文章编号：1000－8152(2006)02－0315－04

具分段损失函数的支持向量机回归及在投资决策中的应用

胡根生，邓飞其
（华南理工大学 自动化科学与工程学院，广东 广州 510640）

摘要：支持向量机回归模型的性能与所选用的损失函数有很大关系. 本文提出一种具分段损失函数的支持向量机回归模型，其分段损失函数对落在不同区间的误差项采用不同的惩罚函数形式，并将该模型应用于投资决策问题中，估计收益率向量的联合概率密度函数和最优投资组合. 仿真实验表明，其性能要优于一般的支持向量回归方法.

关键词：支持向量机回归；损失函数；投资决策
中图分类号：TP18　　**文献标识码**：A

Support vector regression with piecewise loss function and its application in investment decision

HU Gen-sheng, DENG Fei-qi
(College of Automation Science and Engineering, South China University of Technology, Guangzhou Guangdong 510640, China)

Abstract: The selection of loss function plays an important role to the performance of support vector regression (SVR) model. This paper proposes an SVR model with piecewise loss function, which gives different penalty values for deviation in different regions. The SVR model is applied in the problem of investment decision to estimate the joint probability density function of yield vector and the optimal portfolio. Experiments show that its performance is superior to that of standard SVR method.

Key words: support vector regression; loss function; investment decision

1 引言(Introduction)

支持向量机(SVM)是20世纪90年代由Vapnik等人提出的一种基于统计学习理论的机器学习算法[1]. 它不存在局部极小问题，其计算复杂性与输入样本的维数无关，具有很强的泛化能力和抗噪声能力，在小样本学习下，比其他算法具有更明显的优势，因而近年来，支持向量机已被广泛地应用于分类问题和回归问题.

在支持向量机算法模型中，必须选择目标函数中恰当的损失函数. 标准的支持向量机模型采用Vapnik定义的ε-不敏感损失函数[1]. Suykens等人用平方损失函数代替ε-不敏感损失函数(LS－SVM)[2]. Laplace损失函数用误差项的绝对值作为惩罚函数，Huber损失函数把Laplace损失函数和二次损失函数结合了起来[3]，但这几种损失函数不再具有ε-不敏感损失函数解的稀疏性. 本文结合ε-不敏感损失函数和Huber损失函数的优点，采用多分段损失函数的形式，在不同的区间段内，损失函数具有不同的形式，这样既提高了模型的抗噪声能力，又提高了估计的精度.

在投资决策问题中，决策者需要解决不确定经济系统中最优资产组合问题. 现有的数理金融理论都是假定未来收益概率分布为已知，或者用经验分布来代替实际分布，因而与实际问题存在较大的误差. 在估计未来收益概率分布的各种理论中，由于上述的多种性能，支持向量机无疑具有明显的优势. 本文将一种具分段损失函数的支持向量机回归模型应用于投资决策问题中，获得最优的投资组合.

2 具分段损失函数的支持向量机回归模型(SVR with piecewise loss function)

给定独立同分布的数据集$\{(x_i, y_i), i = 1, \cdots, N\}$，这里$x_i \in X \subset \mathbb{R}^n$，$y_i \in Y \subset \mathbb{R}$是取自未知概率

收稿日期：2004－12－20；收修改稿日期：2005－07－12.
基金项目：国家自然科学基金资助项目(60374023)；广东省自然科学基金资助项目(011629).

分布 $P(X, Y)$ 的样本. 支持向量回归的做法是将数据 $x_i(i = 1, \cdots, N)$ 通过下面的非线性映射变换到线性可分的高维特征空间 H:

$$\Phi: x_i \to \Phi(x_i). \tag{1}$$

根据 Mercer 定理[4],能构造正定核函数 K 作为高维特征空间元素的内积,如果该核函数满足 Mercer 条件,即

$$K(x, x_i) = \langle \Phi(x), \Phi(x_i) \rangle, \text{对任意 } x, x_i \in X. \tag{2}$$

支持向量回归的目标是构造线性回归函数 $f \in F$,这里 $F = \{f | f: X \to Y\}$ 是一个函数类,使得结构风险 $R_{\text{reg}}(f)$ 最小. 这里

$$R_{\text{reg}}(f) = R_{\text{emp}}(f) + C\Omega(f). \tag{3}$$

其中:$R_{\text{emp}}(f)$ 是经验损失函数,$\Omega(f)$ 是正则项.

