改进的 LS-SVM算法及在交通流量预测上的应用

张朝元[1], 陈　丽[2]

(1. 大理学院 数学与计算机学院, 云南 大理 671003; 2. 大理学院 物理与电子信息学院, 云南 大理 671003)

摘要: 对标准的 LS-SVM 算法进行了改进, 得到一种新的学习算法. 这种新的学习算法不仅能减少计算的复杂性, 提高学习速度; 同时能提高函数估计的精确度. 将改进的 LS-SVM 算法应用于交通流量的预测, 同时与传统的多元线性回归及支持向量机方法进行比较, 结果表明改进的 LS-SVM 方法具有较高的预测精度, 且实验取得了较好效果.

关键词: SVM法; LS-SVM法; 多元线性回归; 交通流量; 预测

中图分类号: TP391　　文献标识码: A　　文章编号: 1007-855X(2008)06-0072-04

Improved Algorithm of LS-SVM and Its Application to Traffic Flow Prediction

ZHANG Chao-yuan[1], CHEN Li[2]

(1. Department of Mathematics and Computer, Dali University, Dali, Yunnan 671003, China;
2. Department of Physics and Electronic Information, Dali University, Dali, Yunnan 671003, China)

Abstract: Based on the traditional least squares support vector machine for function estimation, an improved algorithm of LS-SVM is presented in this paper. The proposed method can reduce the computation complexity and increase the learning speed. The accuracy of estimation is also improved. Through its application to traffic flow prediction and through comparison with the traditional multiple linear regression and support vector machine, this method is proved to be more effective and precise.

Key words: SVM; LS-VSM; multiple linear regression; traffic flow; prediction

0 引 言

Vapnik在文献[1]、[2]里提出了非线性函数估计的支持向量机器法 (Support Vector Machine, 简记为 SVM法). SVM是神经网络领域中的一种新的重要的方法. 近年来许多学者对此作了许多深入的研究, 不仅形成了一套较为完备的理论与方法, 而且其应用也是日趋广泛. SVM法目前被广泛地应用于模式识别、函数估计、回归分析、金融时间序列预测等各类问题. SVM法解决了传统神经网络可能得到问题的局部极小点解和神经网络的隐层数的确定问题, 它将一个具体问题转化为一个带不等式约束的二次凸规划问题, 为了解决这个凸规划问题, SVM法将其放到其对偶空间进行处理.

近年来, 最小二乘支持向量机器法 (Least Squares Support Vector Machine, 简记为 LS-SVM法) 被提出来并被用于分类问题[1,2]和非线性函数的估计问题[3], LS-SVM法将 SVM法中的不等式约束修改为等式约束, 并将误差平方和 (Sum Squared Error, 简记为 SSE) 损失函数作为训练集的经验损失. 但是, LS-SVM法有其自身的缺点[3], 文献[3]提出了一种修正的 LS-SVM法以克服其缺点. 尽管 LS-SVM法有缺点, 但用于函数估计仍然是一种十分可行的方法.

然而, 用于函数估计的标准 LS-SVM 在求解大规模问题时存在学习速度过慢的问题. 因此, 如何减少计算时间和存储空间成为用于函数估计的 LS-SVM 学习算法的研究热点[4]. 由于 O. L. Mangasari等人提

收稿日期: 2008-09-08. 基金项目: 大理学院科研基金资助项目 (项目编号: 2005X23).

第一作者简介: 张朝元 (1978-), 男, 硕士, 讲师. 主要研究方向: 神经网络和统计学习理论.

E-mail: zcy_km@163.com

出的用于模式识别的 SOR(Successive Over relaxation for Support Vector Machine)算法[5]适合迭代求解并能用于解决大规模问题.因此,本文考虑将这一方法推广到函数估计问题中,对用于函数估计的 LS-SVM算法的优化式加以改造,得到了一种函数估计的 LS-SVM的改进算法.这一新的算法具有能减少计算复杂性、提高学习速度和在一定程度上能提高回归估计的精度性等方面的优点.

