文章编号:1001 - 9081(2009)10 - 2751 - 04

基于二次 Renyi 熵的非迭代最小二乘支持向量机预测模型

赵冠华[1,2]
(1. 天津大学 管理学院, 天津 300072; 2. 山东财政学院 会计学院,济南 250014)
(sczgh2006@ sina. com)

摘 要:将二次 Renyi 熵应用于企业财务困境预测,提出了一种基于二次 Renyi 熵的最小二乘支持向量机(LS-SVM)模型。通过将该模型与传统的 LS-SVM 模型、标准 SVM 模型以及与二项 Logistic 回归模型、BP 神经网络(BP-ANN)的分析比较,表明了该模型无论是训练样本的数量还是运算时间,都显著优于其他模型,且有较好的稳定性。实证分析表明,将二次 Renyi 熵引入企业财务困境预测领域是成功的,同时,通过对原始输入变量进行显著性检验、因子分析处理,减少了输入变量个数,预测正确率达到了 88%,说明因子分析法是有效的。

关键词:二次 Renyi 熵;最小二乘支持向量机;标准支持向量机;非迭代;因子分析;财务困境预测
中图分类号: TP181 **文献标志码**:A

Prediction model of noniterative least squares SVM based on quadratic Renyi-entropy

ZHAO Guan-hua[1,2]
(1. *School of Management, Tianjin University, Tianjin* 300072, *China*;
2. *School of Accounting, Shandong University of Finance, Jinan Shandong* 25001, *China*)

Abstract: A learning algorithm of noniterative Least Squares Support Vector Machine (LS-SVM) based on quadratic Renyi-entropy was proposed in the article by using quadratic Renyi-entropy in financial distress prediction. By comparing the model of LS-SVM based on quadratic Renyi-entropy with traditional LS-SVM, standard SVM, binomial Logistic regression model and Back Propagation Artificial Neural Network (BP-ANN), this paper concluded that either the number of training samples or the computing time, the model of noniterative LS-SVM based on quadratic Renyi-entropy is remarkably better than the others, as well as the stability. Indicated by demonstration analysis, the model of noniterative LS-SVM based on quadratic Renyi-entropy is successful in financial distress prediction. Meanwhile, although the number of input variable has been reduced by conspicuity test and gene analysis, the accuracy rate of the prognosis still reached 88%. In a word, the factor analysis method has been successfully proved in the article.

Key words: quadratic Renyi-entropy; Least Squares Support Vector Machine (LS-SVM); standard SVM; noniterative; factor analysis; financial distress prediction

0 引言

支持向量机(Support Vector Machine, SVM)[1]是对传统统计学的重要发展和补充,已成为当前机器学习理论的一个突出研究热点,由于其出色的学习性能,尤其是它的良好的泛化能力,已经引起了人们对这一领域的极大关注[2-3]。

标准 SVM 最终归结为求解一个二次规划(Quadratic Programming, QP)问题,但当训练样本数量增大时,二次规划问题将面临着维数灾难[3]。由于内存的限制,大规模问题的求解无法进行。在支持向量机出现之后的短短几年时间里,许多研究人员致力于其理论与算法的研究。文献[4]作者提出了一种分解算法,把标准支持向量机的二次规划问题分解成一系列小规模的子 QP 问题,使得每个子问题都能容易求解,该算法有效地解决了大规模支持向量机问题。在文献[4]的基础上,Joachims 从选择训练集的角度出发提出了具体的实现方案,并用软件实现了该分解算法。文献[5]作者提出了序列最小优化(Sequential Minimal Optimization, SMO)的分解算法,把标准 SVM 的 QP 问题分解成可以解析求解的最小 QP 问题,即每个训练集仅由两个样本组成。文献[6]作者提出的最小二乘支持向量机(Least Squares Support Vector Machines, LS-SVM)把不等式约束换成等式约束,从而使得支持向量机的求解由 QP 问题转化为一个线性方程组,极大地提高了求解效率,同时也降低了求解难度。文献[7 - 10]作者进一步研究了 LS-SVM,并针对 LS-SVM 算法破坏支持向量稀疏性的缺点,提出了稀疏近似的策略。当支持向量谱分布均匀时,该方法不易对支持向量进行取舍。文献[11 - 12]作者提出了固定尺度的 LS-SVM 算法并将该算法用于电力负荷预测。

以上文献在用支持向量机求解分类问题时,是将所有的样本都参与训练,这些样本都作为支持向量,再加上 LS-SVM 的支持向量缺少稀疏性这一缺点,使得当训练样本较大时,求解矩阵的逆在时间和空间上的开销都很大。本文将信息论中熵的概念引入 LS-SVM 算法中,独立推导出了熵的表达式,通过实验给出了适合企业财务困境预测的二次 Renyi 熵核函数。实证结果表明,无论是参与训练的样本数还是训练时间都比传统 LS-SVM 有了明显减少,且模型的稳定性也明显好

