加权支持向量回归的权值确定方法

吴金花,孙德山
(辽宁师范大学 数学学院,辽宁 大连 116029)

摘 要:针对标准支持向量回归中由于噪声和野点造成的回归误差,提出了一种基于线性规划的权值确定方法。该方法的基本思想是首先根据样本偏离数据域距离的不同,采用线性规划下的一类分类算法得到一个权值确定函数,然后将得到的权值确定函数应用于加权支持向量回归,加权的目的是为了减弱噪声和野点对回归结果的影响。实验表明,该权值确定方法与标准支持向量回归相比,可以有效减小回归误差,提高支持向量机抗噪声的能力。

关键词:支持向量回归;线性规划;权值
中图分类号:TP181 **文献标识码**:A **文章编号**:1673-629X(2009)06-0135-03

An Approach of Confirming Weight Values for Reweighted Support Vector Regression

WU Jin-hua, SUN De-shan
(Institute of Mathematics, Liaoning Normal University, Dalian 116029, China)

Abstract: In order to overcome the error of regression accused by noise and outliers in standard support vector regression, an approach of confirming weight values based on linear programming is presented. The essential idea of the approach is to reduce the influence caused by noise and outliers in support vector regression. Firstly, the weight value of each input sample is confirmed according to its distance to database by linear programming. The proposed method is applied to reweighted support vector regression. Experimental results show that this approach compared with standard support vector regression can reduce the error of regression effectively, and improve the anti-noise capability of SVM.

Key words: support vector regression; linear programming; weight value

0 引 言

支持向量机(Support Vector Machine, SVM)[1,2]是20世纪90年代Boser, Guyon, Vapnik提出的一种基于统计学习理论[3,4]的机器学习算法,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,在国内外学术界日益受到广泛重视。支持向量机的核心思想是将结构风险最小化原则引入到分类中,在属性空间中构建最优分类超平面,使分类器得到全局最优解。它是从线性可分情况下的最优超平面发展而来的。对于线性不可分的情况,可以通过非线性映射将低维空间的样本映射到高维特征空间,从而转化为线性可分的情况。

支持向量回归是Vapnik在定义了ε-不敏感损失函数的基础上提出来的支持向量回归(简称ε-SVR)算法。在ε-SVR中,输入样本被等同对待,每个样本的松弛项被赋予相同的惩罚因子,所以当样本中存在噪声和野点时,导致SVM在这些点较为敏感,由此产生过拟合现象。

针对这种情况,2002年Lin C F等将模糊隶属度的概念引入到SVM分类中,提出了模糊支持向量机(fuzzy support vector machine, FSVM)[5,6]的概念。文献[7]针对支持向量机中由于噪声和孤立点带来的过拟合问题,提出了一种基于数据域描述的模糊隶属度函数模型,根据样本到特征空间最小超球球心的距离确定其模糊隶属度。文献[8]给出一种奇异值软剔除的加权稳健支撑向量回归方法(WRSVR)。

笔者在此引入权值,根据每个样本偏离数据域程度的不同赋予不同的权值,使噪声点的权值接近于一个很小的实数,以减小对回归函数的影响。在确定权值的训练中,采用线性规划下的一类分类方法。实验证明,该方法减小了回归误差,提高了SVM的抗噪能力。同时在一类分类中,线性规划比二次规划节省很多时间。

收稿日期:2008-09-22;**修回日期**:2008-12-03
基金项目:辽宁省高等学校科研项目资助(2008343)
作者简介:吴金花(1982-),女,河北唐山人,研究方向为统计学习;孙德山,博士,研究方向为应用统计学。

1 加权支持向量回归

设给定的训练样本为：

$$\{(x_i,y_i),i=1,2,\cdots,l\}$$

其中，$x_i \in R^N$ 为输入值，$y_i \in R$ 为对应的目标值，l 为训练样本个数。

支持向量回归[9] 的基本思想是寻找一个从输入空间到输出空间的一个非线性映射 $\phi(x):R^N \to H$，将输入数据 x 映射到高维特征空间 H 中，采用适当的核函数 $K(x_i,x_j)$ 代替高维特征空间中的向量内积 $<\phi(x_i),\phi(x_j)>$，并在特征空间中用下式来寻求最优回归函数：

$$f(x)=<w,\phi(x)>+b \tag{1}$$

其中，w,b 分别为回归函数的权重和偏置。

基于支持向量机的最优回归函数是指满足结构风险最小化原理，即极小化优化问题是最小化下面的函数：

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

约束为：

$$f(x_i)-y_i \leqslant \varepsilon_i+\varepsilon,i=1,2,\cdots,l \tag{3}$$

$$y_i-f(x_i) \leqslant \xi_i^*+\varepsilon,i=1,2,\cdots,l \tag{4}$$

$$\xi_i,\xi_i^* \geqslant 0,i=1,2,\cdots,l \tag{5}$$

其中，第一项使函数更为平坦，从而提高泛化能力，第二项为减小误差，常数 c 为惩罚系数，对两者做出折中。ε 为一正常数，控制回归精度。

引入拉格朗日函数，得到优化问题的对偶形式为：

$$\max -\frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)K(x_i,x_j)+\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)y_i-\sum_{i=1}^{l}(\alpha_i+\alpha_i^*)\varepsilon \tag{6}$$