1 标准的 LS-SVM函数估计

函数估计问题是一类很古老的问题.设被估计的未知函数为 $f(x)$,由 $f(x)$ 给定数据点为 $(x_1, y_1), \cdots, (x_N, y_N)$,其中 $x_k \in R^n$ 为输入向量,$y_k \in R$ 为输出变量,且 $y_k = f(x_k)$, $k = 1, \cdots, N$.函数估计问题就是在这些给定的数据上寻找一个函数 $f_N(x)$,使 $f_N(x)$ 与 $f(x)$ 能尽可能地“接近”.

函数估计问题最终就是求解待估计的未知函数 $f(x)$.作非线性映射:$\varphi: R^n \to H$,其中 φ 称为特征映射,H 为特征空间(一般,H 为高维空间或无穷维空间),则被估计函数 $f(x)$ 有如下形式:$y = f(x) = w^T \varphi(x) + b$,其中 w 为空间 H 中的权向量,$r \in R$ 为偏值.于是,LS-SVM法估计非线性函数为如下特征空间中的最优问题:

$$\min_{w, b, e} J(w, e) = \frac{1}{2}w^T + \frac{1}{2}\gamma\sum_{k=1}^{N} e_k^2 \qquad \text{s.t.} \qquad y_k = w^T\varphi(x_k) + b + e_k \tag{1}$$

其中 $e_k \in R$ 为误差变量.注意到 $J(w, e)$ 是由正则化项 $\frac{1}{2}w^T w$ 和 SSE项 $\frac{1}{2}\sum_{k=1}^{N} e_k^2$ 组成,其中 γ 为一实值常量,它决定了二者的相对重要性.

一般地,由于 w 可能为无限维的,于是直接计算规划(1)是非常困难的,因此将这一规划问题转化到其对偶空间中.定义 Lagrange函数:

$$L(w, b, e; \alpha) = J(w, e) - \sum_{k=1}^{N}\alpha_k[w^T\varphi(x_k) + b + e_k - y_k] \tag{2}$$

其中 $\alpha_k \in R$ 为 Lagrange乘子.于是最优解的条件如下:

$$\begin{cases} \dfrac{\partial L}{\partial w} = 0 \to w = \sum_{k=1}^{N}\alpha_k\varphi(x_k) \\ \dfrac{\partial L}{\partial b} = 0 \to \sum_{k=1}^{N}\alpha_k = 0 \\ \dfrac{\partial L}{\partial e_k} = 0 \to \alpha_k = \gamma e_k, \ k = 1, \cdots, N \\ \dfrac{\partial L}{\partial \alpha_k} = 0 \to y_k = w^T\varphi(x_k) + b + e_k, \ k = 1, \cdots, N \end{cases} \tag{3}$$

这些条件除了 $\alpha_k = \gamma e_k$ 之外,与标准的 SVM最优条件很相似.其中 $\alpha_k = \gamma e_k$ 使得 LS-SVM不再具有 SVM所具有的稀疏性.

利用(3)消 w 去 e_k 与得规划(1)的解的方程:

$$\begin{bmatrix} 0 & (q_N)^T \\ 1_N & \Omega + \frac{1}{\gamma}I \end{bmatrix}\begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ Y \end{bmatrix} \tag{4}$$

其中向量 $1_N = (1, 1, \cdots, 1)^T$, $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_N)^T$, $Y = (y_1, y_2, \cdots, y_N)^T$, Ω 为一矩阵,其定义为:$\Omega = (\Omega_{ij})_{N\times N}$,其中 $\Omega_{ij} = \varphi(x_i)^T\varphi(x_j)$.

注意到(4)为一线性方程组,利用(4)可求得 α 与 b 的值,于是获得被估计函数 $f(x)$ 的表达式为:

$$y = f(x) = \sum_{k=1}^{N}\alpha_k K(x, x_k) + b \tag{5}$$

其中 $K(x, y)$ 为核函数,其定义为 $K(x, y) = \varphi(x)^T\varphi(Y)$.一般地,核函数 $K(x, y)$ 应满足 Mercer条件.

