基于支持向量机分类的回归方法*

陶 卿[1,2], 曹进德[3], 孙德敏[4]

[1](中国科学院 自动化研究所,北京 100080);
[2](中国人民解放军炮兵学院 一系,安徽 合肥 230031);
[3](东南大学 应用数学系,江苏 南京 210096);
[4](中国科学技术大学 自动化系,安徽 合肥 230027)
E-mail: q_tao@sohu.com; qing.tao@mail.ia.ac.cn
http://www.ia.ac.cn

摘要: 支持向量机(support vector machine,简称 SVM)是一种基于结构风险最小化原理的分类技术,也是一种新的具有很好泛化性能的回归方法.提出了一种将回归问题转化为分类问题的新思想.这种方法具有一定的理论依据,与 SVM 回归算法相比,其优化问题几何意义清楚明确.

关 键 词: 回归;分类;支持向量;最大边缘
中图法分类号: TP18 文献标识码: A

统计学习理论起源于 20 世纪 60 年代晚期[1,2],但在 1990 年以前,它仅仅是进行函数估计的理论分析工具.到了 90 年代中期,人们提出了理论严谨的结构风险最小化原理,并在此基础上创造性地产生出了一种新的机器学习算法——SVM(support vector machines)[3~5],SVM 的近期发展及成功应用使得统计学习理论已成为研究估计高维函数算法的理论和实用工具.

SVM 学习算法现已成为训练多层感知器、多项式和 RBF 神经网络的替代性方法[6].对线性可分(二分类)情形,SVM 算法最后归结为一个二次规划问题,这个规划问题具有一定的代表性和理论体系统一性.首先对线性不可分问题,只要对规划问题线性可分情形下的约束条件适当松弛,就可得到不可分情形下的线性分类器,这正是软边缘算法[5];而对非线性分类器的设计问题,可通过输入空间到特征空间的非线性映射将其转化为线性可分情形加以解决,而决定非线性分类器的优化问题正是线性可分情形时的适当变形,即将输入空间的欧氏内积变为核函数[6~8].基于结构风险最小化原理的思想同样被成功地应用于函数回归,出现了理论依据更好的回归方法[7].

从神经网络系统理论的发展来看[9],线性可分问题是最基本的,Rosenblatt 感知器(peceptron)的分类算法为三层前馈神经网络能以任意精度逼近 L^2 中的任意函数奠定了理论基础,而这种逼近能力正是前馈网络被广泛应用于建模预测和多种控制问题的理论依据.我们打算遵照前馈神经网络的理论体系对 SVM 进行研究.受 SVM 算法是最大边缘算法的启发,文献[10]对线性可分情形提出一种基于闭凸集间的距离优化的算法,而将线性不可分的情形,通过一种闭凸包收缩的方法,将其归结为线性可分情形.文献[10]的优化问题集可分性判断和解分类超平面于一体,其中支持向量的几何意义非常清晰.

分类问题的样本点明确地属于某一类,而回归问题样本点属于的类别事先是不知道的,这正是分类问题与

* 收稿日期: 2000-09-15; 修改日期: 2001-04-17
基金项目: 国家自然科学基金资助项目(60175023);中国博士后科学基金资助项目(5030436);安徽省自然科学基金资助项目(01042304);安徽省优秀青年基金资助项目

作者简介: 陶卿(1965－),男,安徽长丰人,博士,副教授,主要研究领域为神经网络,支持向量机;曹进德(1963－),男,安徽和县人,博士,教授,主要研究领域为应用数学,神经网络;孙德敏(1939－),男,辽宁新民人,教授,博士生导师,主要研究领域为模式识别与智能系统,控制理论及其应用.

回归问题的区别所在.本文通过对样本点集的适当变换,提出一种将回归问题转化为二分类问题的新思想,从而可用文献[10]的方法求解,一方面这与前馈神经网络的理论体系相一致,另一方面也使得回归问题中支持向量的几何意义更明显,为分类问题的研究成果应用于回归问题奠定了理论基础.

1 SVM 回归方法

本节将简介基于结构风险最小化原理的 Support Vector 回归方法[7,11].

