

用于回归的临近支持向量机

杜喆¹ 胡廷锋² 刘三阳¹

(西安电子科技大学应用数学系 西安 710071)¹ (洛阳师范学院数学科学学院 洛阳 471022)²

摘要 将临近支持向量分类机应用在回归问题上,提出临近支持向量回归机,给出线性与非线性情况下的回归函数,该方法比支持向量回归机(SVR)问题减少了参数和一半变量,比最小二乘支持向量回归机(LSSVM)求解公式更加简单,且核函数不需要满足 Mercer 条件。数值实验结果表明,与 SVR 和 LSSVM 相比,该方法的学习速度更快,且泛化能力较之不相上下。

关键词 临近支持向量机,回归,支持向量机

中图分类号 TP181 **文献标识码** A

Proximal Support Vector Machine for Regression

DU Zhe¹ HU Ting-feng² LIU San-yang¹

(Department of Applied Mathematics, Xidian University, Xi'an 710071, China)¹

(College of Mathematics, Luoyang Normal University, Luoyang 471022, China)²

Abstract A new technique—proximal support vector regression machine (PSVRM) was obtained when extended proximal support vector machine classifier towards regression problem, the linear and nonlinear regression function of PSVRM was proposed. PSVRM has less arguments and half variables than support vector regression(SVR), the solving formula is much simpler than that of least square support vector machine for regression(LSSVM), and the kernel is no need to satisfy the Mercer's condition. Numerical simulations show that, comparing with SVR and LSSVM, PSVRM is faster and the generalization ability is sometimes faster than the former two.

Keywords Proximal support vector machine, Regression, Support vector machine

1 引言

支持向量机(Support Vector Machine, SVM)^[1]是基于统计学习理论提出的一种机器学习方法,避免了传统方法的一些缺点,有着良好的性能。SVM 最早是针对模式识别问题提出的,随着对 ϵ 不敏感损失函数的引入, SVM 被推广到回归估计并展现了极好的学习性能^[2],在金融时间序列预测^[3]、系统建模^[4]得到广泛应用。但标准 SVM 需求解一个约束二次规划^[1,2],计算复杂度,为了提高训练速度,通常采用 Chunks, Osunna 算法、SMO 等方法求解^[5,6]。为将原问题简化,从变形角度也有很多学者做了相关研究。文献[7]提出基于线性规划的 SVM 变型,计算速度大大加快,泛化能力逼近标准 SVM。文献[8]提出的最小二乘支持向量机(Least Square Support Vector Machine, LSSVM)将标准 SVM 的损失函数由一次变为二次,将不等式约束改变为等式约束,从而通过解方程组得到问题的解,避免了求解二次规划而大大降低计算复杂度,文献[9,10]对其在回归方面进行了研究。文献[11]提出临近支持向量机(Proximal Support Vector Machine, PSVM),其思想是将两类样本分别集中在两个平行超平面附近,然后使这两个超平面的间隔达到最大。由于 PSVM 算法

简单、训练速度快,引起广泛的关注,文献[12-15]针对 PSVM 做了不同方面的研究,但都是基于分类问题,本文将 PSVM 推广到回归估计得到临近支持向量回归机(Proximal Support Vector Regression Machine, PSVRM),给出线性与非线性情形的 PSVRM 表达式,与 LSSVM 一样, PSVRM 比 SVR 减少一个参数和一半的问题变量,降低计算复杂度,但是其求解比 LSSVM 更加简单快速,经在人工数据和真实数据集上的数值实验表明, PSVRM 有着良好的函数估计能力。

2 线性临近支持向量回归机

假定训练集为 $\{(x_i, y_i)\}_{i=1}^n \in R^m \times R$, 记为矩阵 $A_{n \times m} \times Y$, 这里 A 的每一行为一个 x_i , Y 为样本输出值; ξ 为松弛因子, e 为一个元素全为 1 的列向量, C 为一正实数, 设所求的回归函数为 $f(x) = w^T x + \gamma$ 则问题可描述为:

$$\min \frac{C}{2} \|\xi\|^2 + \frac{1}{2} (\|w\|^2 + \gamma^2) \\ s. t. Aw + e\gamma + \xi = Y \quad (1)$$

其 Lagrangian 函数为:

$$L(w, \gamma, \xi, u) = \frac{C}{2} \|\xi\|^2 + \frac{1}{2} (\|w\|^2 + \gamma^2) - u^T (Aw + e\gamma + \xi - Y) \quad (2)$$