2 改进的 $LS-SVM$ 函数估计

本文将 O. L. Mangasari等人提出的用于模式识别的 SOR 算法推广到函数估计问题中, 对目标函数(1)进行简单的改进得到如下优化问题:

$$\min_{w,b,e} J(w,b,e)=\frac{1}{2}w^Tw+\frac{1}{2}b^2+\frac{1}{2}\gamma\sum_{k=1}^{N}e_k^2 \quad \text{s. t.} \quad y_k=w^T\varphi(x_k)+b+e_k \tag{6}$$

将该规划问题转化到其对偶空间中, 定义 Lagrange函数:

$$L(w,b,e,\alpha)=J(w,b,e)-\sum_{k=1}^{N}\alpha_k[w^T\varphi(x_k)+b+e_k-y_k] \tag{7}$$

其中 $\alpha_k\in R$ 为 Lagrange乘子. 于是最优解的条件如下:

$$\begin{cases}\dfrac{\partial L}{\partial w}=0\rightarrow w=\sum_{k=1}^{N}\alpha_k\varphi(x_k)\\ \dfrac{\partial L}{\partial b}=0\rightarrow b=\sum_{k=1}^{N}\alpha_k\\ \dfrac{\partial L}{\partial e_k}=0\rightarrow \alpha_k=\gamma e_k,\ k=1,\cdots,N\\ \dfrac{\partial L}{\partial \alpha_k}=0\rightarrow y_k=w^T\varphi(x_k)+b+e_k,\ k=1,\cdots,N\end{cases} \tag{8}$$

利用(8)消去 w、e_k 及 b 得到规划(6)的解的方程:

$$(\Omega+\frac{1}{\gamma}E)\alpha=Y \tag{9}$$

其中向量 $\alpha=(\alpha_1,\alpha_2,\cdots,\alpha_N)^T$, $Y=(y_1,y_2,\cdots,y_N)^T$, Ω 为一矩阵, 其定义为: $\Omega=(\Omega_{ij})_{N\times N}$, 其中 $\Omega_{ij}=\varphi(x_i)^T\varphi(x_j)+1$.

注意到(9)为一线性方程组, 利用(9)可求得 α 的值, 于是获得被估计函数 $f(x)$ 的表达式为:

$$y=f(x)=\sum_{k=1}^{N}\alpha_k[K(x,x_k)+1] \tag{10}$$

可以看出, 改进得到的方程组(9)式明显比改进前的方程组(4)要简单, 而且很容易求解, 在某种意义上就是提高了速度, 减少了复杂性. 同时, 得到了估计函数表达式(10)与(5)比较也可以看出, 估计函数也变得简单明了.

3 应用实例

所谓预测就是利用过去和当前的观测值去估计未来值, 这实际上是基于这样一个假设, 即未来值与过去值存在某种确定的函数关系. 所以预测的目的就是试图寻找一个函数以确定未来值与过去值之间的关系, 也就是说预测问题与函数逼近和估计问题在本质上是等价的. 本文将选择利用交通流量来进行预测.

路段上的交通流量与前几个时段的交通流量有着必然的联系, 同时路段是路网中的一个部分, 路段的交通状况必然受到上下游路段的交通状况的影响. 所以路段上的交通流量势必与相连路段前几个时段的交通流量有着内在的联系. 这样就可以利用路段前几个时段的交通流量数据列去预测未来时段的交通流量, 也可以利用上下游路段前几个时段的交通流量预测路段未来时段的交通流量.

设 $V_i(\tau)$ 为路段 i 上的 τ 时刻的交通流量向量, $V_i(\tau-1)$ 为路段 i 上的 τ 时刻前一时段的交通流量向量. 令 $\hat{V}_i(\tau)=[V_1(\tau),V_2(\tau),\cdots,V_d(\tau)]$, d 为所考虑路段的总数, 若只考虑研究路段的交通流量, 则 $d=1$. 考虑到路段的长度和交通流的特性, 采用当前时间段和前 s 个时间段的交通流量对未来时间段的交通流量进行预测(通常只考虑 3个时间段的交通流量的影响, 即 $s=2$). 这样, 将 $\hat{V}_i(\tau),\hat{V}_i(\tau-1),\cdots,\hat{V}_i(\tau-s)$ 作为第 τ 个输入样本, $\hat{V}_i(\tau+1)$ 作为第 τ 个样本输出值. 故我们的目的是要在 $\hat{V}_i(\tau+1)$ 与路段上 i 前 $s+1$ 个时间段的交通流量(即 $\hat{V}_i(\tau),\hat{V}_i(\tau-1),\cdots,\hat{V}_i(\tau-s)$)之间寻找一个函数关系或者它的一个逼近.