考虑下列线性回归问题:

$$(y_1,x_1),\ldots,(y_l,x_l),\ x_i\in R^n,\ y_i\in R,\ i=1,2,\ldots,l,$$

求回归线性函数

$$f(x)=\langle w,x\rangle+b,$$

其中 $w\in R^n$, $b\in R$.

基于 Support Vector 的最优回归函数是指满足结构风险最小化原理,即极小化

$$\Phi(w)=\frac{1}{2}\|w\|^2+C\cdot R_{emp}[f], \tag{1}$$

其中 C 是预先指定的常数, $R_{emp}[f]$ 是经验风险.

对于 $R_{emp}[f]$,可以采用不同的代价函数来描述,如二次函数、Huber 函数和 ε -insensitive 函数,其中 Vapnik 提出的 ε -insensitive 函数具有很好的性质[7].当回归测度函数为 ε -insensitive 代价函数:

$$|\xi|_\varepsilon=\begin{cases}0 & \text{if } |\xi|_\varepsilon\le\varepsilon\\ |\xi|-\varepsilon & \text{otherwise}\end{cases}$$

时,式(1)可表示为

$$\Phi(w)=\frac{1}{2}\|w\|^2+\frac{1}{l}\sum_{i=1}^{l}|y_i-f(x_i)|_\varepsilon\ . \tag{2}$$

特别地,当 $|y_i-\langle w,x_i\rangle-b|\le\varepsilon$, $i=1,2,\ldots,l$ 满足时,式(2)显然等价于

$$\min\frac{1}{2}\|w\|^2, \tag{3}$$

$$\text{subject to}\ \begin{cases}y_i-\langle w,x_i\rangle-b &\le &\varepsilon\\ \langle w,x_i\rangle+b-y_i &\le &\varepsilon\end{cases}.$$

当优化问题式(2)的约束条件不满足时,它显然是无解的.为了克服这一缺陷,用类似于 Cortes 的松弛方法[5]来处理式(3),此时式(2)变为

$$\min\frac{1}{2}\|w\|^2+C\sum_{i=1}^{l}(\xi_i+\xi_i^*), \tag{4}$$

$$\text{subject to}\begin{cases}y_i-\langle w,x_i\rangle-b &\le &\varepsilon+\xi_i\\ \langle w,x_i\rangle+b-y_i &\le &\varepsilon+\xi_i^*\\ \xi_i,\xi_i^* &\ge &0\end{cases}.$$

松弛回归方法的几何意义如图 1 所示.对优化问题式(4),通过采用数学规划中的对偶方法,可得到最优回归线性函数的 w 和支持向量[7,11].

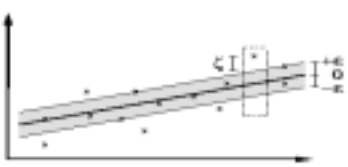

Fig.1 The relaxation method for regression

图 1 松弛回归方法

2 基于分类的回归方法

本节首先将回归问题转化为线性可分情形下的分类问题,进而用文献[10]的方法加以解决.

受图 1 的启发,选取 $\varepsilon>0$,第 1 节中的回归问题显然可以转化为 $Q_1=\{(x_i,y_i+\varepsilon),i=1,...,l\}$ 和 $Q_2=\{(x_i,y_i-\varepsilon),i=1,...,l\}$ 的线性分类问题.下面来分析这种转化的合理性.

首先,当选取的 ε 充分大时,Q_1 和 Q_2 总是线性可分的;其次,当 Q_1 和 Q_2 线性可分时,根据线性可分情形下的 SVM 理论[3,6,7],Q_1 和 Q_2 的最大边缘分类超平面 $\langle\hat{w},z\rangle+\hat{b}=0$ 中的 $\hat{w}=(\hat{w}_1,\hat{w}_2)$ 由下列优化问题决定:

$$\min\frac{1}{2}\|\hat{w}\|^2, \tag{5}$$

$$\text{subject to } \begin{cases}\langle\hat{w},z_i\rangle+\hat{b} & >0 \quad z_i\in Q_1\\ \langle\hat{w},z_i\rangle+\hat{b} & <0 \quad z_i\in Q_2\end{cases}.$$

按照超平面的函数表示习惯,令 $\hat{w}_2=-1$.此时,优化问题式(5)和优化问题式(3)等价,这种等价性表明,当 $|y_i-\langle w,x_i\rangle-b|\le\varepsilon$, $i=1,2,...,l$ 满足时,回归问题和转化后分类问题的解是一致的.

