

多项式光滑孪生支持向量回归机

黄华娟¹, 丁世飞^{1,2}

(1 中国矿业大学 计算机科学与技术学院, 江苏 徐州 221116;

2 中国科学院 计算技术研究所智能信息处理重点实验室, 北京 100190)

摘 要: 针对光滑孪生支持向量回归机(Smooth Twin Support Vector Regression, STSVR)中 Sigmoid 函数逼近精度不高的问题, 将正号函数展开为无穷多项式级数, 由此得到一族光滑函数. 采用该多项式光滑函数逼近孪生支持向量回归机的不可微项, 并用 Newton-Armijo 算法求解相应的模型, 提出了多项式光滑孪生支持向量回归机(Polynomial Smooth Twin Support Vector Regression, PSTSVR). 不仅从理论上证明了 PSTSVR 的收敛性和满足任意阶光滑的性能, 而且在人工数据集和 UCI 数据集上的实验表明了 PSTSVR 比 STSVR 具有更好的回归性能.

关键词: 孪生支持向量回归机; 多项式; 光滑; Newton-Armijo 算法

中图分类号: TP183

文献标识码: A

文章编号: 1000-7180(2013)10-0005-04

Polynomial Smooth Twin Support Vector Regression

HUANG Hua-juan¹, DING Shi-fei^{1,2}

(1 School of Computer Science and Technology, China University of Mining and Technology, Xuzhou 221116, China;

2 Key Laboratory of Intelligent Information Processing, Institute of Computing Technology,

Chinese Academy of Science, Beijing 100190, China)

Abstract: Sigmoid function is used as the smoothing function in Smooth Twin Support Vector Regression (STSVR). However, the approximation accuracy of Sigmoid function is low. In order to overcome this problem, in this paper, a new smooth twin support vector regression, term as Polynomial Smooth Twin Support Vector Regression (PSTSVR), is proposed. In PSTSVR, plus function is transformed to an infinite polynomial series. Thus a family of smoothing functions is derived. Polynomial function is used to approximate the non-differential term of twin support vector regression. Then Newton-Armijo algorithm is used to solve the corresponding model. We have proved that PSTSVR is not only convergent, but also can meet the arbitrary order smooth performance. Meanwhile, the experimental results on several artificial and benchmark datasets show that PSTSVR has better regression performance than STSVR.

Key words: twin support vector regression; polynomial function; smooth; Newton-Armijo

1 引言

支持向量机(Support Vector Machine, SVM)是近年来提出的一种基于统计学习理论的机器学习方法^[1]. 一些经典的改进方法虽然在一定程度上提高了 SVM 的学习性能, 但算法实现比较困难^[2]. 近年来, 很多学者开始研究基于标准 SVM 的变形算法. 文献[3]中提出了孪生支持向量机(Twin Sup-

port Vector Machines, TWSVM). TWSVM 将 SVM 问题转换为两个较小的凸二次规划问题, 能将训练时间约减到 SVM 的 1/4. 接着文献[4]中提出了孪生支持向量回归机(Twin Support Vector Regression, TSVR). 使训练速度大为提高. 然而, TSVR 是在对偶空间中求解两个二次规划问题, 这种求解方法对于样本数目较大的问题将受到时间和内

收稿日期: 2013-01-12; 修回日期: 2013-03-30

基金项目: 国家重点基础发展计划(2013CB329502); 国家自然科学基金项目(41074003)

存的制约. 文献[5]中提出了光滑孪生支持向量回归机 (Smooth Twin Support Vector Regression, STSVR). 然而, Sigmoid 函数的逼近精度较差, 使得 STSVR 模型的效果有待进一步提高. 为了改善其回归效果, 寻找性能更好的光滑函数是研究 STSVR 的一个关键问题. 本文引入一族多项式函数作为光滑函数来逼近 TSVR 的目标函数, 最后采用 Newton-Armijo 方法求解得到的新模型, 从而提出多项式光滑孪生支持向量回归机 (Polynomial Smooth Twin Support Vector Regression, PSTSVR). 不仅从数学理论角度上证明了该模型的收敛性和任意阶光滑性, 还从实验角度验证了该模型的有效性.