令 $x(\tau) = [\hat{V}_i(\tau), \hat{V}_i(\tau-1), \cdots, \hat{V}_i(\tau-s)]^T$, $y(\tau) = \hat{V}_i(\tau+1)$ 则交通流量预测模型为:

$$y(\tau) = < w, \varphi(x(\tau)) > + b \tag{11}$$

其中 w, b 则是我们要寻求的模型参数. 在预测模型中, w 表示被估计函数的系数, 称为权向量; b 表示被估计函数的常数, 称为偏值. 当我们通过实际数据由改进的 *LS - SVM* 函数估计算法可以得到 w, b, 这样就可以用预测模型 (11) 式来对未来的流量进行预测. 因此, w, b 的获得直接影响我们的预测模型.

表 1　预测所得各项误差指标

Tab. 1　The errors of forecast result

误差指标	平均相对误差 /%	平均绝对相对误差 /%	最大绝对相对误差 /%	相对误差平方和均值平方根 /%	均等系数
MLR	0.52	7.35	23.53	8.94	0.9234
SVM	- 0.94	4.71	25.16	7.41	0.9596
LS - SVM	0.49	5.34	19.55	6.94	0.9634

昆明市一二一大街是昆明市的主要交通要道, 对它的研究和预测具有重要的现实意义. 经过对一二一大街 (建设路交叉路口至圆通北路交叉路口) 由西向东方向机动车 (除摩托车外) 交通流量的人工统计, 得到了一系列实际观测数据. 利用改进的 LS - SVM 方法对一二一大街的交通流量进行预测和模拟, 同时与传统的多元线性回归 (MLR) [6] 及支持向量机方法 (SVM) [7] 进行比较, 预测结果如表 1和图 1.

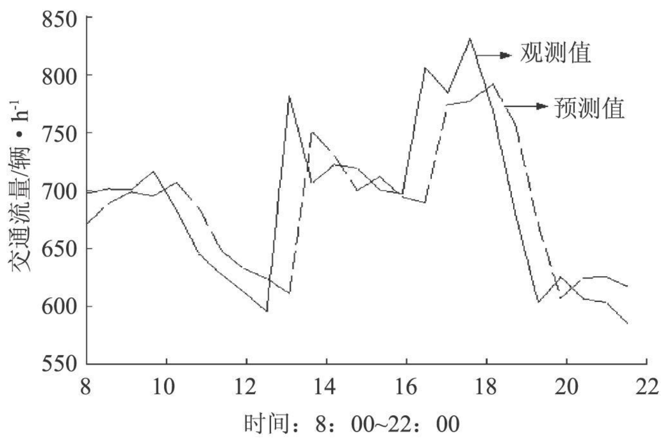

图1　预测结果和真实结果的比较曲线图

Fig.1　The comparison of the forecast result and the real result

4 结　论

基于 O. L Mangasari 等人提出的用于模式识别的 SOR 算法, 本文对用于函数估计的 LS - SVM 算法加以改造, 得到了一种新的函数估计的 LS - SVM 算法. 将该方法应用于交通流量的预测和模拟. 从表 1的结果和图 1的预测曲线可以看出改进的 LS - SVM 算法更具有优越性. 这一新的算法具有能减少计算复杂性、提高学习速度, 且能提高函数估计的精确度等方面的优点. 这一方法有望在交通流量时间序列预测模拟方面得到广泛的推广和应用.

参考文献:

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

[2] Vapnik V N. Statistical learning theory[M]. Wiley, New York, 1998.

[3] Suykens J A K, Brabanter J D, Lukas L, et al. Weighed least squares support vector machines: robustness and sparse approximation[J]. Neurocomputing, 2002, 48: 85 - 105.

[4] 杜树心, 吴铁军. 用于回归估计的支持向量机方法 [J]. 系统仿真学报, 2003, 15(11): 1580 - 1585.

[5] Mangasarian O L, David R Musicant. Lagrangian support vector machine[J]. Journal of Machine Learning Research, 2001, (1): 161 - 177.

[6] 刘剑平, 陆元鸿. 概率论与数理统计方法 [M]. 上海: 华东理工大学出版社, 1999: 166 - 172.

[7] Vladimir N. Vapnik. 统计学习理论的本质 [M]. 张学工, 译. 北京: 清华大学出版社, 2000: 126 - 130.