根据以上的分析和文献[10],可用闭凸集间距的方法来求解第 1 节中的回归问题:

$$\begin{cases}\min\|\lambda_1p_1+\lambda_2p_2+...+\lambda_lp_l-\beta_1q_1-\beta_2q_2-...-\beta_lq_l\|^2\\ \lambda_1+\lambda_2+...+\lambda_l=1,\ \beta_1+\beta_2+...+\beta_l=1\\ \lambda_i\ge0,\beta_j\ge0,i=1,2,...,l\end{cases}, \tag{6}$$

其中 $p_i=(x_i,y_i+\varepsilon)$, $q_i=(x_i,y_i-\varepsilon)$, $i=1,2,...,l$.与优化问题式(5)相比,式(6)总是有解的,从解的结果还可以判断 Q_1 和 Q_2 的线性可分性.设 $\lambda_1^*,\lambda_2^*,...,\lambda_l^*$, $\beta_1^*,\beta_2^*,...,\beta_l^*$ 是式(6)的一组解,则 Q_1 和 Q_2 的最大边缘线性分类器为过 $\lambda_1^*p_1+\lambda_2^*p_2+...+\lambda_l^*p_l$ 和 $\beta_1^*q_1+\beta_2^*q_2+...+\beta_l^*q_l$ 连线中点且与这条连线垂直的超平面,可用点法式求得其方程.对应于 $\lambda_1^*,\lambda_2^*,...,\lambda_n^*$, $\beta_1^*,\beta_2^*,...,\beta_m^*$ 中非零数的向量称为相应回归问题的支持向量.

最后,我们讨论一下 ε 的选取问题.首先 ε 不能选取得过大,尽管选取充分大的 ε 可保证 Q_1 和 Q_2 的线性可分,但它同时导致 Q_1 和 Q_2 的范围过大,从而使分类集合 VC 维的上界增大[3,6,7],与结构风险最小化原理相矛盾.如果规划问题式(6)目标函数的最优值为 0,表明 Q_1 和 Q_2 线性不可分,这说明 ε 选取得过小,这时可用文献[10]中闭凸集收缩的方法来解决.

3 方法应用举例

例 1:考虑下列线性回归问题[7].

x	y
1.0	−1.6
3.0	−1.8
4.0	−1.0
5.6	1.2
7.8	2.2
10.2	6.8
11.0	10.0
11.5	10.0
12.9	10.0

取 $\varepsilon=5$,得 $\lambda_5=1$,其余 $\lambda_i=0$, $\beta_1=0.0276$, $\beta_7=0.9724$,其余 $\beta_i=0$. $w=(2.9237,-2.5205)$, $b=-12.1083$.分类结果如图 2 和图 3 所示.

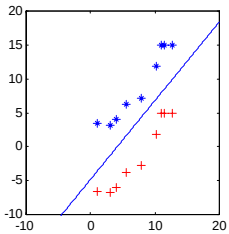

Fig.2 The classification after transformation(ε =5)

图 2 转化后的分类图(ε =5)

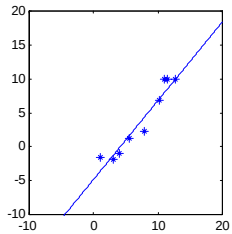

Fig.3 The linear regression function based on classification

图 3 基于分类的线性回归函数

4 结 论

遵照前馈神经网络的理论体系,文献[10]和本文将线性不可分情形下的线性分类器设计问题和回归问题都转化为线性可分条件下的分类问题.转化后的问题支持向量的几何意义明确,并可用文献[10]的方法进行求解.

References:

[1] Cherkassky, V., Mulier, F. Vapnik-Chervonenkis learning theory and its applications. IEEE Transactions on Neural Networks, 1999, 10(5):985~988.

[2] Vapnik, V. An overview of statistical learning theory. IEEE Transactions on Neural Networks, 1999,10(5):988~999.

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

[4] Cortes, C., Vapnik, V. Support vector networks. Machine Learning, 1995,20(1):1~25.

[5] Cortes, C. Prediction of generalization ability in learning machines [Ph.D. Thesis]. Department of Computer Science, University of Rochester, 1995.