收稿日期:2009 - 04 - 28。 **基金项目**:国家自然科学基金资助项目(70840018);山东省科技攻关计划资助项目(2008GG30009005);山东省软科学研究计划资助项目(2008RKA223)。

作者简介:赵冠华(1962 -),男,江苏徐州人,副教授,博士研究生,主要研究方向:统计学习理论、数据挖掘、财务决策支持系统。

于其他算法。

1 最小二乘支持向量机模型

LS-SVM 主要是根据优化问题目标函数的不同,推出一系列不同的等式约束。该模型的一般性描述为:

给定 l 个训练样本的集合 $\{(x_i,y_i),i=1,2,\cdots,l\}$,第 i 个输入数据 $x_i\in\mathbf{R}^n$,输出数据 $y_i\in\{+1,-1\}$,y_i 为二分变量。LS-SVM 的目标就是要构造一个如下形式的分类器:

$$f(x)=\mathrm{sgn}(\boldsymbol{w}^{\mathrm{T}}\varphi(x)+b) \tag{1}$$

使得样本 x 能够被 $f(x)$ 正确分类。LS-SVM 算法就是要求解下面的优化问题[13]:

$$\min\frac{1}{2}\|\boldsymbol{w}\|^2+\frac{1}{2}C\sum_{i=1}^{l}\xi_i^2 \tag{2}$$

$$\text{s.t. } y_i(\boldsymbol{w}^{\mathrm{T}}\varphi(x_i)+b)=1-\xi_i;\ i=1,2,\cdots,l \tag{3}$$

其中:$\boldsymbol{w}=(w_1,w_2,\cdots,w_l)^{\mathrm{T}}$ 为权重向量,它是一个垂直于分类超平面($\boldsymbol{w}^{\mathrm{T}}x_i+b=0$)的向量(图1),$b$ 为常数,这里代表阈值。ξ_i 是大于零的松弛变量,它是样本点离该类的距离,ξ_i 用来度量样本点 x_i 违反约束 $y_i(\boldsymbol{w}^{\mathrm{T}}\varphi(x_i)+b)=1-\xi_i$ 的程度。$\varphi(x_i)$ 是映射函数,当求解的问题线性不可分时,通过该函数将样本的输入空间映射到特征空间,所求解的问题将变成线性可分了。C 是惩罚因子,C 越大,对错误的惩罚越重。C 对于支持向量机最终的判定效果有很大的影响,是支持向量机一个非常重要的参数。C 的大小直接影响到支持向量机的泛化(预测)性能。

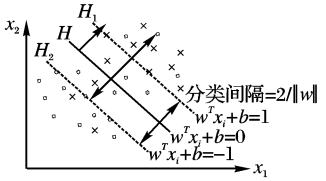

图1 分类支持向量机示意图

式(2)、(3)是一个带有等式约束的二次规划问题,为了便于求解,可以引入 Lagrange 函数,它的二次规划问题的最优解为下面 Lagrange 函数的鞍点:

$$L(\boldsymbol{w},b,\xi,\alpha)=\frac{1}{2}\|\boldsymbol{w}\|^2+\frac{1}{2}C\sum_{i=1}^{l}\xi_i^2-\sum_{i=1}^{l}\alpha_i[y_i(\boldsymbol{w}^{\mathrm{T}}\varphi(x_i)+b)+\xi_i-1] \tag{4}$$

其中 α_i 为 Lagrange 乘子,由于是等式约束,其值可以是正的也可以是负的,最优化条件为将式(4)分别对 $\boldsymbol{w}$、b、ξ_i、α_i 求偏导数并令其等于零:

$$\frac{\partial L}{\partial w}=\boldsymbol{w}-\sum_{i=1}^{l}\alpha_i y_i\varphi(x_i)=0 \tag{5}$$

$$\frac{\partial L}{\partial b}=-\sum_{i=1}^{l}\alpha_i y_i=0 \tag{6}$$

$$\frac{\partial L}{\partial \xi_i}=C\sum_{i=1}^{l}\xi_i-\sum_{i=1}^{l}\alpha_i=0 \tag{7}$$

$$\frac{\partial L}{\partial \alpha_i}=-\sum_{i=1}^{l}[y_i(\boldsymbol{w}^{\mathrm{T}}\varphi(x_i)+b)+\xi_i-1]=0 \tag{8}$$

