letters, 1999, 24(1-2): 15-23.

[16] Holland, J.. Adaptation in Natural and Artificial Systems [M]. Ann Arbor: The University of Michigan Press, 1975.

[17] Weise, T., Achler, S., et al. Evolving Classifiers-Evolutionary Algorithms in Data Mining [OL]. http://www.it-weise.de/documents/files/WAGVZ2007DM, 2007.

[18] Suykens, J. A. K., Gestel, T. V., et al. Least Squares Support Vector Machines [M]: World Scientific Publication Company, 2002.

[19] Brito, A. E.. Iterative Adaptive Extrapolation Applied to SAR Image Formation and Sinusoidal Recovery[D]. Department of Electrical and Computer Engineering, 2001.

[20] Asuncion, A., Newman, D. J.. UCI Machine Learning Repository [Z]. University of California, School of Information and Computer Science, 2007.

[21] Kou, G., Peng, Y., et al. A new multi-criteria convex quadratic programming model for credit analysis [C]. In V. N. Alexandrov et al (Eds): ICCS 2006, LNCS 3994, Springer-Verlag Berlin, 2006: 476-484.

[22] Braga, P. L., Oliveira, A. L. I., et al. A GA-based feature selection and parameters optimization for support vector regression applied to software effort estimation [C]. Proceedings of the 2008 ACM symposium on Applied computing, Fortaleza, Ceara, Brazil, ACM, 2008.

[23] Shi, Y.. Multiple criteria optimization-based data mining methods and applications: A systematic survey [J]. Knowledge and Information Systems, 2009: 1-23.

[24] 云庆夏. 进化算法[M]. 北京: 冶金工业出版社, 2000.

[25] Thomas, L. C.. Consumer finance: Challenges for operational research [J]. Journal of the Operational Research Society, 2010, 61(1): 41-52.

A Robust Weighted Adaptive L_p LS-SVM Method for Credit Risk Assessment

LIU Jing-li[1,2], LI Jian-ping[2], XU Wei-xuan[2], SHI Yong[3]

(1. School of Management, University of Science and Technology of China, Hefei 230026, China;
2. Institute of Policy and Management, Chinese Academy of Sciences, Beijing 100190, China;
3. Research Center on Fictitious Economy and Data Science, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: Consumer credit risk assessment is an important aspect of financial risk management and credit industry competition. Credit database often contains noisy data, which makes the data uncertain. Least squares support vector machines, a widely used binary classification model, is simple and easy to be applied. In this paper, we propose a robust weighted adaptive L_p least squares support vector machines, which can deal with unbalanced data sets and noisy data. The empirical test on simulation and three credit data sets have shown the model has outstanding robustness and generalization ability.

Key words: credit risk assessment; robust; adaptive; least squares support vector machines