收稿日期:2008-01-30 本文受国家自然科学基金项目(60574075, 60705004)资助。

杜喆(1982—),男,博士研究生,研究方向是机器学习、最优化计算方法;胡廷锋(1964—),男,硕士,副教授,研究方向是多元统计与计量经济;刘三阳(1959—),男,博士,教授,博导,研究方向是人工智能、最优化方法及其在通信网络中应用、决策支持系统。

根据 KKT 条件,令对变量求偏导为 0 有:

$$\begin{cases} \frac{\partial L}{\partial w} = w - A^T u = 0 \\ \frac{\partial L}{\partial \gamma} = \gamma - e^T u = 0 \\ \frac{\partial L}{\partial \xi} = C\xi - u = 0 \\ \frac{\partial L}{\partial u} = -Aw - e\gamma - \xi + Y = 0 \end{cases} \quad (3)$$

即可得出: $w = A^T u$, $\gamma = e^T u$, $\xi = C^{-1} u$ 。将他们代入式(3)的最后一个等式有:

$$AA^T u + ee^T u + C^{-1} u = Y \quad (4)$$

$$\text{所以有: } u = (AA^T + ee^T + C^{-1} I)^{-1} Y \quad (5)$$

这里 I 为单位矩阵,如果令 $H = [A \ e]$,则有:

$$u = (C^{-1} I + HH^T)^{-1} Y \quad (6)$$

如果样本的个数大于样本的维数,则使用 SMW (Sherman-Morrison-Woodbury) 公式^[10]来求矩阵的逆,可以降低求逆矩阵的维数,大大减少计算复杂度。得到:

$$u = C(I + H(C^{-1} I + H^T H)^{-1} H^T) Y \quad (7)$$

则线性临近支持向量回归机的函数:

$$f(x) = w^T x + \gamma = (Ax + e)^T u \quad (8)$$

3 非线性临近支持向量回归机

对于非线性情形,使用 $w = A^T u$ 替换目标函数,使用核函数 $K(A, A^T) = K$ 代替约束中的 AA^T ,则问题(1)变为:

$$\min \frac{C}{2} \|\xi\|^2 + \frac{1}{2} (\|u\|^2 + \gamma^2)$$

$$\text{s. t. } K(A, A^T)u + e\gamma + \xi = Y \quad (9)$$

其 Lagrangian 函数为:

$$L(w, \gamma, \xi, v) = \frac{C}{2} \|\xi\|^2 + \frac{1}{2} (\|u\|^2 + \gamma^2) - v^T (Ku + e\gamma + \xi - Y) \quad (10)$$

同样令偏导为 0,有:

$$\begin{cases} \frac{\partial L}{\partial u} = u - K^T v = 0 \\ \frac{\partial L}{\partial \gamma} = \gamma - e^T v = 0 \\ \frac{\partial L}{\partial \xi} = C\xi - v = 0 \\ \frac{\partial L}{\partial v} = -Ku - e\gamma - \xi + Y = 0 \end{cases} \quad (11)$$

得: $u = K^T v$, $\gamma = e^T v$, $\xi = C^{-1} v$,代入后消去有:

$$KK^T v + ee^T v + C^{-1} v = Y \quad (12)$$

$$\text{即: } v = (KK^T + ee^T + C^{-1} I)^{-1} Y \quad (13)$$

令 $G = [K \ e]$,则

$$v = (C^{-1} I + GG^T)^{-1} Y \quad (14)$$

这样可以对式(6)和式(14)使用相同的程序代码。得到非线性临近支持向量机的回归函数为

$$f(x) = (KK(A, x) + e)^T v \quad (15)$$

问题(1)和(9)中目标函数的第一项为平方错误值的和,第2项为使回归函数光滑的正则项,则参数 C 是经验风险和泛化能力的一个平衡因子,一般通过交叉验证法选取。

与标准 SVR 相比,本文的 PSVRM 减少了一个调整参数 ϵ ,且减少了 n 个变量,从而大大降低了计算时间空间复杂度。与 LSSVMR 相比,PSVRM 的目标函数是严格凸的,问题求

解更加简单,速度更快。另外对于非线性情形,由于 GG^T 对称正定,故 PSVRM 在选择核函数时可以不满足 Mercer 的正定条件^[1,8]。

4 数值实验

以下数值实验采用 matlab7.0.1 软件在 CPU3.0G,内存 1G 的 PC 机上进行。首先对含正态噪声的线性函数 $y = x + \tau$, $\tau \sim N(0, 1)$,进行回归估计得到图 1,图 1(a)中虚线是 $y = x$,图 1(b)中实线为线性 PSVRM 的拟合线,可看出二者几乎没有差别。再对含正态噪声的非线性函数 $y = \sin x/x + \tau$, $\tau \sim N(0, 0.2)$ 进行回归估计得到图 2,采用径向基核函数 (RBF, $K(x, y) = \exp(-\mu \|x - y\|^2)$),虚线为 $y = \sin x/x$,实线为非线性 PSVRM 的拟合曲线,同样可看出本文方法的有效性。