同时考虑到 $y_i\in\{+1,-1\}$ 可得到:

$$\begin{cases}\boldsymbol{w}=\sum_{i=1}^{l}\alpha_i y_i\varphi(x_i)\\ \sum_{i=1}^{l}\alpha_i y_i=0\\ \alpha_i=C\xi_i\\ y_i[\boldsymbol{w}^{\mathrm{T}}\varphi(x_i)+b]+\xi_i-1=0\end{cases} \tag{9}$$

式(9)可以写成如下的方程组:

$$\begin{bmatrix}\boldsymbol{I}&0&0&-\boldsymbol{Z}^{\mathrm{T}}\\0&0&0&-\boldsymbol{Y}^{\mathrm{T}}\\0&0&CI&-\boldsymbol{I}\\\boldsymbol{Z}&\boldsymbol{Y}&\boldsymbol{I}&0\end{bmatrix}\begin{bmatrix}\boldsymbol{w}\\b\\\boldsymbol{\xi}\\\boldsymbol{\alpha}\end{bmatrix}=\begin{bmatrix}0\\0\\0\\\overline{1}\end{bmatrix} \tag{10}$$

这里,$\boldsymbol{\xi}=(\xi_1,\xi_2,\cdots,\xi_l)^{\mathrm{T}}$,$\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\cdots,\alpha_l)^{\mathrm{T}}$,$\overline{1}=(1,1,\cdots,1)^{\mathrm{T}}$,$\boldsymbol{Z}=[\varphi(x_1)y_1,\varphi(x_2)y_2,\cdots,\varphi(x_l)y_l]^{\mathrm{T}}$,$\boldsymbol{Y}=(y_1,y_2,\cdots,y_l)^{\mathrm{T}}$,$\boldsymbol{I}$ 是单位矩阵。消去 $\boldsymbol{\xi}$ 和 $\boldsymbol{w}$,再利用 Mercer 条件:

$$\boldsymbol{\Omega}_{ij}=y_iy_j\varphi^{\mathrm{T}}(x_i)\varphi(x_j)=y_iy_jK(x_i,x_j) \tag{11}$$

则方程组(10)转化为:

$$\begin{bmatrix}0&-\boldsymbol{Y}^{\mathrm{T}}\\\boldsymbol{Y}&\boldsymbol{\Omega}+C^{-1}I\end{bmatrix}\begin{bmatrix}b\\\boldsymbol{\alpha}\end{bmatrix}=\begin{bmatrix}0\\\overline{1}\end{bmatrix} \tag{12}$$

设 $\boldsymbol{A}=\boldsymbol{\Omega}+C^{-1}I$,由于 $\boldsymbol{A}$ 是一个对称的半正定矩阵,因此 $\boldsymbol{A}^{-1}$ 存在。求解线性方程组(12),得到的解如下:

$$b=\frac{\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{A}^{-1}\overline{1}}{\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{Y}}\boldsymbol{\alpha}=\boldsymbol{A}^{-1}(\overline{1}-\boldsymbol{Y}b) \tag{13}$$

因此,所求的分类函数为:

$$f(x)=\mathrm{sgn}\left(\sum_{i=1}^{l}\alpha_i y_iK(x_i,x)+b\right) \tag{14}$$

2 基于二次 Renyi 熵的非迭代 LS-SVM 模型

2.1 熵及二次 Renyi 熵

熵是对热力学系统中随机程度的一种度量,这个概念最早用来研究热力学中气体[14]。后来 Shannon 创建了信息论,第一次用熵的概念来研究信息在信道中的传输。在信息论中,信息是由一个所谓的信息源输出的[15],若用 $I(A)=-\log p(A)$(其中 $p(A)$ 表示事件 A 发生的概率)来度量事件 A 给出的信息量,称为事件 A 的自信息量。设某信息源输出几个相互独立的消息 $x_i(i=1,2,\cdots,N)$,当每个消息出现的概率为 $p_i(i=1,2,\cdots,N)$ 时,则可用 $H_i=-p_i\log p_i$ 来度量一个消息所给出的平均信息量,则整个事件的平均信息量为:

$$H=-\sum_{i=1}^{N}p_i\log p_i \tag{15}$$

H 称为信息熵或 Shannon 熵,Shannon 熵是对信源输出信息的随机性的度量,这种度量是基于所有可能输出状态的概率 p_i。当所有状态出现的概率都相同时,熵最大,这时对应系统的随机程度越高。相反,当某个状态 i 出现的概率为 $p_i=1$ 时,熵为0,这时对应的系统为确定性系统。Renyi 进一步扩展了 Sannon 熵的概念,并将概率密度函数 $p(x)$ 的 $\alpha(\alpha\geqslant 0,\alpha\neq 1)$ 阶 Renyi 熵定义为:

$$H_{R\alpha}=\frac{1}{1-\alpha}\log\int p^{\alpha}(x)\,\mathrm{d}x \tag{16}$$

本文采用二次 Renyi 熵,它的表达式为:

$$H_{R2}=-\log\int p^2(x)\,\mathrm{d}x \tag{17}$$

其中,$\int p^2(x)\,\mathrm{d}x$ 可以用式(18)估计:

$$\int p^2(x)\,\mathrm{d}x\approx\int\hat{p}^2(x)\,\mathrm{d}x=\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}K(x_i,x_j) \tag{18}$$