2 多项式光滑孪生支持向量回归机

2.1 PSTSVR 原理及其性质

令 $A_{l \times n}$ 为训练样本输入数据集, 即 $\{x_k\}_{k=1}^l$; 令 $Y_{l \times 1}$ 为训练样本输出数据集, 即 $A_{l \times n}$ 对应的回归为 $Y_{l \times 1} = [y_1, y_2, \dots, y_l]^T$.

孪生支持向量回归机是要在训练数据点两侧产生一对不平行的函数, 分别确定回归函数的 ϵ - 不敏感上、下界.

对于线性的情况, TSVR 通过训练数据的 ϵ_1 不敏感下界

$$f_1(x) = w_1^T x + b_1 \quad (1)$$

与 ϵ_2 不敏感上界

$$f_2(x) = w_2^T x + b_2 \quad (2)$$

确定最终的回归函数, 而这对函数可以通过求解下面的一对 QP 问题得到:

$$\begin{aligned} \min \quad & \frac{1}{2} \|Y - e\epsilon_1 - (Aw_1 + eb_1)\|^2 + C_1 e^T \xi \\ \text{s. t.} \quad & Y - (Aw_1 + eb_1) \geq e\epsilon_1 - \xi, \xi \geq 0 \end{aligned} \quad (3)$$

$$\begin{aligned} \min \quad & \frac{1}{2} \|Y + e\epsilon_2 - (Aw_2 + eb_2)\|^2 + C_2 e^T \eta \\ \text{s. t.} \quad & (Aw_2 + eb_2) - Y \geq e\epsilon_2 - \eta, \eta \geq 0 \end{aligned} \quad (4)$$

式中, $C_1, C_2 > 0, \epsilon_1, \epsilon_2 > 0$ 为常数, ξ, η 为松弛变量, e 为 $l \times 1$ 维的单位列向量.

在求解式(3)和式(4)这两个二次规划问题时, 一般地, 我们都在对偶空间中进行求解. 当训练样本数目较多时, 这种求解方法将使得算法的效率降低. 为了提高 TSVR 的效率, 本文先将它转化为无约束的最优化问题, 然后再运用求解速度快的优化算法进行求解.

由 KKT 条件, 式(3)和式(4)的松弛变量 ξ 和 η 可以变为

$$\xi = \max(0, e\epsilon_1 - Y + Aw_1 + eb_1) \quad (5)$$

$$\eta = \max(0, e\epsilon_2 + Y - (Aw_2 + eb_2)) \quad (6)$$

将式(5)和式(6)分别带入式(3)和式(4), 则可以得到无约束的 TSVR 模型:

$$\begin{aligned} \min \quad & \frac{1}{2} \|Y - e\epsilon_1 - (Aw_1 + eb_1)\|^2 + \\ & C_1 e^T \max(0, e\epsilon_1 - Y + Aw_1 + eb_1) \end{aligned} \quad (7)$$

$$\begin{aligned} \min \quad & \frac{1}{2} \|Y + e\epsilon_2 - (Aw_2 + eb_2)\|^2 + \\ & C_2 e^T \max(0, e\epsilon_2 + Y - (Aw_2 + eb_2)) \end{aligned} \quad (8)$$

$$\text{令 } \varphi_1 = \max(0, e\epsilon_1 - Y + Aw_1 + eb_1) = (u_1)_+ \quad (9)$$

$$\varphi_2 = \max(0, e\epsilon_2 + Y - (Aw_2 + eb_2)) = (u_2)_+ \quad (10)$$

式中, $(u_1)_+$ 和 $(u_2)_+$ 为正号函数. 则式(7)和式(8)可以写成

$$\begin{aligned} \min \quad & \frac{1}{2} \|Y - e\epsilon_1 - (Aw_1 + eb_1)\|^2 + \\ & C_1 e^T (u_1)_+ \end{aligned} \quad (11)$$

$$\begin{aligned} \min \quad & \frac{1}{2} \|Y + e\epsilon_2 - (Aw_2 + eb_2)\|^2 + \\ & C_2 e^T (u_2)_+ \end{aligned} \quad (12)$$

