文章编号: 1001-0920(2009)01-0125-04

基于无约束优化的非线性支持向量回归

张军峰, 胡寿松
(南京航空航天大学 自动化学院, 南京 210016)

摘 要: 提出利用牛顿法以及共轭梯度法解决非线性支持向量回归学习问题,不仅可以加速模型选择的过程,而且能够提高训练速度. 将该方法应用于煤气炉数据集建模以及 Mackey-Glass 混沌时间序列预测,仿真结果表明了该方法的有效性.

关键词: 支持向量回归; 无约束优化; 牛顿法; 共轭梯度法
中图分类号: TP18 **文献标识码**: A

Nonlinear SVR based on unconstrained optimization

ZHANG Jun-feng, HU Shou-song
(College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China. Correspondent: ZHANG Jun-feng, E-mail: wufeng7919@163.com)

Abstract: A learning strategy based on Newton method and conjugate gradient method is proposed in this paper to solve the nonlinear support vector regression (SVR) training problem, which is able to accelerate not only the model selection process but also the training speed. By applying it into gas furnace data set modeling and Mackey-Glass chaotic time series prediction, the simulation results indicate the effectiveness of the proposed learning strategy.

Key words: Support vector regression; Unconstrained optimization; Newton method; Conjugate gradient method

1 引 言

Vapnik[1]创立的支持向量机(SVM),遵循了结构风险最小化原理,可将非线性问题转化为线性问题进而得到全局最优解. 近年来在模式识别、系统建模和时间序列预测等领域得到了广泛的应用.

对偶原理可以方便地解决约束条件,相应的优化问题可由内积的形式表示,并可引入核函数,因此,目前主要利用对偶原理设计 SVM 学习算法[2]. 事实上,不利用对偶原理,同样可以训练 SVM. Keerthi 和 DeCoste[3] 基于无约束优化训练线性 SVM; Chapelle[4] 将这种方法推广到非线性领域.

然而,上述工作均局限于支持向量分类(SVC). 本文将其推广到支持向量回归(SVR)领域,着重阐述利用牛顿法以及共轭梯度法实现非线性 SVR 的训练,并通过仿真实验说明该方法在模型选择与训练速度方面的有效性与优越性.

2 非线性 SVR 对偶优化

假设训练集为 $\{(x_1, y_1), \cdots, (x_n, y_n)\} \in \mathbf{X} \times \mathbf{R}$, $\mathbf{X} \subset \mathbf{R}^d$,其中 $\mathbf{X}$ 表示输入样本空间. 在特征空间 $\mathbf{J}$ 中,构造如下线性回归函数:

$$f(x) = w^{\mathrm{T}} \Phi(x) + b. \tag{1}$$

其中: $\Phi: x \mapsto \Phi(x) \in \mathbf{J}$, $w \in \mathbf{J}$.

考虑二次 ε 不敏感损失函数作用,并引入松弛变量,式(1)中的 w, b 可通过如下优化问题求解:

$$\min_{w,\xi,\xi^*} w^{\mathrm{T}} w + C \sum_{i=1}^{n} (\xi_i + \xi_i^*). \tag{2}$$

$$\text{s.t. } y_i - w^{\mathrm{T}} \Phi(x) - b \le \varepsilon + \xi_i, \ \xi_i \ge 0;$$

$$w^{\mathrm{T}} \Phi(x) + b - y_i \le \varepsilon + \xi_i^*, \ \xi_i^* \ge 0.$$

利用对偶原理,上述优化问题可转化为

$$\max_{\alpha,\alpha^*} -\sum_{i,j=1}^{n} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\left(k(x_i, x_j) + \delta_{i,j}/C\right) - 2\varepsilon \sum_{i=1}^{n} (\alpha_i + \alpha_i^*) + 2\sum_{i=1}^{n} y_i (\alpha_i - \alpha_i^*), \tag{3}$$

$$\text{s.t. } \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) = 0, \ \alpha_i, \alpha_i^* \ge 0.$$

其中: $k(x_i, x_j)$ 为核函数, $\delta_{i,j}$ 表示 Kronecker δ.