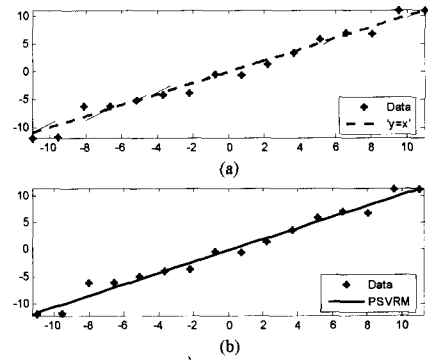


图 1 线性 PSVRM 的回归线图

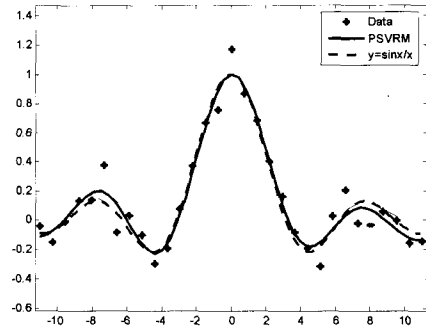


图 2 非线性 PSVRM 回归线图

此外再在 UCI 数据库^[16]中的真实数据集上进行 SVR, LSSVMR 与 PSVRM 这 3 个算法的回归性能比较,SVR 代码取自软件包 Libsvm2.82^[17],LSSVMR 与 PSVRM 采用 Matlab 语言编写。随机取数据集中一定数量的样本作为训练集,余下的作为测试集,核函数取为 RBF 核,评价指标采用 mean squared error(MSE)和 linear correlation coefficient(R)以及训练时的 CPU 时间(s),这里:

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - f_i)^2,$$

$$R = \frac{\sum_{i=1}^N (y_i - \bar{y})(f_i - \bar{f})}{\sqrt{\sum_{i=1}^N (y_i - \bar{y})^2 \sum_{i=1}^N (f_i - \bar{f})^2}}$$

f_i 为回归函数的预测值, $\bar{f}$ 为预测值的均值, $\bar{y}$ 为真实样本数值 y_i 的均值。MSE 越小、 R 越接近 1 表明回归效果越好。采用交叉验证来确定参数 C 和 μ , 3 个算法在相同的参数下

(下转第 157 页)

出版社,2004

[4] Louden K C. 程序设计语言-原理与实践(第二版)[M]. 黄林鹏, 毛宏燕, 黄晓琴, 等译. 北京: 电子工业出版社, 2004

[5] Partt T W, Zelkowitz M V. 程序设计语言: 设计与实现(第四版)[M]. 傅育熙, 张冬荣, 黄林鹏, 译. 北京: 电子工业出版社, 2001

[6] Winskel G. 程序设计语言的形式语义[M]. 宋国新, 邵志清, 等译. 北京: 机械工业出版社, 2004(1): 45-60

[7] 刘晓杰. 智能神经网络程序设计语言中规则的指称语义[J]. 计算机应用研究, 2004(10)

[8] Zhao Xiaofang, Li Fanzhang. The Frame of DFL Programming Language. Journal of Communication And Computer, 3(1): 1-15

[9] 李凡长, 等. 动态模糊逻辑及其应用. 昆明: 云南科技出版社, 1997

[10] 李凡长, 等. 动态模糊集及其应用. 昆明: 云南科技出版社, 1997

[11] 李凡长, 等. 动态模糊逻辑引论. 昆明: 云南科技出版社, 2005: 8-40

[12] 李凡长, 郑家亮. 动态模糊数据的设计语言[J]. 计算机科学,

1997(增): 54-56

[13] 李凡长, 等. 动态模糊数据模型研究. 计算机研究与发展, 1998, 35(8): 714-718

[14] 李凡长. 动态模糊数据运算及模型设计. 计算机工程, 2001, 27(3): 100-102

[15] 李凡长, 等. 动态模糊数据模型及其设计语言. 计算机科学, 1997, 24(12): 49-53

[16] Li Fanzhang. Dynamic Fuzzy Logic and Its Applications. Nova Science Publishers, 2007(12)

[17] Zhao Xiaofang, Fan Hui, Liu Xiaohua. Data Types of DFL Programming Language. Fuzzy Systems And Knowledge Discovery. Haikou, Hainan, China, 2007: 116-120