不难看出, 式(11)和式(12)的第二项不光滑, 无法用快速的数值优化方法进行求解. 我们需要采用光滑函数逼近式(11)和式(12)的不可微项, 再采用梯度算法进行求解.

2.2 多项式光滑函数

Weierstrass 定理 设任意的连续函数 $f(x), x \in [m, n]$, 则存在多项式 $P_n(x)$, 使得 $\lim_{n \rightarrow \infty} \max_{m \leq x \leq n} |f(x) - P_n(x)| = 0$.

Weierstrass 定理^[6]说了闭区间上的任意连续实值函数, 可用多项式函数任意地一致逼近. 因为正号函数是连续函数, 因此可以用多项式函数去逼近它. 文献[6]中只是给出了一阶和二阶的两个光滑多项式函数的形式, 本文依据类似的思想, 采用函数的级数展开法给出此类光滑函数的通用公式.

定理 1 正号函数 x_+ 在 $[-\frac{1}{k}, \frac{1}{k}]$ 上可以展开成多项式级数为

$$\begin{aligned} x_+ = \frac{1}{2k} \left(\frac{1+k^2x^2}{2} - \sum_{n=2}^{\infty} \frac{(2n-3)!!}{(2n)!!} \right. \\ \left. (1-k^2x^2)^n \right) + \frac{x}{2} \end{aligned} \quad (13)$$

证明略.

定理 2 x_+ 在 $[-\frac{1}{k}, \frac{1}{k}]$ 上的多项式逼近函数为

$$P_n(x, k) = \begin{cases} x, & x \geq \frac{1}{k} \\ \frac{1}{2k} \left(\frac{1+k^2x^2}{2} - \sum_{l=2}^n \frac{(2l-3)!!}{(2l)!!} (1-k^2x^2)^l \right) + \frac{x}{2}, & |x| < \frac{1}{k}, k > 0 \\ 0, & x \leq -\frac{1}{k} \end{cases} \quad (14)$$

式中, n 为正整数.

定理 3 $P_n(x, k)$ 由定理 3 给出, 则:

- (1) $P_n(x, k)$ 关于 x 具有 n 阶光滑性;
- (2) $\lim_{n \rightarrow \infty} \max (P_n(x, k) - x_+) = 0$

证明略.

定理 3 说明, 在 n 取足够大时, 采用展开级数法的多项式光滑函数对正号函数的逼近是可以达到任意给定的精度的.

2.3 最优光滑因子

在式(14)中有一个参数 k , 称为光滑因子. 下面我们给出光滑因子最优值的公式.

定义 1 任意给定逼近精度 E , 如果光滑函数 $P_n(x, k)$ 对 x_+ 的逼近满足 $|P_n(x, k) - x| \leq E$, 则称使得该式成立的光滑因子的取值为最优光滑因子, 记为 $k_{\text{opt}}(n, E)$.