总假定 H 是一个再生核希尔伯特空间. 根据支持向量机理论,最小化式(3)等价于下面的约束优化问题:

$$\begin{cases} \min \|f\|^2 + C\sum\limits_{i=1}^{N} l(\xi_i) \\ \text{s. t. } |y_i - f(x_i)| \leqslant \varepsilon + \xi_i, \ \xi_i \geqslant 0, \\ \quad i = 1, \cdots, N. \end{cases} \tag{4}$$

求解具有多分段损失函数的支持向量机回归,可以利用 F. Perez-Cruz 等人提出的变权迭代算法[5],采用迭代的方法获得回归函数的权系数.

定义 Lagrange 乘子

$$L = \|f\|^2 + C\sum_{i=1}^{N} l(\xi_i) - \sum_{i=1}^{N} \alpha_i [(\varepsilon + \xi_i)^2 - (y_i - f(x_i))^2] - \sum_{i=1}^{N} \mu_i \xi_i. \tag{5}$$

根据表示定理[6],$f = \sum\limits_{j=1}^{N} \gamma_j K(x_j, \cdot)$,因而由 KKT 条件得到

$$\Phi\gamma + \Phi DK\gamma = \Phi D y. \tag{6}$$

这里

$$\Phi = (K(x_1, \cdot), K(x_2, \cdot), \cdots, K(x_N, \cdot)),$$

$$\gamma = (\gamma_1, \gamma_2, \cdots, \gamma_N)^{\mathrm{T}}, y = (y_1, y_2, \cdots, y_N)^{\mathrm{T}},$$

$$D = \text{diag}(\alpha_1, \alpha_2, \cdots, \alpha_N), K = K(x_i, x_j)_{N \times N}.$$

在式(6)的两边同时乘以 (ΦD) 的广义逆 $(\Phi D)^+$,可得 $(D^+ + K)\gamma = y$,由于 K 是正定的,即对任意非零列向量 V,有 $V^{\mathrm{T}}KV > 0$,所以 $V^{\mathrm{T}}(D^+ + K)V > 0$,因而 $(D^+ + K)$ 是正定矩阵,因而有

$$\gamma = (D^+ + K)^{-1} y. \tag{7}$$

式(7)中 D 里面的 α_i 的取值,与式(4)中损失函数的形式有关. 本文结合 ε-不敏感损失函数和 Huber 损失函数的优点,采用多分段损失函数的形式,在不同的区间段内,损失函数具有不同的形式,其定义为

$$l(e_i) = \begin{cases} 0, & e_i \in [0, \beta_0\varepsilon], \\ \vdots & \vdots \\ (e_i - \varepsilon)\prod\limits_{j=l}^{p} \dfrac{(e_i - \varepsilon)}{(\beta_j - 1)\varepsilon}, & e_i \in (\beta_{l-1}\varepsilon, \beta_l\varepsilon], \\ \vdots & \vdots \\ e_i - \varepsilon, & e_i \in (\beta_p\varepsilon, +\infty). \end{cases} \tag{8}$$

这里:$e_i = |y_i - f(x_i)|$,$1 = \beta_0 \leqslant \beta_1 \leqslant \cdots \leqslant \beta_p < +\infty$. 当 $\beta_p = 1$ 时,上述损失函数等价于 ε-不敏感损失函数;当 $\beta_{p-1} = 1$ 时,损失函数式(8)等价于下述函数形式

$$l(e_i) = \begin{cases} 0, & e_i \in [0, \varepsilon], \\ \dfrac{(e_i - \varepsilon)^2}{(\beta - 1)\varepsilon}, & e_i \in (\varepsilon, \beta\varepsilon], \beta > 1. \\ (e_i - \varepsilon), & e_i \in (\beta\varepsilon, +\infty), \end{cases} \tag{9}$$

损失函数式(8)对落在越小区间值的误差项,其惩罚值也越小,且小于相应的 ε-不敏感损失函数的惩罚值,因而容许有较小的误差,提高了它的抗噪声能力.