[18] Zhao Xiaofang, Li Fanzhang. Denotational Semantics of Dynamic Fuzzy Logic Programming Language // 2006 IEEE International Conference on Granular Computing [C]. Atlanta, USA. May 2006: 409-412

(上接第 127 页)

运行 10 次后取平均值得到结果见表 1。

需要指出的是 Libsvm2.82 的内部函数采用 C++ 语言编写, 而这里 PSVRM 与 LSSVMR 都采用 Matlab 语言, 但是

仍然能看出本文方法的速度优势。从表 1 中可以看出 PSVRM 与 LSSVMR 和 SVR 的拟合能力不相上下, 但是学习速度最快。

表 1 PSVRM 与 SVR, LSSVMR 的性能比较

Datasets (train/test/dim.)	PSVRM			SVR($\epsilon=0.01$)			LSSVMR		
	MSE	R	Time	MSE	R	Time	MSE	R	Time
Housing(300/206/13)	12.125	0.920	0.016	12.594	0.917	0.218	12.188	0.929	0.048
M-g(800/585/6)	0.015	0.850	0.625	0.017	0.833	2.864	0.014	0.853	1.297
Body-fat(200/52/14)	4.5-e5	0.932	0.015	3.0-e5	0.951	0.165	2.7-e5	0.959	0.088
Space-ga(2000/1107/6)	0.011	0.842	1.406	0.010	0.851	19.62	0.010	0.843	3.156

结束语 本文将用于模式分类的 PSVM 方法, 推广到回归估计, 给出线性与非线性情形下的 PSVRM 函数估计公式, PSVRM 简单快速, 核函数可以不一定满足正定条件。数值实验表明本文方法可行有效, 与 SVR 和 LSSVMR 相比, 其速度最快, 且泛化能力不相上下。下一步的研究工作可基于 PSVRM 在非线形系统辨识、金融时间序列预测、预报建模与控制方面的应用进行。

参 考 文 献

[1] Vapnik V. The Nature of Statistical Learning Theory[M]. New York, Springer, 1995

[2] Drucker H, Burges C J C, Kaufman L. Support vector regression machines. Advances in Neural Information Processing Systems, Cambridge: MIT Press, 1997: 155-161

[3] Tay F E H, Cao L. Application of support vector machines in financial time series forecasting[J]. Omega, 2001, 29(4): 309-317

[4] Chan W C, Chan C W, Cheung K C, et al. On the modeling of nonlinear dynamic systems using support vector neural networks[J]. Engineering Application of Artificial Intelligence, 2001, 14(2): 105-113

[5] 汪西莉, 刘芳, 焦李成. 基于大规模数据的支撑矢量机的训练和分类[J]. 西安电子科技大学学报, 2002, 29(1): 123-127

[6] Flake G W, Lawrence S. Efficient SVM regression training with SMO[J]. Machine Learning, 2002, 41(1): 271-290

[7] 周伟达, 张莉, 焦李成. 线性规划支撑矢量机[J]. 电子学报, 2001, 29(11): 1-5

[8] Suykens J A K, Vandewalle J. Least Square Support Vector Ma-

chine Classifiers[J]. Neural Processing Letters, 1999, 9: 293-300

[9] Gestel T V, et al. Financial time series prediction using least squares support vector machines within the evidence framework[J]. IEEE Trans on Neural Networks, 2001, 12(4): 809-821

[10] 吴青, 刘三阳, 杜喆. 回归型模糊最小二乘支持向量机[J]. 西安电子科技大学学报, 2007, 34(5): 773-778

[11] Fung G, Mangasarian O L. Proximal support vector machine classifiers // Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, CA, USA, 2001(8): 77-82

[12] Agarwal D K. Shrinkage Estimator Generalization of Proximal Support Vector Machine [C] // Proceeding of the 8th ACM SIGKDD Intl Conference on Knowledge Discovery and Data Mining. Edmonton, Canada, 2002: 173-182

[13] Fung G, Mangasarian O L. Multicategory Proximal Support Vector Classifiers[J]. Machine Learning, 2005, 59: 77-97

[14] Li K, Huang H. Incremental Learning Proximal Support Vector Machine Classifiers [C] // Proceedings of the 1st International Conference on Machine Learning and Cybernetics. Beijing, 2002, 3: 1635-1637

[15] 陶晓燕, 姬红兵, 马志强. 基于样本分布不平衡的近似支持向量机[J]. 计算机科学, 2007, 34(5): 174-176

[16] Chang C C, Lin C J. LIBSVM-A Library for Support Vector Machines [CP]. 2005. <http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html>

[17] Murphy P M, Aha D W. UCI repository of machine learning databases [DB]. 2006. <http://www.ics.uci.edu/~mllearn/ML-Repository.html>