第 18 卷 第 1 期　　控 制 与 决 策　　2003 年 1 月
Vol 18 No. 1　　Control and Decision　　Jan. 2003

文章编号: 1001-0920(2003)01-0089-03

支持向量机回归在线建模及应用

王定成[1,2], 方廷健[2], 高理富[1,2], 马永军[1]
(1. 中国科学技术大学 自动化系, 安徽 合肥 230026; 2. 中国科学院 合肥智能机械研究所, 安徽 合肥 230031)

摘　要: 支持向量机(SVM)回归理论与神经网络等非线性回归理论相比具有许多独特的优点。讨论了建模中 SVM 核函数、损失函数的选取和容量控制等问题, 并用实验加以验证。将 SVM 回归动态建模理论应用于非线性、时变、大时延温室环境温度变化的建模和预测, 模型简单, 预测效果好。

关键词: 支持向量机; 回归; 建模; 非线性
中图分类号: TP18　　**文献标识码**: A

Support vector machines regression on-line modelling and its application

WANG Ding-cheng[1,2], *FANG Ting-jian*[2], *GAO Li-fu*[1,2], *MA Yong-jun*[1]
(1. Department of Automation, University of Science and Technology of China, Hefei 230026, China;
2. Hefei Institute of Intelligent Machines, Chinese Academy of Science, Hefei 230031, China)

Abstract: The support vector machines theory is shown to have excellent performance compared with other non-linear regression, such as neural networks. The problems how to select the kernel function, loss function and control capacity, and so on, are discussed with simulation demonstration. The dynamic SVM regression modelling is applied to the process of greenhouse temperature change which is non-linear, time-varying, dead-time. The model is simplified and the result of prediction is fine.

Key words: Support vector machines; Regression; Modelling; Non-linear

1　引　　言

尽管线性系统建模的理论和方法比较成熟, 但实际的模型大多是非线性模型, 因此非线性模型更具一般的表达能力, 能更精确地表达真实系统的模型。对于非线性系统而言, 系统模型的建立并没有统一的方法, 用得较多的方法有神经网络中的前馈神经网络和 RBF 神经网络。但神经网络的局部极小点、过学习以及结构和类型的选择过分依赖于经验等固有的缺陷, 严重降低了其应用和发展的效果。支持向量机成功地克服了神经网络的这些缺陷[1], 因而, 采用支持向量机回归算法建立模型是一个新颖而有发展前途的研究方向。

本文将支持向量机回归算法用于非线性建模, 并将其应用于建立具有非线性、时变、大时延的温室环境温度变化的模型。

2　支持向量机回归在线建模

2.1　支持向量机

支持向量机最初用于解决模式识别问题。在模式识别中, 为了发现具有推广能力的决策规则, 所选择的训练数据的一些子集称为支持向量。最佳的支持向量分离等效于所有数据的分离。支持向量机是从线性可分情况下的最优分类面发展而来的, 其基本思想可参阅文献[1～5]。支持向量机形式上类似于一个神经网络, 输出是中间节点的线性组合, 每个

收稿日期: 2001-10-16; **修回日期**: 2001-12-27。
作者简介: 王定成(1967—), 男, 安徽霍山人, 博士生, 从事机器人传感器、信号处理等研究; 方廷健(1939—), 男, 上海人, 研究员, 博士生导师, 从事模式识别、人工智能等研究。

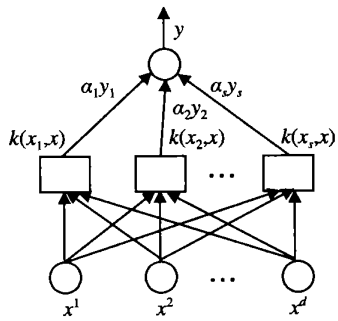


图 1 支持向量机结构

中间节点对应于一个支持向量。其结构如图 1 所示。

2.2 支持向量机回归

支持向量机回归的基本思想是通过一个非线性映射 Φ, 将数据 x 映射到高维特征空间 F, 并在这个空间进行线性回归。即

$$f(x) = (\omega \cdot \Phi(x)) + b$$
$$\Phi: R^n \to F, \quad \omega \in F \tag{1}$$

其中 b 是阈值。这样, 在高维特征空间的线性回归便对应于低维输入空间的非线性回归, 免去了在高维空间 ω 和 $\Phi(x)$ 点积的计算。由于 Φ 是固定不变的, 因此影响 ω 的有经验风险的总和 R_{emp}, 以及使其在高维空间平坦的 $\|\omega\|^2$。则有

$$R(\omega) = R_{emp} + \lambda\|\omega\|^2 = \sum_{i=1}^{l} e(f(x_i) - y_i) + \lambda\|\omega\|^2 \tag{2}$$