利用 KKT 条件,可以得到

$$\xi_i = \max(0, (e_i - \varepsilon)],$$

$$\alpha_i = \begin{cases} 0, & \xi_i = 0, \\ \vdots & \vdots \\ \dfrac{C(p-l+2)}{2e_i}\prod\limits_{j=l}^{p} \dfrac{\xi_i}{(\beta_j - 1)\varepsilon}, & \xi_i \in ((\beta_{l-1} - 1)\varepsilon, (\beta_l - 1)\varepsilon], \\ \vdots & \vdots \\ C/2e_i & \xi_i \in ((\beta_p - 1)\varepsilon, +\infty). \end{cases} \tag{10}$$

综上所述,计算这种具有多分段损失函数的支持向量机回归的步骤为

1) 给定模型参数 C, ε, β_i,值,设置初始值 γ, e_i;

2) 对于每个 $i = 1, \cdots, N$,利用式(10)计算相应的 α_i 值;

3) 计算 $\gamma = (D^+ + K)^{-1} y$, $f(x_i) = \sum\limits_{j=1}^{N} \gamma_j K(x_i, x_j)$, $e_i = |y_i - f(x_i)|$,如果所有的 $e_i \leqslant \varepsilon (i = 1, \cdots, N)$,则停止计算,否则返回 2).

3　基于支持向量机回归的投资决策分析(Analysis of investment decision based on SVR)

3.1　投资决策问题中收益率向量的密度函数估计(Density function estimation of yield vector in investment decision problem)

假设市场上仅有 n 种风险资产,其收益率向量记为 $X = (X^1, \cdots, X^n)$. 记其相应的联合概率分布函数和联合概率密度函数分别为 $F(x) = F(x^1, \cdots, x^n)$ 和 $f(x) = f(x^1, \cdots, x^n)$,取其独立同分布的样本集 $x_j^i, i = 1, \cdots, n, j = 1, \cdots, N$, 用经验分布函数 $F_N(x)$ 近似代替 $F(x)$,这里

$$F_N(x) = F_N(x^1, \cdots, x^n) = \frac{1}{N}\sum_{j=1}^{N}\prod_{i=1}^{n}\theta(x^i - x_j^i), \tag{11}$$

则经验分布函数 $F_N(x)$ 相应的密度函数为

$$f_N(x) = f_N(x^1, \cdots, x^n) = \frac{\partial^n F_N(x^1, \cdots, x^n)}{\partial x^1 \cdots \partial x^n}. \tag{12}$$

记 $x_j = (x_j^1, \cdots, x_j^n)$,将 $f_N(x_i)$ 替换式(4)中的 y_i,利用第 2 节介绍的方法求解,得到下述表示形式的回归函数

$$f(x) = \sum_{j=1}^{N}\gamma_j K(x, x_j). \tag{13}$$

其中的核函数满足第 2 节开始的假设条件,例如我们可以用高斯径向基函数核

$$K(x, x_j) = \exp(-\|x - x_j\|^2/2\sigma^2). \tag{14}$$

3.2　均值、方差投资组合(Mean-variance portfolio)

利用式(13)表示形式的回归函数,可以求出此 n 种风险资产的期望收益率 $\mathrm{E}(X)$ 及其协方差矩阵 Σ. 设投资者投资此 n 种风险资产的资产组合向量为 $w = (w_1, \cdots, w_n)^{\mathrm{T}}$,投资者的效用函数为均方效用函数 $U(\mathrm{E}(X)w, w^{\mathrm{T}}\Sigma w)$,则均值、方差、投资组合可以表示成以下的二次规划问题[7]

$$\begin{cases}\min \dfrac{1}{2}w^{\mathrm{T}}\Sigma w \\ \text{s. t. } 1^{\mathrm{T}}w = 1,\ \mathrm{E}(X)w = \mu.\end{cases} \tag{15}$$