收稿日期: 2007-12-25; **修回日期**: 2008-03-12.
基金项目: 国家自然科学基金重点项目(60234010); 航空科学基金项目(05E52031).
作者简介: 张军峰(1979—),男,江苏建湖人,博士生,从事机器学习、故障预报的研究; 胡寿松(1937—),男,浙江慈溪人,教授,博士生导师,从事复杂系统的可靠控制等研究.

对于优化问题(3)的解,回归函数(1)可表示为

$$f(x) = \sum_{i}^{nsv}(\alpha_i - \alpha_i^*)\,k(x_i, x) + b, \quad (4)$$

其中 nsv 表示支持向量(SVs)的个数.

3 非线性 SVR 无约束优化

3.1 无约束优化目标函数

将优化问题(2)写成如下无约束的形式:

$$\min_{w} w^{\mathrm{T}}w + C\sum_{i}^{n}L_{\varepsilon}^{2}(y_i, w^{\mathrm{T}}\Phi(x_i) + b). \quad (5)$$

通过引入核函数 k,以及相应的再生核希尔伯特空间(RKHS) H,式(5)可改写为

$$\min_{f\in H}\lambda\|f\|_H^2 + \sum_{i}^{n}L_{\varepsilon}^{2}(y_i, f(x_i)), \quad (6)$$

其中 $\lambda = 1/C$ 为正则化参数. 为便于阐述问题,首先假定回归函数 $f(x_i)$ 中不含阈值 b.

假设损失函数 L 对于第 2 个变量是可微的,利用再生性质 $f(x_i) = \langle f, k(x_i, \cdot)\rangle_H$,可以对式(6)求导,并且在最优解 f^* 处,有如下等式成立:

$$2\lambda f^* + \sum_{i}^{n}\frac{\partial L}{\partial f}(y_i, f^*(x_i))\,k(x_i, \cdot) = 0. \quad (7)$$

利用表示定理[2],联合式(7),可将最优解表示为关于训练数据核函数的线性组合,即

$$f(x) = \sum_{i}^{n}\beta_i k(x_i, x). \quad (8)$$

设核矩阵为 K, $K_{ij} = k(x_i, x_j)$, $K_i \in \mathbf{R}^n$ 表示 K 的第 i 列. 此时式(6)可改写为

$$\min_{\beta}\lambda\beta^{\mathrm{T}}K\beta + \sum_{i}^{n}L_{\varepsilon}^{2}(y_i, K_i^{\mathrm{T}}\beta). \quad (9)$$

因此,无约束优化的目标函数为

$$\Omega(\beta) = \lambda\beta^{\mathrm{T}}K\beta + \sum_{i}^{n}\max(0, |y_i - K_i^{\mathrm{T}}\beta| - \varepsilon)^2.$$

3.2 牛顿优化法

当损失函数 L 可微时,可通过梯度下降法、牛顿法以及共轭梯度法对 $\Omega(\beta)$ 进行优化[5]. 首先运用与对偶优化问题密切相关的牛顿法实现优化.

对于给定的 β 值,将 $|y_i - f(x_i)| > \varepsilon$ 对应的样本点定义为支持向量集 sv. 假设对训练样本作重新排序,使得前 nsv 个样本为支持向量. 设 I^{sv} 为 $n\times n$ 的对角矩阵,其中前 nsv 个元素为 1,其余元素为零,并设一符号向量 $s \in \mathbf{R}^n$,满足如下条件:

$$s(i) = \begin{cases} 1, & y_i - K_i^{\mathrm{T}}\beta > \varepsilon; \\ -1, & y_i - K_i^{\mathrm{T}}\beta < -\varepsilon; \\ 0, & \text{otherwise}. \end{cases}$$