因此,二次 Renyi 熵可以用下面的表达式近似:

$$H_{R2}\approx-\log\left(\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}K(x_i,x_j)\right) \tag{19}$$

$K(x_i,x_j)$ 称为核函数,对于特定的应用场合,它的表达式会有所不同,一般通过实验的方法确定。

2.2 基于二次 Renyi 熵的非迭代 LS-SVM 模型及算法

在传统 LS-SVM 的模型中,支持向量数目的确定缺少相

应的理论指导,而本文提出的算法在选取支持向量时,采用二次 Renyi 熵作为评价标准,选择使得二次 Renyi 熵改变最大的样本加入到训练集中。

首先构造一个初始样本训练集 $w = \{(x_i, y_i), i = 1,2,\cdots,n\}$,利用式(13)、(14)计算前 n(这里 $n = 8$)个样本的 b、α、$f(x)$。令 $b_n = b, \alpha_n = \alpha, A_n^{-1} = A^{-1}$,则每增加一对样本后,式(13)、(14)变成:

$$b_{n+1} = \frac{Y_{n+1}^{\mathrm{T}} A_{n+1}^{-1} \overline{1}_{n+1}}{Y_{n+1}^{\mathrm{T}} A_{n+1}^{-1} Y_{n+1}} \tag{20}$$

$$\alpha_{n+1} = A_{n+1}^{-1}(\overline{1}_{n+1} - Y_{n+1} b_{n+1}) \tag{21}$$

$$f_{n+1}(x) = \mathrm{sgn}\left(\sum_{i=1}^{n+1} \alpha_{n+1,i} y_i K(x_i, x) + b_{n+1}\right) \tag{22}$$

利用式(20 ~ 22)计算 b_{n+1}、α_{n+1}、$f_{n+1}(x)$ 以及误判率,若误判率没有达到事先设定的值,再从剩下的训练样本中抽出一对样本,计算它们的二次 Renyi 熵,找出最大的 Renyi 熵,构建新的训练集。重复以上操作直到误判率小于预先设定的值为止。算法的实现步骤为。

1)预先设定一个误判率 ε,$\varepsilon = \left|\frac{f(x) - y}{y}\right|$,$\varepsilon$ 可取普通 LS-SVM 算法计算出的误判率值,这里假定为 E。设初始训练集为 $w = \{(x_i, y_i), i = 1,2,\cdots,n\}$,根据 $\boldsymbol{A} = \boldsymbol{\Omega} + C^{-1}I$ 解析地计算出它的逆矩阵 $\boldsymbol{A}_n^{-1}$,根据式(20) ~ (22),算出参数 α_{n+1}、b_{n+1}、$f_{n+1}(x)$;

2)根据总训练样本数 m(这里 $m = 102$)及初始训练样本数 n(这里 $n = 8$),计算出剩余训练样本数 $p(p = m - n)$,并假定每次取出的一对样本构成的集合为 s_i,且 $s_i \notin w$;

3)当 $\varepsilon \geqslant E$ 时,执行 4);

4)对于循环变量 $i = 1,2,\cdots,p$,执行 5);

5)将二次 Renyi 熵最大的一对样本集合 s_i 加入到初始训练集 w 中,构成临时训练集 w_i,即 $w_i = w \cup \{s_i\}$;

6)根据式(19)计算熵 $H_{R2}(w_i)$;

7)找出 w_i 的二次 Renyi 熵 $H_{R2}(w_i)$ 中最大者,记作 $j = \max_i\{H_{R2}(w_i)\}$;

8)将二次 Renyi 熵最大的一对样本集合 s_j 加入到初始训练集 w 中,构成新的临时训练集,即 $w = w \cup \{s_j\}$;

9)根据 w 中的样本数据,利用式(20 ~ 22)计算出 b_{n+1}、α_{n+1}、$f_{n+1}(x)$,并依据已知分类函数值(ST 公司为 +1,非 ST 公司为 -1)计算出误判率 ε;

10)判断误判率 $\varepsilon > E$ 是否成立,若成立,转 4);否则,转 11);