其中: l 表示样本的数目, $e(\cdot)$ 是损失函数, λ 是调整的常数。最小化 $R(\omega)$ 便得到用数据点表示的 ω

$$\omega = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*)\Phi(x_i) \tag{3}$$

其中 α_i 和 α_i^* 是最小化 $R(\omega)$ 的解。考虑方程(1) 和(3), $f(x)$ 可表示为

$$f(x) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*)(\Phi(x_i) \cdot \Phi(x)) + b = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*)k(x_i, x) + b \tag{4}$$

其中 $k(x_i, x) = \Phi(x_i) \cdot \Phi(x)$ 称为核函数, 它是满足 Mercer 条件的任何对称的核函数对应于特征空间的点积[5]。核函数的种类较多, 有多项式函数 $k(x_i, x) = [(x \cdot x_i) + 1]^q$, RBF 函数 $k(x_i, x) = \exp\{-|x - x_i|^2/2\sigma^2\}$, Sigmoid 函数 $k(x_i, x) = \tanh(v(x \cdot x_i) + c)$ 等[1,3]。

根据文献[1～3], 式(2) 中的损失函数 $e(\cdot)$ 有以下几种:

1) 线性 ϵ 不敏感损失函数

$$e(f(x) - y) = \begin{cases} 0, & |f(x) - y| \le \epsilon \\ |f(x) - y| - \epsilon, & \text{其他} \end{cases} \tag{5}$$

2) 二次 ϵ 不敏感函数

$$e(f(x) - y) = \begin{cases} 0, & |f(x) - y| \le \epsilon \\ |f(x) - y|^2 - \epsilon, & \text{其他} \end{cases} \tag{6}$$

3) Huber 损失函数

$$e(f(x) - y) = \begin{cases} \epsilon|f(x) - y| - \epsilon^2/2, & |f(x) - y| > \epsilon \\ \frac{1}{2}|f(x) - y|^2, & \text{其他} \end{cases} \tag{7}$$

在如下约束条件下

$$\sum_{i=1}^{l} (\alpha_i - \alpha_i^*) = 0, \quad \alpha_i, \alpha_i^* \in [0, C]$$

求最小化 $R(\omega)$, 即得式(4) 中的 $\alpha_i - \alpha_i^*$。当式(4) 中的 b 取在边界上的一点, 便可进行计算, 但出于稳定性的考虑, 推荐使用边界点上的平均值[2]

$$b = \text{average}_k\left\{\delta_k + y_k - \sum_{i=1}^{l} (\alpha_i - \alpha_i^*)k(x_i, x_k)\right\} \tag{8}$$

其中 δ_k 为预测误差。对于 ϵ 不敏感损失函数, $\delta_k = \epsilon\,\text{sign}(\alpha_k - \alpha_k^*)$; 对于 Huber 损失函数, $\delta_k = (1/C)(\alpha_k - \alpha_k^*)$。

2.3 支持向量机回归在线建模

1) 选取适当的支持向量机模型。2.2 节所述的 $f(x)$ 即为从数据中在线建立的模型。支持向量回归算法主要由核函数、损失函数和容量控制来确定。

① 核函数的选择: 核函数除了上面介绍的 3 种函数外, 还有多层感知器核、Fourier 级数核、B-样条核等多种。从这些核函数中选择一个最好的核函数, 方法之一是通过比较各种核函数的 VC 维的上界, 但这种方法要在非线性特征空间计算包含数据的超平面的半径。比较受欢迎的方法是采用 Bootstraping 或 Cross-validation 来选择核函数。

② 参数的选择: 若有足够的数据采用 Cross-validation, 便可得到核的参数。然而文献[6] 新近提出一种模型选择方法, 不需任何合格的数据便可从理论上来确定参数。

③ 损失函数的选择: 损失函数主要根据实际模型的特点来选择, 例如 ϵ 不敏感损失函数具有稀疏性, 而最小二乘误差准则、最小模损失函数和 Huber 损失函数等则不具有稀疏性。

④ 容量控制: 容量控制在某些情况下直接与调

整的参数相关, 但它与数据中的噪声有关, 因此容量控制取决于数据中的噪声。

2) 如式(4) 所示, 初始m 个样本点建立系统模型。

3) 根据需要预测n 步数据$y_p(m+1), y_p(m+2), \cdots, y_p(m+n)$。

4) 计算实时采集的数据$y(m+1), y(m+2), \cdots, y(m+n)$ 的误差$e(m+1), e(m+2), \cdots, e(m+n)$。