则梯度 ∇ 可表示为

$$\nabla = 2\lambda KI^{sv}\beta + 2KI^{sv}(KI^{sv}\beta - (Y - \varepsilon s)I^{sv}). \quad (10)$$

通过对梯度 ∇ 求导可得 Hessian 矩阵为

$$H = 2\lambda KI^{sv} + 2KI^{sv}KI^{sv}. \quad (11)$$

牛顿法由 $\beta \leftarrow \beta - H^{-1}\nabla$ 实现每一步的更新过程. 然而若初始点距离最优点 β^* 较远时,所产生的点列可能不收敛到 β^*. 此时可以考虑修正牛顿法,即 $\beta \leftarrow \beta - \gamma H^{-1}\nabla$,其中 γ 由一维搜索法[5]确定.

由式(10)和(11)可得

$$\nabla = H\beta - 2KI^{sv}(Y - \varepsilon s)I^{sv}.$$

因每一步更新后都有 $\nabla = 0$ 成立,故

$$\beta = (\lambda I_n + KI^{sv})^{-1}(Y - \varepsilon s)I^{sv}. \quad (12)$$

由于矩阵 $\lambda I_n + I^{sv}K$ 的左下角块为零,则有

$$\beta_{sv} = (Y_{sv} - \varepsilon s_{sv})^{-1}(Y_{sv} - \varepsilon s_{sv}), \quad (13)$$

其中 I_{nsv} 为 nsv ×nsv 的单位矩阵.

至此,可以给出基于修正牛顿法的 SVR 无约束问题优化算法的具体内容如下:

算法 1 基于修正牛顿法的 SVR

Function: $\beta = \mathrm{PrimalSVR}_{\mathrm{ENM}}(K, Y, \lambda)$

sv $\leftarrow \{1, \cdots, n\}$; $\beta\_\mathrm{old} = 0$

Repeat

$\beta = 0$;

$\beta_{sv} = (\lambda I_{nsv} + K_{sv})^{-1}(Y_{sv} - \varepsilon s_{sv})$;

step $= \beta - \beta\_\mathrm{old}$; $\gamma = 1$;

While $\Omega(\beta + \gamma\cdot\mathrm{step}) > \Omega(\beta) + \mu\gamma\cdot(\nabla^{\mathrm{T}}\mathrm{step})$

$\gamma \leftarrow \rho\cdot\gamma$;

End

$\beta \leftarrow \beta + \gamma\cdot\mathrm{step}$; $\beta\_\mathrm{old} \leftarrow \beta$;

sv $\leftarrow i$, 其中 i 满足 $|y_i - K_i^{\mathrm{T}}\beta| > \varepsilon$;

Until sv 不再发生改变.

算法 1 中,$\mu \subset (0, 0.5)$,$\rho \in (0, 1)$ 为对步长 γ 进行一维搜索所需的参数.

若考虑加入阈值,待优化参数为 $[b, \beta^{\mathrm{T}}]^{\mathrm{T}}$,则相应的增广 Hessian 矩阵为

$$H = 2\begin{bmatrix} \mathbf{1}^{\mathrm{T}}I^{sv}\mathbf{1} & \mathbf{1}^{\mathrm{T}}KI^{sv} \\ KI^{sv}\mathbf{1} & \lambda K + KI^{sv}KI^{sv} \end{bmatrix}, \quad (14)$$

其中 $\mathbf{1}$ 为具有适当维数的向量且所有元素均为 1.

此时,对应于式(13)的更新方程为

$$\begin{bmatrix} b \\ \beta_{sv} \end{bmatrix} = \begin{bmatrix} 0 & \mathbf{1}^{\mathrm{T}} \\ \mathbf{1} & \lambda I_{nsv} + K_{sv} \end{bmatrix}^{-1}\begin{bmatrix} 0 \\ Y_{sv} - \varepsilon s_{sv} \end{bmatrix}. \quad (15)$$

3.3 共轭梯度法

共轭梯度法本质上是求取近似解,也可以优化 $\Omega(\beta)$. 但值得注意的是,共轭梯度法的性能主要取决于待优化问题的参数化. 其收敛速度大约等于矩阵 K^2 的条件数,因此速度较慢,可通过 K 预处理解决上述问题.