11)将最后一次计算出的 $f_{n+1}(x)$ 值作为最终的预测值,结束程序运行。

该算法已在 Matlab R2007 上编程实现。

3　实验验证

3.1　输入变量与响应变量的选取

根据上市公司的年报披露制度,上市公司公布当年年报的截止日期为下一年度的 4 月 30 日,故上市公司 $t-1$ 年的年报与在第 t 年是否被特别处理几乎是同时发生的,为了避免夸大预测的正确率,本文采用上市公司 $t-2$ 年的财务数据来预测公司在第 t 年是否因财务困境被特别处理。

输入变量选取了六个一级财务指标,包括:短期偿债能力、长期偿债能力、营运能力、盈利能力、成长能力以及现金流量,共计 23 个二级财务指标;同时还选取了四个一级非财务指标,包括:股本结构、公司治理、资产规模以及地域环境,计 8 个二级非财务指标,最终确定输入变量为 31 个。

响应变量取为第 t 年是否被特别处理,如是取 +1,否则取 -1。

3.2　样本来源和预处理

本文采用随机方式从沪、深两市抽取 200 家 A 股上市公司 2002 ~ 2007 六年的数据构成初始研究样本,其中 100 家是 ST 公司,100 家是非 ST 公司,它们构成配对样本。剔除数据不全的样本,最终确定的研究样本数为 152 家,其中,102 家构成训练样本集,50 家构成测试样本集。所有样本数据均取自深圳国泰安 CSMAR 数据库。

做实证分析前,必须要对这 31 个初始输入变量进行预处理。首先,用 K-S 检验法对初始输入变量进行正态分布检验,因样本总体上不符合联合正态分布,无法使用成对 T 检验,而只能采用非参数检验法;其次,用 Wilcoxon 符号秩法对样本进行显著性检验,结果发现,两组样本有 19 个变量在 5% 水平上存在显著性差异;最后,再对这些变量进行因子分析,以消除变量间存在的多重共线性。经因子分析,最终确定了 8 个公共因子,这些公共因子就是模型的输入变量。

3.3　预测模型的构造及计算

为了便于比较,本文除了给出基于二次 Renyi 熵的非迭代 LS-SVM 模型、标准 SVM 模型以及传统 LS-SVM 模型外,还给出了二项 Logistic 回归模型以及 BP 神经网络(Back Propagation Artificial Neural Network,BP-ANN)模型。

1)二项 Logistic 回归模型:$P = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}$

其中:$\boldsymbol{x} = (x_1, x_2, \cdots x_l)^{\mathrm{T}}$ 为输入的样本数据,$\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots \beta_t)$ 为一组与 x 对应的回归系数,α 为模型的截距。如果 $p > 0.5$,该企业为 ST 型;如果 $p < 0.5$,该企业为正常型。

2)BP-ANN 模型:

$$W_{jh}(t+1) = W_{jh}(t) - \eta \frac{\partial E}{\partial W_{jh}} + \alpha(W_{jh}(t) - W_{jh}(t-1))$$

$$W_{hi}(t+1) = W_{hi}(t) - \eta \frac{\partial E}{\partial W_{hi}} + \alpha(W_{hi}(t) - W_{hi}(t-1))$$

其中:α 为势态因子,η 为学习效率,t 为迭代次数,E 为定义误差,W_{jh} 为输入层节点与隐含层节点之间的连接权值;W_{hi} 为隐含层节点与输出层之间的连接权值。经试算,本文的各参数设定为:$\eta = 0.15$,$\alpha = 0.8$,输入层节点数为 8,隐含层节点数为 5,输出层节点数为 1。

3)标准 SVM 模型:核函数分别取高斯径向基核函数 $k(x, x_i) = \exp\{-\sigma^2 |x - x_i|^2\}$、多项式核函数 $K(x_i, x_j) = (1 + x_i^{\mathrm{T}} x_j)^d$,采用交叉验证方法确定 $\sigma^2 = 1$,$d = 1$,参数 b 取样本计算值的平均值,这里 $b = 0.0693$。

4)传统 LS-SVM 模型:核函数的选取与标准 SVM 相似,分别取高斯径向基核函数和多项式核函数。这里,惩罚因子 $C = 5$,$\sigma^2 = 3.33$,$b = 0.1438$,多项式核中的 $d = 1$。

5)基于 Renyi 熵的 LS-SVM 模型:高斯径向基核函数不适合本文的财务困境预测,核函数取多项式核 $K(x_i, x_j) = (1 + x_i^T x_j)^d$,$d = 1$。用式(19)计算训练样本的熵,初始样本数取 8。