约束为：

$$\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)=0 \tag{7}$$

$$0 \leqslant \alpha_i,\alpha_i^* \leqslant \mu_i c,i=1,\cdots,l \tag{8}$$

解这个二次优化问题，得到回归函数 $f(x)$ 的表达式为：

$$f(x)=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)K(x_i,x)+b \tag{9}$$

2 权值的确定

权值的确定采用线性规划下的一类分类算法。一类分类的基本思想是对给定的样本点集 $\{x_i,i=1,\cdots,l\}$，用一个非线性映射将样本点映射到高维特征空间，在高维空间中找一个超平面，使之以尽可能大的距离 ρ 将尽可能多的样本从原点分离开，即估计一个决策函数 $f(x)=\sum_{i=1}^{n}\alpha_i k(x_i,x)$，当一个样本 x 满足 $f_w(x) \geqslant \rho$ 时，它被确定属于该类。为了获得 w 和 ρ 的值，并根据结构风险最小化原则，将问题归结为下面的优化：

$$\min \frac{1}{2}\|w\|_2^2-\rho+C\sum_{i=1}^{l}\xi_i \tag{10}$$

约束为：

$$<w,\phi(x_i)> \geqslant \rho-\xi_i,\xi_i \geqslant 0,i=1,\cdots,l \tag{11}$$

将优化问题化为对偶形式：

$$\min \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j K(x_i,x_j) \tag{12}$$

约束为：

$$0 \leqslant \alpha_i \leqslant C,i=1,\cdots,l \tag{13}$$

$$\sum_{i=1}^{l}\alpha_i=1 \tag{14}$$

解出 α 值后，可得决策函数：

$$f(x)=\sum_{i=1}^{l}\alpha_i K(x_i,x) \tag{15}$$

以上是二次规划下的一类分类算法。线性规划下的一类分类算法是在目标函数(10) 中采用 l_∞ - 范数代替 l_2 - 范数，文献[10] 中给出了优化函数：

$$\min -\rho+C\sum_{i=1}^{l}\xi_i \tag{16}$$

约束为：

$$\sum_{i=1}^{l}\alpha_i K(x_i,x_j)=\rho-\xi_i,j=1,\cdots,l \tag{17}$$

$$\sum_{i=1}^{l}\alpha_i=1 \tag{18}$$

$$\alpha_i,\xi_i \geqslant 0,i=1,\cdots,l \tag{19}$$

由此定义权值如下：

$$\mu_i=\begin{cases}(1-\dfrac{f_{\max}-f(x_i)}{f_{\max}-f_{\min}})^2+\sigma & f_{\min} \leqslant f(x_i)<\rho \\ 1-\dfrac{f_{\max}-f(x_i)}{f_{\max}-f_{\min}} & \rho \leqslant f(x_i) \leqslant f_{\max}\end{cases} \tag{20}$$

其中，$f_{\max}=\max(f(x_i)\mid x_i \in X)$，$f_{\min}=\min(f(x_i)\mid x_i \in X)$，$\sigma<1$，为足够小的正实数。

从以上定义可以看出，当 $\rho \leqslant f(x_i) \leqslant f_{\max}$ 时，表示 x_i 是区域内的样本，在支持向量回归中，样本在回归间隔附近；当 $f_{\min} \leqslant f(x_i)<\rho$ 时，表示 x_i 是区域外的样本，其权值接近于一个非常小的实数 σ，这样可以减小这些点对回归函数的影响。

3 实验分析

取 $x \in [-2,2]$，其中间隔为 0.1，因变量 $y=$

$\mathrm{sinc}(x)$,然后在因变量的前 10 个样本中加入噪声 $N(0,0.8)$,其余样本中加入噪声 $N(0,0.1)$,如图 1 所示。分别用标准的支持向量回归和加权支持向量回归建立预测模型。