按照定义 1, 有 $|P_n(x, k) - x| \leq E$, 再由定理 3 可知, $P_n(x, k) - x_+ \leq E$, 因为 $P_n(x, k)$ 对 x_+ 的逼近积分在 $x = 0$ 时误差最大, 所以只要保证在 $x = 0$ 时 $P_n(x, k) - x_+ \leq E$ 成立即可. 将 $x = 0$ 代入式(6), 得到

$$k_{\text{opt}}(n, E) \geq \frac{\frac{1}{2} - \sum_{l=2}^n \frac{(2l-3)!!}{(2l)!!}}{2E} \quad (15)$$

2.4 基于 Newton-Armijo 法的 PSTSVR

对所有的 $n \geq 2$, $P_n(x, k)$ 具有二阶光滑, 因此可以采用 Newton-Armijo 法求解下面两个无约束优化问题, 具体的实现步骤如算法 1 所示.

$$\Phi^1 = \min \frac{1}{2} \|Y - e\epsilon_1 - (Aw_1 + eb_1)\|^2 + C_1 e^T P((e\epsilon_1 - Y + Aw_1 + eb_1), k) \quad (16)$$

$$\Phi^2 = \min \frac{1}{2} \|Y + e\epsilon_2 - (Aw_2 + eb_2)\|^2 +$$

$$C_2 e^T P((e\epsilon_2 + Y - (Aw_2 + eb_2)), k) \quad (17)$$

算法 1 基于 Newton-Armijo 法的 PSTSVR:

Input: 给定初始点 $(w^0, \gamma^0) \in R^{n+1}$, η , 令迭代步骤 $i = 0$;

Output: 目标函数的最优值;

Step1: 计算 $\Phi^i = \Phi(w^i, b^i)$ 和 $g^i = \nabla \Phi(w^i, b^i)$;

Step2: 如果 $\|g^i\| \leq \eta$, 则选取 $(w^*, b^*) = (w^i, b^i)$, 停机; 否则由方程 $\nabla^2 \Phi(w^i, b^i) d^i = -g^i$, 计算下降方向 d^i ;

Step3: 取 $\delta \in (0, \frac{1}{2})$, $\lambda_i = \max\{1, \frac{1}{2}, \frac{1}{4}, \dots\}$

使得

$\Phi(w^i, b^i) - \Phi(w^i, b^i) + \lambda_i d^i \geq -\lambda_i g^i d^i$ 成立, 令 $(w^{i+1}, b^{i+1}) = (w^i, b^i) + \lambda_i d^i$;

Step4: 令 $i \leftarrow i + 1$, 转 Step2.

2.5 非线性 PSTSVR

PSTSVR 不仅可以解决线性分类问题, 如果将前面的结论推广到非线性光滑 PSTSVR, 就可以处理非线性问题.

基于核空间 TSVR 的两个超平面可以表示为

$$f_1(x) = K(x^T, A^T)w_1 + b_1 \quad (18)$$

$$f_2(x) = K(x^T, A^T)w_2 + b_2 \quad (19)$$

引入多项式光滑函数后, 则非线性 PSTSVR 的优化问题为

$$\min \frac{1}{2} \|Y - e\epsilon_1 - (K(A, A^T)w_1 + eb_1)\|^2 + C_1 e^T P_n(x, k) \quad (20)$$

$$\min \frac{1}{2} \|Y + e\epsilon_2 - (K(A, A^T)w_2 + eb_2)\|^2 + C_2 e^T P_n(x, k) \quad (21)$$

前面的定理和定义也适用于非线性 PSTSVR.

3 实验与分析

为了测试提出的 PSTSVR 算法的性能, 我们将对人工数据集和 UCI 数据集进行测试, 并和 SVR、TSVR 和 STSVR 的测试结果进行比较. 以上几种方法都是在 Intel(R) Core(TM) 2 Duo CPU E4500、2 GB 内存和 MATLAB7. 11. 0 的环境中运行的.

3.1 人工数据集

$\sin x(x)$ 函数经常被用来测试各种机器学习方法的回归性能, 其表达式为

$$y = \sin x(x) = e^{-\frac{x}{2.5}} \sin 3x, x \in [0, 4\pi].$$

为了更有效地测试我们的算法, 在训练样本中,

分别加入不同的零平均值高斯噪声和均匀分布噪声. 从实验结果可以看出, PSTSVR 和 STSVR 能得到比较小的 SSE, SSE/SST 和较大的 SSR/SST, 这说明 PSTSVR 的逼近精度比 STSVR 的逼近精度高. 此外, PSTSVR 的 CUP 时间少于 STSVR 的时间. 这些实验结果表明, 和 STSVR 相比, PSTSVR 具有更好的回归效果.