[6] Osuna, E.E., Freund, R., Girosi, F. Support vector machines: training and applications. A. I. Memo 1602, MIT Artificial Intelligence Laboratory, 1997.

[7] Gunn, S. Support vector machine for classification and regression. ISIS Report, Image Speech & Intelligent Systems Group, University of Southampton, 1998.

[8] Scholkopf B., Mika, B., Burges, C.J.C., *et al*. Imput space versus feature space in kernel-based methods, IEEE Transactions on Neural Networks. 1999,10(5): 999~1017.

[9] Mehrotha, K., Mohan C.K., Ranka, S. Elements of Artificial Neural Network. Cambridge, MA: MIT Press, 1997.

[10] Tao Qing, Sun De-min, Fan Jin-song, *et al*. The maximal margin linear classifier based on the contraction of the closed convex hull. Journal of Software, 2002,13(3):404~409 (in Chinese).

[11] Smola, A.J. Learning with kernel [Ph.D. Thesis]. Technical University of Berlin, 1998.

附中文参考文献:

[10] 陶卿,孙德敏,范劲松,等.基于闭凸包收缩的最大边缘线性分类器.软件学报,2002,13(3):404~409.

A Regression Method Based on the Support Vectors for Classification*

TAO Qing[1,2], CAO Jin-de[3], SUN De-min[4]

[1](*Institute of Automation*, *The Chinese Academy of Sciences*, *Beijing* 100080, *China*);

[2](1*st Department*, *Artillery Academy of PLA of China*, *Hefei* 230031, *China*);

[3](*Department of Applied Mathematics*, *Southwest University*, *Nanjing* 210096, *China*);

[4](*Department of Automation*, *University of Science and Technology of China*, *Hefei* 230027, *China*)

E-mail: q_tao@sohu.com; qing.tao@mail.ia.ac.cn

http://www.ia.ac.cn

Abstract: The support vector machine is a classification technique based on the structural risk minimization principle, and it is also a class of regression method with good generalization ability. In this paper, a new idea that each regression problem can be changed into a classification problem is presented. The proposed method has some theoretical foundations. Compared with SVM regression method, the geometric meaning of optimization problem in this paper is very clear and obvious.

Key words: regression; classification; support vector machines; maximal margin

* Received September 15, 2000; accepted April 17, 2001

Supported by the National Natural Science Foundation of China under Grant No.60175023; Postdoctoral Science Foundation of China under Grant No.5030436; the Natural Science Foundation of Anhui Province of China under Grant No.01042304; the Excellent Youth Science and Technology Foundation of Anhui Province of China

第 12 届中国计算机学会网络与数据通信学术会议

征 文 通 知

为推动我国在此方向的研究,探讨计算机网络与数据通信技术的发展动态与趋势,促进我国科研人员在此领域的交流与合作,中国计算机学会网络与数据通信专业委员会拟于 2002 年 12 月 2~4 日在武汉举办“第 12 届中国计算机学会网络与数据通信学术会议”.会议由华中师范大学计算机科学系承办,并将邀请该领域的国际知名学者作专题特邀报告.为保证本次会议的学术质量,现向全国科技工作者公开征稿.征稿范围包括计算机通讯网络理论与工程的各个方面.本次会议的论文将结辑出版优秀论文将由计算机学会推荐给有关核心期刊发表.

一、征文要求

(1) 论文应是未公开发表过,一般不超过 6000 字;

(2) 全文电子邮件投稿,要求 Word2000 兼容的电子文档,所有内容放于一个文件中;

(3) 编排格式:

标题: 居中,2 号黑体;作者: 居中,4 号仿宋;作者地址: 5 号楷体;摘要、关键词: 5 号楷体;正文:5 号宋体,分节标题 4 号;参考文献: 小 5 号宋体.

(4) 投稿地址:华中师范大计算机科学系谭连生教授收 E-mail: L.Tan@ccnu.edu.cn

二、重要日期

论文提交截止日期: 2002 年 8 月 15 日 论文接收通知日期: 2002 年 10 月 1 日 会议注册日期: 2002 年 12 月 2 日

联系人:谭连生教授 湖北省武汉市华中师范大学计算机科学系(430079)

电话: 027–87673277 传真: 027–87876070 E-mail: L.Tan@ccnu.edu.cn

0871~1028