

加权支持向量回归算法^{*}

孙德山^{1,2} 吴今培³ 侯振挺¹ 肖健华³

(中南大学铁道校区科学研究所 长沙410075)¹ (辽宁师范大学数学系 大连116029)²

(五邑大学智能技术与系统研究所 广东江门529020)³

Weighting Support Vector Regression Algorithm

SUN De-Shan^{1,2} WU Jin-Pei³ HOU Zhen-Ting¹ XIAO Jian-Hua³

(Research Department of Railway Compus, Central South University, Changsha, 410075)¹

(Department of Mathematics, Liaoning Normal University, Dalian, 116029)²

(Institute of Intelligent Technology & System, Wuyi University, Guangdong, Jiangmen, 529020)³

Abstract Support vector regression machines based on structural risk minimization have a good generalization performance. However, its effect is not good if there exists heterogeneity of variance in the regression models. In order to solve the problem, a kind of weighting support vector regression is proposed in this paper. The results of simulation experiments show the feasibility and effectiveness of the method.

Keywords Support vector machine, Regression, Kernel function, Heterogeneity of variance

1 引言

Vapnik 等人根据统计学习理论提出的支持向量机学习方法^[1],近年来受到了国际学术界的广泛重视。支持向量机的最大特点是根据 Vapnik 结构风险最小化原则,尽量提高学习机的泛化能力,即由有限的训练集样本得到的小的误差能够保证对独立的测试集仍然保持小的误差。另外,由于支持向量机算法是一个凸优化问题,因此局部最优解一定是全局最优解。这些特点是其它学习算法,如神经网络学习算法所不及的。由于其出色的性能,支持向量机已广泛应用于模式分类和回归分析中^[2~5]。在支持向量回归中,有个基本假设,就是样本来自于同一分布,也就是说随机误差项应该具有相同的方差,独立或不相关。但由于实际问题是错综复杂的,因而在建立实际问题的回归分析模型时,经常会出现某一因素随着解释变量观测值的变化而对被解释变量产生不同的影响,导致随机误差项产生不同方差。当一个回归问题存在异方差性时,如果仍然用标准的支持向量回归来建立模型,往往达不到好的效果。本文提出加权支持向量回归算法来解决这个问题,文章最后给出了一个实例,用标准回归方法和加权回归方法进行预测分析,结果表明加权支持向量回归算法的预测结果优于标准支持向量回归算法。

2 加权支持向量回归

首先考虑线性的情况,设给定的训练样本为 $\{(x_1, y_1), \dots, (x_l, y_l)\} \subset R^n \times R$,并设线性回归模型为:

$$f(x_i) = \langle w, x_i \rangle + b + \varepsilon_i \quad i=1, \dots, l$$

其中 $x_i \in R^n, y_i \in R, i=1, \dots, l$ 。方差具有下面的性质,

$$\begin{cases} E(\varepsilon_i) = 0, i=1, \dots, l \\ \text{cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} \sigma_i^2, i=j \\ 0, i \neq j \end{cases} \end{cases}$$

标准的支持向量回归是在假设所有 σ_i^2 都相同的情况下,最小化下面式子:

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

条件为

$$\begin{aligned} \langle w, x_i \rangle + b - y_i &\leq \xi_i^* + \varepsilon, \quad i=1, \dots, l \\ y_i - \langle w, x_i \rangle - b &\leq \xi_i + \varepsilon, \quad i=1, \dots, l \\ \xi_i, \xi_i^* &\geq 0, \quad i=1, \dots, l \end{aligned}$$

在 σ_i^2 不相同的情况下,上面的优化式子就会出现一定的偏差,因为上面式子中各个 ξ_i, ξ_i^* 地位是相同的,也就是说对误差的考虑是相同的尺度,这就导致回归线被拉向方差大的项,而方差小的项的拟合程度就差。为了解决这个问题,我们引入一个适当的权数 λ_i ,以调整各项在回归中的作用。变换式(1)为下面的形式:

$$\frac{1}{2} \|w\|^2 + C \sum_{i=1}^l \lambda_i (\xi_i + \xi_i^*) \quad (2)$$

优化条件没有变。其中的权数应该是观测误差项方差的倒数,即

$$\lambda_i = \frac{1}{\sigma_i^2}$$

在实际应用中,根据情况各个权数按上面式子成比例选取。

引入拉格朗日函数

$$\begin{aligned} L(w, b, \xi, a, a^*, \gamma) &= \frac{1}{2} w \cdot w + C \sum_{i=1}^l \lambda_i (\xi_i + \xi_i^*) - \sum_{i=1}^l a_i [\xi_i \\ &\quad + \varepsilon - y_i + \langle w, x_i \rangle + b] - \sum_{i=1}^l a_i^* [\xi_i^* + \varepsilon + y_i - \langle w, x_i \rangle - \\ &\quad b] - \sum_{i=1}^l (\gamma_i \xi_i + \gamma_i^* \xi_i^*) \end{aligned} \quad (3)$$

其中 $a_i, a_i^*, \gamma_i \geq 0, i=1, \dots, l$

函数 L 的极值应满足条件

^{*} 广东自然科学基金资助(021349),孙德山 博士研究生,研究方向:支持向量机理论、时间序列预测,吴今培 教授,博士生导师,研究方向:人工智能和电子技术等研究,侯振挺 教授,博士生导师,研究方向:马氏过程理论及数理金融等,肖健华 博士,研究方向:人工智能及故障诊断等。

$$\frac{\partial}{\partial w} L = 0, \frac{\partial}{\partial b} L = 0, \frac{\partial}{\partial \xi_i} L = 0, \frac{\partial}{\partial \xi_i^*} L = 0$$

得到下面的式子

$$w = \sum_{i=1}^l (a_i - a_i^*) x_i$$

$$\sum_{i=1}^l (a_i - a_i^*) = 0$$

$$C\lambda_i - a_i - \gamma_i = 0, i = 1, \dots, l$$

$$C\lambda_i - a_i^* - \gamma_i^* = 0, i = 1, \dots, l$$

将以上4式代入式(3),得到优化问题的对偶形式,最大化函数:

$$W(a, a^*) = -\frac{1}{2} \sum_{i,j=1}^l (a_i - a_i^*)(a_j - a_j^*) \langle x_i, x_j \rangle + \sum_{i=1}^l (a_i - a_i^*) \gamma_i - \sum_{i=1}^l (a_i + a_i^*) \varepsilon$$

其约束为

$$\sum_{i=1}^l (a_i - a_i^*) = 0$$

$$0 \leq a_i, a_i^* \leq \lambda C, i = 1, \dots, l$$

这也是一个二次优化问题, w 可由下式给出:

$$w = \sum_{i=1}^l (a_i - a_i^*) x_i$$

于是 $f(x)$ 可表示为:

$$f(x) = \sum_{i=1}^l (a_i - a_i^*) \langle x_i, x \rangle + b$$

对于非线性回归,只要用 $K(x_i, x_j)$ 代替 $\langle x_i, x_j \rangle$ 就可以了。此时的优化函数为:

$$W(a, a^*) = -\frac{1}{2} \sum_{i,j=1}^l (a_i - a_i^*)(a_j - a_j^*) K(x_i, x_j) + \sum_{i=1}^l (a_i - a_i^*) \gamma_i - \sum_{i=1}^l (a_i + a_i^*) \varepsilon$$

约束条件同上面。函数 $f(x)$ 可如下表示:

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

按照 Kuhn-Tucker 定理,在鞍点有

$$a_i [\xi_i + \varepsilon - y_i + f(x_i)] = 0, i = 1, \dots, l$$

$$a_i^* [\xi_i^* + \varepsilon + y_i - f(x_i)] = 0, i = 1, \dots, l$$

$$(C\lambda_i - a_i) \xi_i = 0, i = 1, \dots, l$$

$$(C\lambda_i - a_i^*) \xi_i^* = 0, i = 1, \dots, l$$

由上式可知,对应于 $a_i \in (0, C\lambda_i)$ 或 $a_i^* \in (0, C\lambda_i)$ 的 $f(x_i)$

与 y_i 的误差等于 ε , 因此有:

$$\varepsilon - y_i + f(x_i) = 0, \text{ 对于 } 0 < a_i < C\lambda_i$$

$$\varepsilon + y_i - f(x_i) = 0, \text{ 对于 } 0 < a_i^* < C\lambda_i$$

由以上两式可以求出 b 。

3 算法

将上面的优化问题化为下面的标准二次优化形式:

$$\min_x \frac{1}{2} x^T H x + c^T x$$

其约束为

$$a^T x = 0$$

$$0 \leq x \leq C\lambda$$

根据本文的优化问题,其中的参数取下面的值:

$$x = \begin{bmatrix} a \\ a^* \end{bmatrix}, c = \begin{bmatrix} \varepsilon - y \\ \varepsilon + y \end{bmatrix}, H = \begin{bmatrix} D & -D \\ -D & D \end{bmatrix},$$

$$D = K(x_i, x_j), i = 1, \dots, l, j = 1, \dots, l$$

$$a = [1 \ 1 \ \dots \ 1 \ -1 \ -1 \ \dots \ -1]^T, \lambda = [\lambda_1 \ \dots \ \lambda_l \ \lambda_1 \ \dots \ \lambda_l]^T$$

在实际问题的研究中,误差项的方差 σ^2 通常是未知的,这就要根据实际情况来确定。关于异方差性的检验,统计学家们进行了大量的研究,提出的诊断方法已有多种,常用的有残差图分析法与等级相关系数检验法等。

4 举例

以一个经常被采用的例子(toy)来说明本文方法的有效性。取自变量 $x \in [-1, +1]$, 其中间隔为 0.1, 因变量 $y = \text{sinc}(x)$, 然后在因变量 y 的前 10 个样本中加入噪声 $N(0, 0.8)$, 其余样本中加入噪声 $N(0, 0.1)$ 。分别用传统支持向量回归和加权支持向量回归建立预测模型, 两类支持向量回归中的参数都取为 $\varepsilon = 0.0001, C = 5$, 核函数为 $K(x_i, x_j) = \exp(-\frac{\|x_i - x_j\|^2}{2v^2})$, 本文取 $v^2 = 0.75$ 。加权支持向量回归中的加权值

取前 10 个为 $1/8$, 其余为 1, 然后在 -1 到 $+1$ 之间任选 10 个测试样本(不同于训练样本), 测试指标采用均方误差: $(\text{MSE}) =$

$\sqrt{\frac{1}{K} \sum_{i=1}^K (y_i - \hat{y}_i)^2}$, 其中 y_i 为实际值, $\hat{y}_i$ 为预测值, K 为测试样本的数量, 这里 y_i 取未加入噪声的真实数据。用 Matlab 6.1 编程实现, 标准的支持向量回归算法预测结果是 $\text{MSE} = 0.1319$, 而加权回归算法预测结果是 $\text{MSE} = 0.0303$ 。这就说明了当数据中存在异方差性时, 采用加权支持向量回归算法所获得预测值更接近真实值。

结论 近年来, 支持向量机理论得到了广泛的发展和应, 尤其在模式识别中取得了很好效果^[4], 在很多问题上它都有着其他统计学习技术所难以比拟的优越性, 并在很多领域都得到了成功的应用, 如人脸检测、手写体数字识别、文本自动分类等。另外支持向量机回归算法在函数逼近和时间序列预测中也取得了很好的效果^[3]。本文针对实际回归问题中经常出现异方差性, 而标准的支持向量回归算法在处理这类问题时一般效果不佳, 提出了一种加权支持向量回归算法。仿真实验说明了所给方法的可行性和有效性。

参考文献

- 1 Vapnik V N. Statistical Learning Theory [M]. New York, Wiley, 1998
- 2 Smola A J, Scholkopf B. A tutorial on support vector regression [R]. [NeuroCOLT TR NC-TR-98-030]. Royal Holloway College University of London, UK, 1998
- 3 Fernandez R. Predicting time series with a local support vector regression machine [C]. IN ACAI 99, 1999
- 4 Burges C J C. A Tutorial on Support Vector Machines for Pattern Recognition [R]. Knowledge Discovery and Data Mining, 1998, 2 (2)
- 5 Cortes C, Vapnik V. Support-vector networks [J]. Machine Learning, 1995, 20(3): 273~297