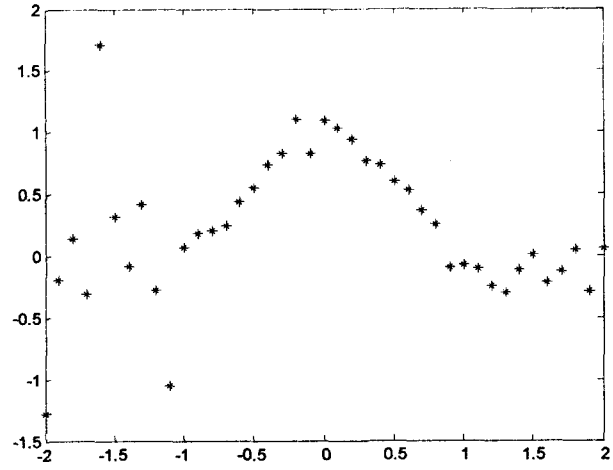

图 1　加入噪声的样本

首先对输入样本进行训练,取 $c = 1/15, \sigma^2 = 0.12$,利用线性规划得到 $\rho = 0.0931, f_{\max} = 0.1377, f_{\min} = 2.8616e-008$,则确定权值为:

$$\mu_i = \begin{cases} (1 - \dfrac{0.1377 - f(x_i)}{0.1377 - 2.8616e-008})^2 + 0.0005 \\ 2.8616e-008 \leqslant f(x_i) < 0.0931 \\ 1 - \dfrac{0.1377 - f(x_i)}{0.1377 - 2.8616e-008} \\ 0.0931 \leqslant f(x_i) \leqslant 0.1377 \end{cases} \tag{21}$$

进行 ε - SVR 训练,取 $c = 5, \sigma^2 = 1.8$,在 -2 到 2 之间取间隔为 0.13 的样本为测试样本(不同于训练样本),测试指标采用均方误差:

$$\mathrm{MSE} = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(y_i - \hat{y}_i)^2} \tag{22}$$

其中,y_i 为实际值,$\hat{y}_i$ 为预测值,k 为测试样本的数量。

标准支持向量回归的结果 MSE = 0.3530,加权支持向量回归的结果 MSE = 0.0858,说明当数据中存在噪声时,加权支持向量回归算法得到的预测值更接近真实值。

4　结束语

标准的支持向量回归在样本中无噪声点时,具有很好的学习和泛化能力。但当样本中存在噪声和野点时,回归间隔偏向噪声点移动,从而出现过拟合现象,产生误差。

文中通过引入权值,减小了噪声点的影响。在确定权值时,考虑了样本离决策超平面的距离,对不同的样本采用不同的公式计算其权值。实验表明,该方法与标准支持向量回归相比,减小了回归误差,提高了支持向量机的抗噪能力。

参考文献:

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

[2] 张　楠,范玉妹.关于支持向量机几何算法的研究[J].计算机技术与发展,2007,17(1):142 - 144.

[3] 张学工.统计学习理论的本质[M].北京:清华大学出版社, 2000.

[4] 杨　斌,陆　游.基于统计学习理论的支持向量机的分类方法[J].计算机技术与发展,2006,16(11):56 - 58.

[5] Lin C F, Wan S D. Fuzzy support vector machine[J]. IEEE Tran. on Neural Network, 2002, 13(2): 464 - 471.

[6] Huang H P, Lin Y H. Fuzzy support vector machine for pattern recognition and data mining[J]. Int'l Journal of Fuzzy Systems, 2002, 4(3): 826 - 835.

[7] 张　英,苏宏业,褚　建.基于数据域描述的模糊支持向量机回归[J].信息与控制,2005,34(1):1 - 6.

[8] 张讲社,郭　高.加权稳健支撑向量回归方法[J].计算机学报,2005,28(7):1171 - 1177.

[9] Bao YuKun, Zhang Rui. Fuzzy support vector machines regression for business forecasting: An application[J]. Computer science, 2006(9): 1313 - 1317.

[10] 孙德山.支持向量机分类与回归方法研究[D].长沙:中南大学,2004.

(上接第 134 页)

[2] van der Aalst W, van Hee K. Workflow Management[M]. [s.l.]: The MIT Press, 2002.

[3] 周建涛, 史美林, 叶新铭.一种基于 Petri 网化简的工作流过程语义验证方法[J].软件学报,2005,16(7):1242 - 1251.

[4] 雷丽晖,段振华.一种基于扩展有限自动机验证组合 Web 服务的方法[J].Journal of Software, 2007,18(12):2980 - 2990.

[5] 单卓为,鱼　滨.基于 SPIN 的 CSCW 系统的验证[J].计算机技术与发展,2008,18(4):9 - 12.

[6] 范玉顺.工作流管理技术基础[M].北京:清华大学出版社,2001.

[7] Anderson A J. Data Flow Systems, In Multiple Processing: A System's Overview[M]. [s.l.]: [s.n.], 1989:441 - 488.

[8] Laprie J, Randell B, Landwehr G. Basic Concepts and Taxonomy of Dependable and Secure Computing[J]. IEEE Transactions on Dependable and Secure Computing, 2004, 1(1): 11 - 33.