5) 如果$e < \delta$(δ为允许误差), 则转 3)。

6) 将采集到的数据添加到在 2) 中计算出的支持向量集合, 并重新建立模型。

7) 转入 3)。

2.4　支持向量机回归实验仿真

设有样本

$$X = [1.0, 3.0, 4.0, 5.6, 7.8, 10.2, 11.0, 11.5, 12.7]$$

$$Y = [-1.6, -1.8, -1.0, 1.2, 2.2, 6.8, 10.0, 10.0, 10.0]$$

采用多项式ϵ不敏感损失函数的支持向量机对样本数据进行回归建模, 通过优化计算得支持向量数为 9, $b = 0$, 而

$$\beta = \alpha_i - \alpha_i^* = [-0.2736, 0.9791, -2.0000, 1.8688, -1.2098, 0.1766, 2.0000, -2.0000, 0.2637]$$

仿真结果如图 2 所示。实验中, 取多项式的阶数为 6, 容量控制C 为 4, ϵ不敏感损失函数为 0.1。

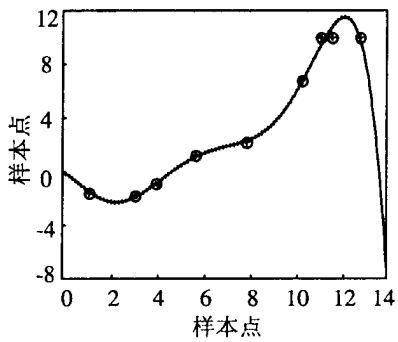


图 2　支持向量机回归实验仿真结果

实际数据到ϵ控制边界的平均误差为 0.1532。从仿真结果可以看出, 模型数据与实际数据几乎完全逼近。

3　支持向量机回归建模的应用

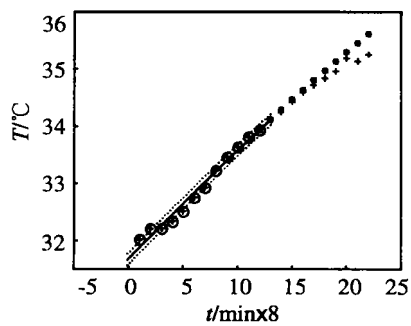
预测控制和模型参考自适应控制等控制方法, 其精度和稳定性取决于模型的准确性, 因此建立实时的动态模型对这类控制具有重要意义。对于多数工业过程, 已有较为成熟的模型, 但对化工、生物、气象、温室环境等过程, 因其非线性、大时延、时变等特性, 很难建立准确的模型。本文采用支持向量机建立温室环境温度的动态模型。

在不降低产量和质量的情况下, 使能量消耗量最小是温室作物生产的最终目的。为达到这一目的, 一些研究者采取了多种控制策略[7, 8], 但这些策略都依赖于温室温度变化的模型。由温室气体能量平衡知识, 可得温室温度变化的简单模型

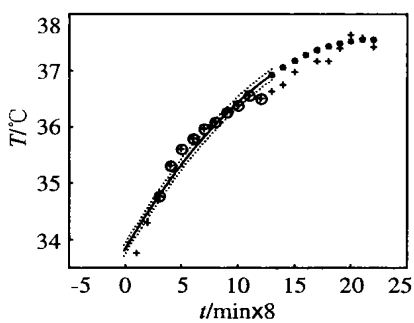
$$\frac{dT_G}{dt} = \frac{1}{C_Q}[K_{out,air}(T_{out} - T_G) + Q_H] \qquad (9)$$

其中: T_G 为温室内温度, C_Q 为温室热容量, $K_{out,air}$ 为从温室内到温室外的热量损失系数, T_{out} 为温室外温度, Q_H 为热动力(未加热情况下为 0)。参数C_Q 和$K_{out,air}$ 取决于温室的建筑材料热容量、质量和风速, 对于不同的温室结构是不一样的, 对于相同结构的温室在不同气候条件下也是不同的。

本文采用支持向量机回归算法对温室的数据进行回归。所采用的支持向量机核函数为径向基函数, 损失函数为二次型损失函数, 径向基函数的 σ= 83, 容量控制为 ∞。所得结果如图 3 所示。



(a)　某天的样本数据和预测数据



(b)　另一天的样本数据和预测数据

图 3　回归模型的温室温度变化

(下转第 95 页)