基于共轭梯度法(采用预处理技术和 Fletcher-Reeves 更新策略)优化 $\Omega(\beta)$ 的伪代码由算

法 2 给出,其中 g 表示式(10) 中的梯度除以 K.

算法 2　基于共轭梯度法的 SVR

Function: β = PrimalSVR$_{CG}$(K, Y, λ)

sv $\leftarrow\{1, \cdots, n\}$; $\beta = 0$; d = $g_{old} = Y - \varepsilon$; iter = 1

Repeat

设 γ^* 为由一维搜索法确定的步长;

$\beta \leftarrow \beta + \gamma^* d$;

sv $\leftarrow i$,其中 i 满足 $o(i) = |y_i - K_i^T\beta| > \varepsilon$.

更新符号向量 s. $g_{new} \leftarrow 2\lambda\beta$;

$g_{new}(i) \leftarrow g_{new}(i) - 2 \cdot s(i) \cdot [o(i) - \varepsilon]$,其中 $i \in$ sv;

$$d \leftarrow -g_{new} + \frac{g_{new}^T K g_{new}}{g_{old}^T K g_{old}} d;$$

$g_{old} \leftarrow g_{new}$; iter $\leftarrow$ iter + 1;

Until iter 大于某一设定常数.

3.4　复杂度分析

基于修正牛顿法的 SVR 学习算法中,步长 γ 可以选为 1,且通常仅需几步迭代(不多于 10 步) 即可保证算法收敛,迭代次数与样本集规模 n 无关. 由于初始迭代设 $\beta = 0$,该算法的复杂度主要体现在求取 Hessian 矩阵的逆($O(n^3)$). 对偶优化问题(3) 中要同时优化 α 和 α_i,算法的复杂度可近似为 $O((2n)^3)$.

共轭梯度法的每一次迭代无需求逆,仅需计算 $K\beta$($O(n^2)$). 然而,共轭梯度法的收敛速度取决于核矩阵 K 的条件数,也即由核参数确定. 且通过算法 2 获得的支持向量数目与训练样本的个数相等,这必将加剧回归函数的复杂度. 因此,共轭梯度法的使用必须联合其他有效的稀疏化策略[6].

4　仿真实验

仿真实验包括以下内容:1) 利用牛顿法实现 SVR 模型选择;2) 证明共轭梯度法的收敛速度受核参数的影响;3) 对比基于两种不同策略下非线性 SVR 的训练速度. 仿真实验中所涉及的对偶优化是利用 MOSEK 基于内点法[7] 进行求解的,并以 RBF 核为非线性 SVR 的核函数.

4.1　模型选择

煤气炉数据集[8] 由 296 对输入 - 输出数据构成,输入表示空气流入速度,输出表示煤气炉排放气体中 CO_2 的浓度. 为构建煤气炉模型,可设 $y(i)$ 为系统的期望输出,输入向量由下式构成:

$$x_i = [y(i-1), \cdots, y(i-3), u(i-1), \cdots, u(i-3)].$$

如此可获得 293 个样本,选取其中下标为奇数的作为训练样本,下标为偶数的作为测试样本.

牛顿法的优势之一在于模型选择. 当采用留一法[9] 时,需对矩阵 $\lambda I_{nsv} + K_{sv}$ 求逆. 在牛顿法中,经过一次牛顿迭代,该逆矩阵无需另行求解,由式(13) 即可直接获得. 最终,通过留一法选择参数 $\varepsilon = 0.1$, $\lambda = 0.01$, $\sigma = 2.5$. 预测效果如图 1 所示,预测均方根误差(RMSE) 为 0.3102,支持向量的个数为 24.