如果 Σ 为非退化的,$\mathrm{E}(X) \neq k1^{\mathrm{T}}$,应用拉格朗日乘子法,得到最优投资组合的表达式为

$$w^* = \Sigma^{-1}(\lambda_1 1 + \lambda_2 \mathrm{E}(X)^{\mathrm{T}}). \tag{16}$$

其中

$$\lambda_1 = (c - \mu b)/\Delta,\ \lambda_2 = (\mu a - b)/\Delta,\ a = 1^{\mathrm{T}}\Sigma^{-1}1,$$
$$b = 1^{\mathrm{T}}\Sigma^{-1}E(X)^{\mathrm{T}},\ c = \mathrm{E}(X)\Sigma^{-1}\mathrm{E}(X)^{\mathrm{T}},\ \Delta = ac - b^2.$$

3.3　直接基于样本的投资组合(Portfolio based on samples directly)

应用以往的样本数据,利用支持向量方法,可以直接建立样本与投资组合之间函数关系的估计,具体步骤如下:

1) 获取训练样本.

计算 t 时刻为止的样本均值 $\mathrm{E}(X_t)$ 和样本协方差 $\Sigma_t, t = 1, 2, \cdots$,利用式(16)(15)获得 t 时刻的投资组合的最小风险 r_t.

2) 计算回归函数权系数.

以 (t, x_t) 作为输入数据,r_t 作为输出数据,求解式(4)的支持向量回归问题.

3) 预测 $t + 1$ 时刻应采用的最优投资组合.

将 $t + 1$ 时刻的数据 $(t + 1, x_{t+1})$ 代入式(13),获得 $t + 1$ 时刻的估计风险 r_{t+1},因而

$$w_{t+1}^{*\mathrm{T}} = \frac{2r_{t+1}}{\mu}\mathrm{E}(X_{t+1})\Sigma_{t+1}^{-1}. \tag{17}$$

4　实验(Experiments)

本节对上面介绍的方法进行仿真实验,分析其性能. 其中的核函数采用式(14)的高斯径向基函数核,损失函数采用式(9)所定义的分段函数.

取数据 $S = \{(x_i, y_i) \in X \times Y \subset \mathbb{R}^7 \times \mathbb{R}, i = 1, \cdots, 240\}$,其中的七维输入数据来自区间 $(-1, 1)$ 的随机数,输出数据为七维的正态分布,该分布的均值为 $\mathrm{E}(X) = [-0.4, -0.3, -0.1, 0.2, 0.3, 0.5, 0.7]$,协方差矩阵为单位矩阵.

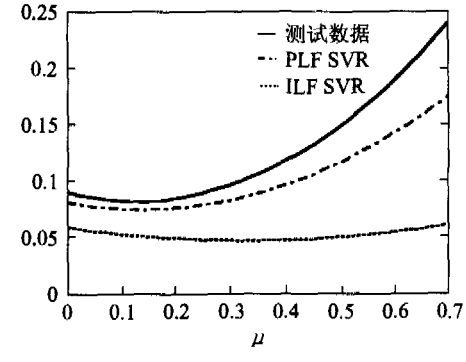

图 1　对于不同的期望收益 μ,用不同的支持向量机回归方法计算的最小风险

Fig. 1　Minimum risk computed by using different SVR method for different expected yield μ.

从中任取 200 组数据作为训练数据,其余的 40 组数据作为测试数据,分别用本文第 2 节介绍的分段损失函数支持向量机回归(PLF-SVR)算法和标准的 ε 不敏感损失函数支持向量机回归(ε-ILF-SVR)算法对训练数据进行训练($c = 1, \sigma = 1, \varepsilon = 0.00001, \beta = 2$),获得式(13)表示形式的回归

函数,并用训练的结果对测试数据进行测试,计算测试结果的平均值如表1所示,它们的协方差矩阵如表2所示,该结果对于不同的期望收益μ,用式(15)计算的最小风险r如图1所示.