表 2　不同联合方法的识别率

联合规则	识别率/%
取　中	93.52
取　大	92.13
求　和	91.22
加权求和	97.52

以上二表说明, 单一分类器的识别率低于联合分类器的识别率, 使用加权求和规则的联合方法的识别率高于所比较的其他联合方法的识别率。实验中发现, 许多被求和规则联合方法误分类的样本已由加权求和规则联合方法所改正。

4　结　　论

本文研究了多分类器联合以及联机图形识别问题, 基于通用的分类器联合的理论框架, 提出一种分类器联合方法。该方法改进了文献[2]的求和规则联合方法, 并从理论上说明了改进的原因。将该分类器联合方法应用于联机图形识别, 提高了分类准确率。所联合的分类器有 3 种, 每种分类器的机制各不相同, 并且都基于不同的模式特征。将该联合方法与现有的几种联合方法进行实验比较, 实验结果表明, 在所比较的方法中这种联合方法的识别率最高。

参考文献(References):

[1] Xu L, Krzyzak A, Suen C Y. Methods of combining multiple classifiers and their applications to handwriting recognition[J]. *IEEE Trans SystMan Cybern*, 1992, 22(3): 418-435.

[2] Josef Kittler, Mohamad Hatef, Robert P W D, et al. On combining classifiers[J]. *IEEE Trans Pattern Anal Mach Intell*, 1998, 20(3): 226-239.

[3] 李昌华, 杨兵, 谢维信. 手绘图形结构的识别方法研究[J]. 西安电子科技大学学报, 2000, 27(S): 98-101.
(Li C H, Yang B, Xie W X. Study of hand-sketched graphics structure recognition methods[J]. *J Xidian Univ*, 2000, 27(S): 98-101.)

[4] Rabiner L, Juang B H. *Fundamentals of Speech Recogition*[M]. Beijing: Press Tsinghua Univ Prentice Hall, 1999. 321-389.

[5] Rubine D. Criteria for gesture recognition technologies[A]. *Neural Networks Pattern Recog Human-comp Int*[C]. Chichester: Ellis Horwood Limited, 1992. 243-263.

(上接第 91 页)

所采用的数据为上午 9 时以后的连续 12 个样本点, 样本点的采样间隔为 8 min, 利用回归模型对未来的 10 个时间点的数据进行预测。图 3(a) 表示一天的样本数据及预测数据, 预测的平均误差为 0.165 6; 图 3(b) 表示另一天的样本数据及预测数据, 预测的平均误差为 0.128 8。图中,"+"点为实测数据,"*"点为预测数据。

4　结　　语

支持向量机回归建模将低维非线性的输入映射到高维线性的输出, 模型简单, 具有良好的应用前景。由于理论较新, 这方面的研究主要局限于理论, 很少应用于实际。本文将其应用于温室温度的变化建模, 取得了良好的效果。目前, 多数有关支持向量机的研究仅仅局限于理论和仿真, 因此将理论应用于解决实际问题的研究具有重要意义。

参考文献(References):

[1] Vapnik V. *The Nature of Statistical Learning Theory*[M]. New York: Springer, 1999.

[2] Müller K R, Smola A J, Rätsch G, et al. Predicting time series with support vector machines[A]. *Proc ICANN'97*[C]. New York: Springer, 1997. 999-1004.

[3] Drucker H, Burges C J, Kaufman L, et al. Support vector regression machines[A]. *Adv Neural Infor Proc Syst*[C]. Cambride: MIT Press, 1997. 155-161.

[4] Vapnik V, Golowich S, Smola A. Support vector method for function approximation, regression estimation and signal processing[A]. *Adv Neural Infor Proc Syst*[C]. Cambride: MIT Press, 1997. 281-287.

[5] Boser B E, Guyon I M, Vapnik V N. A training algorithm for optimal margin classifiers[A]. *5th Annual ACM Workshop COLT*[C]. Pittsburgh: ACM Press, 1992. 144-152.

[6] Campbell C. Algorithmic approaches to training support vector machnies: A survey[A]. *Proc ESANN'2000*[C]. Belgium: D-Facto Publications, 2000. 27-36.

[7] Marsh L S, Albright L D. Economically optimum day temperature for greenhouse hydroponic lettuce production — Part 2: Results and simulations[J]. *Trans ASAE*, 1991, 34(3): 557-562.

[8] Maksarov D, Chalabi Z S. Computing bounds on greenhouse energy requirements using bounded error approach[J]. *Contr Eng Prac*, 1998, (6): 947-995.