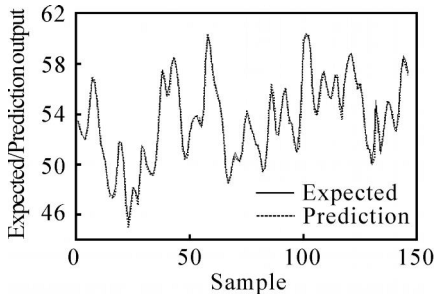

图 1　煤气炉数据集 SVR 模型预测效果

4.2　共轭梯度法

仍以煤气炉数据集建模为例,研究共轭梯度法在非线性 SVR 无约束优化中的应用. 将预测均方根误差作为共轭梯度迭代次数的函数,监测对应于不同 RBF 核参数的函数值,如图 2 所示. 该实验选择参数 $\varepsilon = 0.1$, $\lambda = 0.01$.

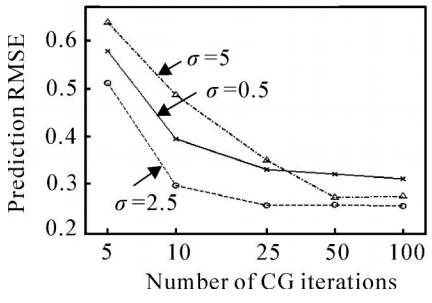

图 2　预测均方根误差随迭代次数的变化规律

由图 2 可以看出,共轭梯度法利用相对较少的迭代次数(25 ～ 50) 便可获得令人满意的解;共轭梯度法的收敛速度取决于 RBF 核参数 σ. 由图 2 还可以发现,当 $\sigma = 2.5$ 时,经过 25 次迭代,其预测均方根误差为 0.2854,明显优于牛顿法的结果(0.3102). 然而其需要 147 个支持向量,而牛顿法仅需 24 个,这是由于共轭梯度法需要采用一维搜索法来确定步长. 因此使用共轭梯度法时,必须联合有效的稀疏化策略.

4.3　训练时间对比

最后,为了对比各种学习方案的非线性 SVR 训练时间,以 Mackey-Glass 混沌时间序列预测为例:

$$\frac{dx(t)}{dt} = -0.1x(t) + \frac{0.2x(t - t_\Delta)}{1 + x(t - t_\Delta)^{10}},$$

其中时滞 $t_\Delta = 30$. 通过积分并选择嵌入维数 $d = 4$,可以获得输入输出数据集. 且本实验中采取的参数为 $\lambda = 0.001$, $\sigma = 0.5$, $\varepsilon = 0.01$.

基于无约束优化的非线性 SVR 训练遵循下述步骤:首先选择小规模的训练样本进行训练;然后选择较大规模的训练样本重新训练;依此类推. 为了便于对比,采用相同规模的训练样本实现基于对偶优化的非线性 SVR 训练,训练时间对比如图 3 所示. 由图 3 可以看出,相对于对偶优化方案,无约束优化方案显得更为高效.

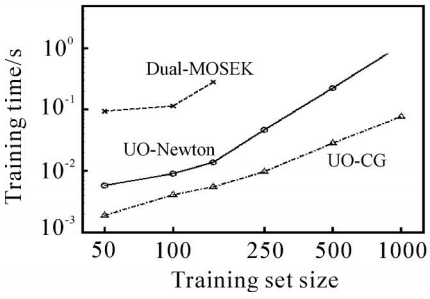

图 3 无约束优化与对偶优化的训练时间对比

5 结 论

本文主要研究了非线性 SVR 的无约束优化方案,并推导出牛顿法和共轭梯度法的更新规则. 实验结果表明,该方法不仅可以加快模型选择的过程,而且还能提高训练速度.