3.4　训练与测试

以上训练模型均在相同的训练集上进行训练,在相同的测试集上进行测试。模型误将 ST 样本判为正常样本,称为犯第 I 类错误;将正常样本判为 ST 样本,称为犯第 II 类错误。样本对训练集和测试集的总误判率取为第 I 类和第 II 类误判率的平均值。

3.5　实验结果与分析

为了便于比较,将 7 种模型的实证对比结果列于表 1。从表 1 可以看出,在测试样本的误判率方面,无论是标准 SVM,还是 LS-SVM 和基于 Renyi 熵的 LS-SVM,其误判率都显

著小于二项 Logistic 回归模型和 BP-ANN(图2)。但前3种模型中,以标准 SVM(多项式核)最差,误判率达到38%,说明了标准 SVM 模型预测的不稳定性。而 LS-SVM 和基于 Renyi 熵的 LS-SVM 误判率均为12%,预测效果最好。表1还表明,虽然本文提出的基于 Renyi 熵的 LS-SVM 模型没有提高预测的正确率,但训练的样本数和训练时间都明显减小。在两类误判率方面,二项 Logistic 回归模型、BP-ANN 以及标准 SVM-多项式核犯第Ⅰ类错误的几率高于犯第Ⅱ类错误的几率,SVM-高斯核、LS-SVM 模型相反,即犯第Ⅰ类错误的几率低于犯第Ⅱ类错误的几率,只有基于 Renyi 熵的 LS-SVM 模型,犯两类错误的几率相同,均为12%,这说明本文提出的基于 Renyi 熵的 LS-SVM 模型不仅大大减少了参与训练的样本个数和时间,而且算法本身具有较好的稳定性。7种模型测试样本的总误判率变化趋势见图2。

表1 不同预测模型各项目比较

模型编号	模型	训练样本误判率/%			测试样本误判率/%			训练样本数	训练时间(CPU s)
		Ⅰ类错误	Ⅱ类错误	总误判率	Ⅰ类错误	Ⅱ类错误	总误判率		
1	二项 Logistic 回归模型	13.0	10.0	11.5	23.4	22.8	23.1	102	26.4
2	BP-ANN	13.0	9.0	11.0	18.7	16.7	17.7	102	38.2
3	标准 SVM(多项式核)	0.0	11.8	5.9	44.0	32.0	38.0	102	16.1
4	标准 SVM(高斯核)	0.0	5.9	2.9	12.0	20.0	16.0	102	15.5
5	LS-SVM(多项式核)	0.0	3.9	2.0	12.0	12.0	12.0	102	13.2
	LS-SVM(高斯核)	5.9	7.8	6.9	8.0	16.0	12.0	102	13.8
6	Renyi 熵 LS-SVM(多项式核)	0.0	3.9	2.0	12.0	12.0	12.0	56	8.3
7	Renyi 熵 LS-SVM(高斯核)	×	×	×	×	×	×	×	×

注:"×"说明高斯核对基于"Renyi 熵的 LS-SVM"算法无效。

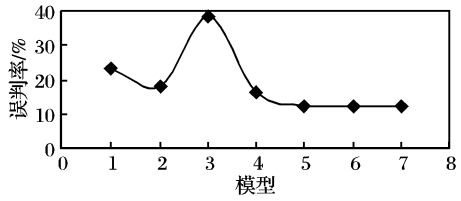

图2 不同预测模型测试样本总误判率

在训练样本的误判率方面,由于直接用参与训练的样本去预测,所以,总的误判率明显低于测试样本。但仍以本文提出的基于 Renyi 熵的 LS-SVM 模型误判率最低,仅为2%,这说明该模型无论是拟合能力,还是泛化能力都最好。7种模型训练样本的总误判率变化趋势见图3。

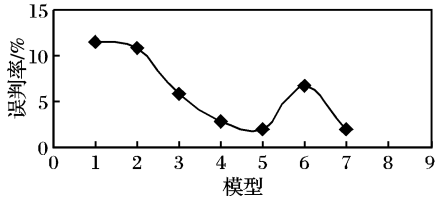

图3 不同预测模型训练样本总误判率

在训练样本数量方面,由于前6种模型是将所有样本都参与训练,样本数为102,而本文提出的算法仅需56个样本就达到了原来的预测正确率,节省了近一半的训练样本。再从训练时间上来看,以 BP-ANN 最长,达到38.2CPU s,二项 Logistic 回归模型次之,而本文的算法仅用了8.3 CPU s,是这7种算法中最短的。7种模型训练时间的变化趋势见图4。