表 1 SVR、TSVR、STSVR 和 PSTSVR 在 UCI 数据集上的比较结果

Dataset	Algorithm	SSE	SSE/SST	SSR/SST	Time/s
Diabetes(43×3)	SVR	0.450 7±0.052 1	0.519 2±0.567 3	0.601 4±0.017 4	3.474
	TSVR	0.368 8±0.198 3	0.416 5±0.580 1	0.610 8±0.051 5	0.982
	STSVR	0.351 2±0.156 2	0.412 9±0.390 1	0.629 4±0.127 4	0.215
	PSTSVR	0.324 5±0.012 6	0.382 5±0.212 1	0.652 8±0.212 7	0.115
Boston Housing(506×14)	SVR	0.405 6±0.017 4	0.128 7±0.035 2	0.905 6±0.154 7	69.842
	TSVR	0.397 4±0.254 7	0.123 5±0.036 2	1.025 7±0.102 5	20.712
	STSVR	0.372 2±0.142 6	0.120 8±0.064 4	0.992 8±0.124 6	0.295 2
	PSTSVR	0.340 5±0.152 2	0.120 1±0.013 5	1.014 5±0.021 2	0.089 9

从表 1 可以看出, PSTSVR 获得的 SSE, SSE/SST 和 SSR/SST 比 STSVR 好. 这说明和 STSVR 相比, PSTSVR 的泛化能力有了一定地提高. 并且, PSTSVR 的 CPU 时间比 STSVR 的要少一些, 并远远小于 TSVR 的时间.

4 结束语

采用多项式光滑函数逼近孪生支持向量回归机的不可微项, 并用 Newton-Armijo 算法求解相应的模型, 提出了一种新的光滑孪生支持向量回归机——多项式光滑孪生支持向量回归机 (PSTSVR). 本文不仅从理论上证明了 PSTSVR 的收敛性和任意阶光滑性, 而且在人工数据集和 UCI 数据集上的实验表明了 PSTSVR 比 STSVR 具有更好的回归性能. 鉴于 PSTSVR 的良好性能, 把 PSTSVR 应用到其他领域是下一步的工作.

参考文献:

[1] Shifei Ding, Bingjuan Qi. Research of granular support vector machine [J]. Artificial Intelligence Review, 2012, 38(1):1-7.

3.2 UCI 数据集

为了进一步检验算法的性能, 我们将对 2 个典型的 UCI 数据集进行测试. 这 2 个 UCI 数据集分别是 Diabetes 和 Boston housing.

表 1 中显示的是 SVR、TSVR、STSVR 和 PSTSVR 分别对 2 个数据集进行 10 次测试的平均结果.

[2] Platt J C. Using analytic QP and sparseness to speed training of support vector machines [C]// Advances in Neural Information Processing Systems 11. Cambridge, MA: MIT Press, 1999:557-563.

[3] Jayadeva, Khemchandni Reshma, Suresh Chandra. Twin support vector machines for pattern classification [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(5):905-910.

[4] Xinjun Peng. TSVR: an efficient twin support vector machine for regression [J]. Neural Networks, 2010 (23):365-372.

[5] Xiaobo Chen, Jian Yang, Jun Liang. Smooth twin support vector regression [J]. Neural Computer & Application, 2012(21):505-513.

[6] 袁玉波, 严杰, 徐成贤. 多项式光滑的支撑支持向量机 [J]. 计算机学报, 2005, 28(1):9-17.

作者简介:

黄华娟 女, (1984-), 博士研究生. 研究方向为模式识别与机器学习.

丁世飞 男, (1963-), 博士后, 教授, 博士生导师. 研究方向为模式识别与机器学习等.