因为目前该方法仅适用于中等规模的数据集建模或预测,所以未来的研究重点在于如何获取稀疏的近似解,特别是对于共轭梯度法,毕竟该方法的解不具备稀疏性. 另外,能否利用无约束优化方案实现在线学习同样值得深入探讨.

参考文献(References)

[1] Vapnik V N. The nature of statistical learning theory [M]. Berlin: Springer-Verlag, 1995.

[2] Scholkopf B, Smola A J. Learning with kernels [M]. Cambridge: MIT Press, 2002.

[3] Keerthi S S, DeCoste D. A modified finite Newton method for fast solution of large scale linear SVMs[J]. J of Machine Learning Research, 2005, 6: 341-361.

[4] Chapelle O. Training a support vector machine in the primal[J]. Neural Computation, 2007, 19(5): 1155-1178.

[5] Boyd S, Vandenberghe L. Convex optimization [M]. Cambridge: Cambridge University Press, 2004.

[6] Keerthi S S, Chapelle O, DeCoste D. Building SVMs with reduced classifier complexity [J]. J of Machine Learning Research, 2006, 7: 1493-1515.

[7] Anderson E D, Anderson A D. The MOSEK interior point optimizer for linear programming [C]. High Performance Optimization. Boston: Kluwer Publishers, 2000: 197-232.

[8] Box G, Jenkins G M, Reinsel G C. Time series analysis, forecasting and control [M]. 3rd ed. New Jersey: Prentice Hall, 2005.

[9] Chapelle O, Vapnik V N, Bousquet O, et al. Choosing multiple parameters for support vector machines [J]. Machine Learning, 2002, 46(1): 131-159.

(上接第 124 页)

参考文献(References)

[1] Hendricks K B, Singhal V R. The effect of supply chain glitches on shareholder wealth [J]. J of Operation Management, 2003, 21(5): 501-523.

[2] Hendricks K B, Singhal V R. An empirical analysis of the effect of supply chain disruptions on long run stock price performance and equity risk of the firm [J]. Production and Operations Management, 2005, 14(1): 35-52.

[3] Eeckhoult L, Gollier C, Schlesinger H. The risk-averse (and prudent) newsboy[J]. Management Science, 1995, 41(11): 786-794.

[4] Fisher M A, Hammond J H, Obermeyer W R, et al. Making supply meet demand in an uncertain world[J]. Harvard Business Review, 1994, 72(3): 83-93.

[5] 索寒生, 储洪胜, 金以慧. 带有风险规避型销售商的供需链协调[J]. 控制与决策, 2004, 19(9): 1042-1044.
(Suo H S, Chu H S, Jin Y H. Supply chain coordination with risk aversion retailers[J]. Control and Decision, 2004, 19(9): 1042-1044.)

[6] 叶飞. 含风险规避者的供应链收益共享契约机制研究[J]. 工业工程与管理, 2006, 11(4): 50-54.
(Ye F. Research on revenue sharing contract mechanism of supply chain with risk averse agent [J]. Industrial Engineering and Management, 2006, 11(4): 50-54.)

[7] Haria Giannocaron, Pierpaolo Pontrandolfo. Supply chain coordination by revenue sharing contracts[J]. Int J of Production Economics, 2004, 89(2): 131-139.

[8] Xianghua Gan, Suresh P Sethi, Houmin Yan. Channel coordination with a risk-neutral supplier and a downside-risk-averse retailer [J]. Production and Operations Management Society, 2005, 1(14): 80-89.

[9] Fishburn P C. Mean-risk analysis with risk associated below-target returns [J]. The American Economic Review, 1977, 67(2): 116-126.

[10] Xianghua Gan, Suresh P Sethi, Houmin Yan. Coordination of supply chains with risk-averse agents [J]. Production and Operations Management, 2004, 13(2): 135-149.

[11] Telser L. Safety-first and hedging [J]. Review of Economic Studies, 1955, 23(Spring): 1-16.