该算法的总误判率变化情况见图5,从图5可以看出,由于每次都是以 Renyi 熵最大的原则加入一对训练样本,所以,随着加入样本数的增加,预测误差逐渐减小,最后趋于一个极限值12%,此时的训练样本数为56,节省了近一半的样本。

4 结语

通过对基于 Renyi 熵的非迭代 LS-SVM 模型与标准 SVM、LS-SVM 模型以及二项 Logistic 回归模型、BP-ANN 的分析比较,可以看出,无论是模型的拟合能力,还是泛化(预测)能力,都以本文提出的模型最佳,而且训练的样本个数和训练时间也都显著小于其他模型,又有较好的稳定性。说明了作者将 Renyi 熵引入财务困境预测领域是成功的。通过实验还发现,高斯核函数不能作为二次 Renyi 熵的表达式,这与其他应用领域有差异。究其原因,可能是高斯核函数 $k(x,x_i) = \exp\{-\sigma^2 \mid x - x_i \mid^2\}$ 中有负的指数项存在,而 ST 和正常上市公司财务指标的差异在数量级上较小,导致计算出的熵差异甚微。实验表明,用多项式核 $K(x_i,x_j) = (1 + x_i^{\mathrm{T}} x_j)^d, d = 1$ 作为二次 Renyi 熵的表达式对本文是适合的。此外,本文的研究还表明,对原始输入变量通过显著性检验、因子分析处理,将31个输入变量用8个公共因子表示,虽然减少了解释变量个数,但预测正确率还是达到了88%,说明了这些公共因子包含了大部分的财务信息,因子分析法对本文是有效的。

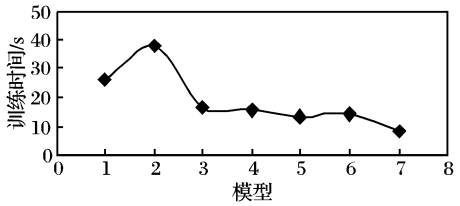

图4 不同预测模型训练时间(CPU s)

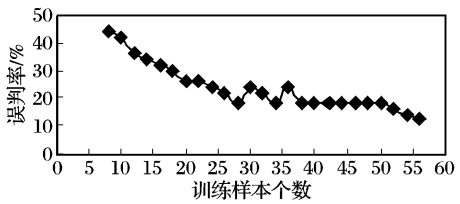

图5 基于 Renyi 熵的 LS-SVM 预测误差趋势

参考文献:

[1] COMES C, VAPNIK V N. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273 - 297.

[2] BORER B, GUYON I, VAPNIK V N. A training algorithm for optimal margin classifiers[C]// Proceedings of the Fifth Annual Workshop on Computational Learning Theory. New York: ACM Press, 1992: 144 - 152.

(下转第2757页)

3) 计算个体适应度函数的值 $F(T)$;

4) 如果种群的适应度函数足够大,或者 T 达到我们预设的终止代数值(25 代) 则转到 8);

5) $T = T + 1$;

6) 应用选择算子从 $P(T-1)$ 中选择新的 $P(T)$;

7) 对 $P(T)$ 进行交叉变异操作之后转到 3);

8) 得出最佳的惩罚因子 C 与核函数参数的组合,并利用优化后的支持向量机对训练样本进行训练,得到全局最优分类面。

3　实验结果与结论

本实验的实验数据来自 CERNET 共享的 2005 年 6 月份的电子邮件(http://www.ccert.edu.cn/spam/sa/datasets.htm),其中垃圾邮件 3 290 份,非垃圾邮件 2 510 份,将数据集分为五个近似相等的子集,四个用作训练集,一个作为测试集,以分类正确率作为评测标准。为方便比较,我们用台湾林智仁等人开发的 LIBSVM 对相同的数据集进行了实验,LIBSVM 包中有一自带的优化程序 grid.py,同样得到优化的参数组合。但是分类结果与本文提出的 GASVM 相比不甚理想。实验结果如表 2 所示。

表 2　两种分类结果对比

使用的分类器	C	r	分类正确率/%
LIBSVM	2048	0.5	81.67
GASVM	114	0.5	89.67

从表 2 可以看出,利用遗传算法优化的支持向量机取得了较好的分类效果,说明本文提出的 GASVM 算法是可行的。

4　结语

本文利用支持向量机对垃圾邮件进行分类,并且利用遗传算法对支持向量机的参数组合进行了优化,获取了最优的参数组合,从而取得了较好的分类正确率。今后的主要工作集中在:综合优化特征提取部分,使得分类的正确率能够获得更大的提高;考虑平衡数据与非平衡数据对分类的影响,并寻求合适的方法来解决这个问题。

参考文献:

[1] CERVANTES J, LI XIAO-OU, YU WEN. SVM classification for large data sets by considering models of classes distribution[C]// Proceedings of the 2007 Sixth Mexican International Conference on Artificial Intelligence, Special Session. Washington, DC: IEEE Computer Society, 2007: 51 - 60.