表1 用不同的支持向量机回归方法所获得的回归函数的平均值

Table 1 Mean value of regression function obtained by using different SVR methods

PLF-SVR	-0.3359	-0.2485	-0.1464	0.1465	0.2592	0.5387	0.7030
ε-ILF-SVR	-0.2780	-0.2640	-0.0798	0.0535	0.1405	0.5910	0.6434

表2 用不同的支持向量机回归方法所获得的回归函数的协方差矩阵

Table 2 Covariance matrix of regression function obtained by using different SVR method

PLF-SVR							ε-ILF-SVR						
1.0448	0.0234	0.0015	-0.0111	0.0034	-0.0028	0.1045	0.8212	0.1634	0.0683	-0.2294	-0.0440	-0.1747	-0.1567
0.0234	0.7308	-0.1669	0.0571	-0.0272	0.0349	0.0777	0.1634	0.4390	-0.2048	0.1147	-0.1068	0.1654	0.0525
0.0015	-0.1669	0.9309	-0.0010	0.0895	0.0763	0.0634	0.0683	-0.2048	0.7587	0.0433	0.0644	0.1449	-0.0104
-0.0111	0.0571	-0.0010	0.8767	-0.0316	-0.0205	-0.0357	-0.2294	0.1147	0.0433	0.6359	-0.0836	-0.0949	0.0226
0.0034	-0.0272	0.0895	-0.0316	0.9515	-0.0502	-0.0678	-0.0440	-0.1068	0.0644	-0.0836	0.8193	0.0840	0.0686
-0.0028	0.0349	0.0763	-0.0205	-0.0502	0.9502	-0.0411	-0.1747	0.1654	0.1449	-0.0949	0.0840	0.7001	-0.0403
0.1045	0.0777	0.0634	-0.0357	-0.0678	-0.0411	0.7782	-0.1567	0.0525	-0.0104	0.0226	0.0686	-0.0403	0.5397

5 结论(Conclusion)

本文讨论了一种具分段损失函数的支持向量机回归模型,并利用变权迭代算法对此模型进行求解,最后将其应用于投资决策问题中,估计收益率向量的概率密度函数和最优的投资组合.仿真实验表明,估计的结果与实际的结果还是相当吻合的,其性能要优于使用标准的ε-不敏感支持向量回归算法.依靠增加样本数的方法,可以进一步缩小它们之间的差距,但这会带来计算机内存不足问题,造成计算速度急速下降,因此大样本下更为有效的支持向量机回归方法有待进一步研究.

参考文献(References):

[1] VAPNIK V. *The Nature of Statistical Learning Theory* [M]. New York: Springer-Verlag, 1995.

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

[3] HUBER P J. *Robust Statistics* [M]. New York: Wiley, 1981.

[4] SCHÖLKOPT B, SMOLA A J. *Learning with Kernels* [M]. Cambridge, MA: The MIT Press, 2002.

[5] PEREZ-CRUZ F, CAMPS G, SORIA E, et al. Multi-dimensional function approximation and regression estimation [C]// *Proc of the 12th Int Conf on Artificial Neural Networks*. Berlin: Springer-Verlag, 2002: 757 - 762.

[6] KIMELDORF G S, WAHBA G. Some results on Tchebycheffian spline functions [J]. *J of Mathematical Analysis and Applications*, 1971, 33(1): 82 - 95.

[7] LEIPPOLD M, TROJANI F, VANINI P. A geometric approach to multiperiod mean variance optimization of assets and liabilities [J]. *J of Economic Dynamics and Control*, 2004, 28(6): 1079 - 1113.

作者简介:

胡根生 (1971—),男,华南理工大学自动化科学与工程学院博士研究生,主要研究方向为支持向量机理论及应用,E-mail: hugs2906@sina.com;

邓飞其 (1962—),男,教授,博士生导师,主要研究方向为大系统、随机系统的控制理论、系统工程、支持向量机及数理金融理论,E-mail: aufqdeng@scut.edu.cn.