[2] NHUNG N P, PHUONG T M. An efficient method for filtering image-based spam[C]// Proceedings of the 2007 IEEE International Conference on Research, Innovation and Vision for the Future. [S.l.]: IEEE Press, 2007: 96 - 102.

[3] KIM D S, NGUYEN H-N. Genetic algorithm to improve SVM based network intrusion detection system[C]// Proceedings of the 19th International Conference on Advanced Information Networking and Applications. Washington, DC: IEEE Computer Society, 2005: 155 - 158.

[4] DRUCKER H, WU DONG-HUI, VAONICK V N. Support vector machines for spam categorization [J]. IEEE Transactions on Neural Networks, 1999, 10(5): 1048 - 54.

[5] VAPNIK V N. An overview of statistical learning theory [J]. IEEE Transactions on Neural Network, 1999, 10(5): 988 - 999.

[6] 刘伍颖, 王挺. 一种多过滤器集成学习垃圾邮件过滤方法[C]// 全国信息检索与内容安全学术会议论文集. 苏州: [出版者不详], 2007.

[7] 李钢, 王蔚, 张胜. 支持向量机在脑电信号分类中的应用[J]. 计算机应用, 2006, 26(6): 1431 - 1433.

[8] 樊兴华, 孙茂松. 一种高性能的两类中文文本分类方法[J]. 计算机学报, 2006, 29(1): 124 - 131.

[9] 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.

[10] 王清祥, 广凯, 潘金贵. 基于支持向量机的邮件过滤[J]. 计算机科学, 2007, 34(9): 93 - 95.

(上接第 2754 页)

[3] SCHOLKOPF B, BURGER C, VAPNIK V N. Extracting support data for a given task[C]// Proceedings of First International Conference on Knowledge Discovery and DataMining, [S. l.]: AAAI Press, 1995: 262 - 267.

[4] OSUNA E, FREUND R, GIROSI F. An improved training algorithm for support vector machines[C]// Proceedings of the 1997 IEEE Workshop on Neural Networks and Signal Processing. Amelia Island: IEEE Press, 1997: 276 - 285.

[5] PLATT J C. Fast training of support vector machines using sequential minimal optimization[C]// Advances in Kernel Methods: Support Vector Learning. Cambridge: MIT Press, 1999: 185 - 208.

[6] SUYKENS J A K, VANDEWALLE J. Least squares support vector machine classifiers[J]. Neural Processing Letter, 1999, 9(3): 293 - 300.

[7] SUYKENS J A K, LUKAS L, WANDEWALLE J. Sparse approximation using least squares support vector machines[C]// ISCAS: Proceeding of the IEEE International Symposium on Circuits and Systems. [S.l.]: IEEE Press, 2000: 757 - 760.

[8] LI YONG-MIN, GONG SHAO-GANG, SHERRAH J, *et al.* Support vector machine based multi-view face detection and recognition [J]. Image and Vision Computing, 2004, 22(5): 413 - 427.

[9] ROMDHANI S, TORN P, SCHOLKOPF B, *et al.* Efficient face detection by a cascaded support-vector machine expansion[J]. Royal Society of London Series A-Mathematical Physical and Engineering Sciences, 2004, 13(4): 3283 - 3297.

[10] SHIH P C, LIU C J. Face detection using discriminating feature analysis and support vector machine[J]. Pattern Recognition, 2006, 39(2): 260 - 276.

[11] ESPINOZA M, SUYKENS J A K, MOOR B D. Load forecasting using fixed-size least squares support vector machines[C]// Computational Intelligence and Bioinspired Systems, LNCS 3512. Berlin: Springer, 2005: 1018 - 1026.

[12] ESPINOZA M, SUYKENS J A K, MOOR B D. Fixed-size least squares support vector machines: A large scale application in electrical load forecasting[J]. Computational Management Science, 2006, 3(2): 113 - 129.

[13] SUYKENS J A K, LUKAS L, WANDEWALLE J. Sparse approximation using least squares support vector machines [C]// SCAS'2000: Proceeding of the IEEE International Symposium on Circuits and Systems. [S.l.]: IEEE Press, 2000: 757 - 760.

[14] VINGAA S, JONAS S, ALMEIDA J A. Renyi continuous entropy of DNA sequences[J]. Journal of Theoretical Biology, 2004, 231(3): 377 - 388.

[15] SHANNON C E. A mathematical heory of communication[J]. The Bell System Technical Journal, 1948, 27(3): 379 - 423; 623 - 656.