文章编号: 1001-0920(2005)06-0621-04

基于模糊最小二乘支持向量机的软测量建模

张 英, 苏宏业, 褚 健
(工业控制技术国家重点实验室, 浙江大学 先进控制研究所, 浙江 杭州 310027)

摘 要: 将模糊隶属度概念引入最小二乘支持向量机, 提出一种基于支持向量数据域描述的模糊隶属度函数模型, 将输入空间中的样本映射到一个高维的特征空间; 然后根据其偏离数据域的程度赋予不同的隶属度. 该方法提高了最小二乘支持向量机的抗噪声能力, 尤其适用于未能完全揭示输入样本特性的情况. 将提出的方法用于催化裂化分馏塔轻柴油凝固点的软测量建模, 仿真结果表明, 该模糊隶属度函数模型能够提高最小二乘支持向量机的预测精度.

关键词: 支持向量机; 数据域描述; 模糊隶属度; 软测量; 建模
中图分类号: TP274 **文献标识码**: A

Soft sensor modeling based on fuzzy least squares support vector machines

ZHANG Ying, SU Hong-ye, CHU Jian
(National Laboratory of Industrial Control Technology, Institute of Advanced Process Control, Zhejiang University, Hangzhou 310027, China. Correspondent: ZHANG Ying, E-mail: zhangying@iipc.zju.edu.cn)

Abstract: A fuzzy membership model based on support vector data description is proposed to fuzzify all the training data. Then the model is introduced into least square support vector machines (LS-SVM). Data samples in input space are mapped into a high dimensional feature space, and then the smallest enclosing hypersphere is found. The fuzzy membership value to each input point is computed according to the distances to the center of hypersphere. The proposed soft sensor model based on fuzzy least squares support vector machines is applied to predict the frozen point of light diesel in fluid catalytic cracking process. Simulation result shows that the proposed method actually increases the accuracy of LS-SVM.

Key words: support vector machines; data description; fuzzy membership; soft sensor; modeling

1 引 言

在工业过程质量控制中, 由于缺乏在线的质量测量仪表, 往往存在一些无法直接测量的变量, 而实验室分析值又具有较大的时间滞后, 使得产品质量难以得到保证. 解决这一问题的主要方法是通过软测量建模, 而目前软测量建模的方法主要包括机理建模、基于数据驱动的建模以及混合建模[1]. 1995年, Cortes 和Vapnik 提出了以有限样本统计学习理论为基础的支持向量机[2](SVM), 通过一个二次规划求取样本的最优分类面. 由于SVM 坚实的理论基础, 良好的泛化性能, 并能有效地解决非线性、过学习、局部极值等一系列难题, 使其受到广泛关注. 近年来, Suhkens 提出一种新的SVM 方法——最小二乘支持向量机[3](LS-SVM)方法. LS-SVM 是标准SVM 的一种扩展. 与传统的SVM 不同, LS-SVM 求解线性方程组, 极大减少了SVM 中由于求解二次规划问题带来的计算复杂性, 而且LS-SVM 的数值稳定性和容量控制的策略, 使得核函数矩阵在非正定的情况下也能取得良好的效果.

与SVM 相比, LS-SVM 虽然具有更快的训练

收稿日期: 2004-08-02; **修回日期**: 2004-11-02
基金项目: 国家"十五"科技攻关项目(2001BA204B07); 国家863 计划项目(2001AA413020).
作者简介: 张英(1975—), 男, 湖北天门人, 博士生, 从事数据挖掘、机器学习等研究; 苏宏业(1969—), 男, 浙江杭州人, 教授, 博士生导师, 从事先进控制理论与应用等研究.

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

速度, 但不能保证解是全局最优解, 而且其训练精度有所下降[4]. 本文将模糊隶属度概念引入LS-SVM中, 提出一种基于支持向量数据域描述(SVDD)的模糊隶属度函数模型, 根据样本偏离数据域的程度赋予不同的隶属度. 该方法提高了LS-SVM的抗噪声能力, 尤其适合于未能完全揭示输入样本特性的情况. 仿真结果表明了该方法的有效性.

2 模糊最小二乘支持向量机

给定训练样本集 $S = \{(x_1, y_1), \cdots, (x_l, y_l)\} \in R^n$, $i = 1, 2, \cdots, l$. 在非线性情况下引入变换 $\varphi: R^n \to H$, 将样本从输入空间 R^n 映射到一个高维特征空间 H. 输入空间中的函数估计可归结为求解下面的二次规划[2,5]:

$$\min \Phi = \frac{1}{2}\|w\|^2 + \frac{1}{2}C\sum_{i=1}^{l}\xi_i^2,$$
$$\text{s.t.}\quad y_i = w \cdot \varphi(x_i) + b + \xi_i. \tag{1}$$

其中: ξ_i 为松弛变量, C 为惩罚因子. 与传统的SVM相比, 这一方法中二次规划约束条件为等式, 且损失函数为二次函数, 故称为最小二乘支持向量机. 引入Lagrange系数 α_i, 定义如下的Lagrange函数:

$$L = \frac{1}{2}\|w\|^2 + \frac{1}{2}C\sum_{i=1}^{l}\xi_i^2 - \sum_{i=1}^{l}\alpha_i(w \cdot \varphi(x_i) + b + \xi_i - y_i). \tag{2}$$

根据Mercer条件, 存在映射 φ 和核函数 $K(\cdot, \cdot)$, 使得 $K(x_i, x_j) = \varphi(x_i) \cdot \varphi(x_j)$. 令 L 对变量 w, b, ξ_i, α_i 的偏导数等于零, 并将得到的等式代入式(2), 可以得到矩阵方程

$$\begin{bmatrix} 0 & 1_v^{\mathrm{T}} \\ 1_v & \Omega + C^{-1}I \end{bmatrix}\begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix}. \tag{3}$$

其中: $y = [y_1, \cdots, y_l]$, $1_v = [1, \cdots, 1]$, $\alpha = [\alpha_1, \cdots, \alpha_l]$, Ω 中的元素为 $\Omega_{ij} = K(x_i, x_j)$, $i, j = 1, 2, \cdots, l$. 求解矩阵方程(3), 最后得到最小二乘支持向量机的函数估计为

$$f(x) = \sum_{i=1}^{l}\alpha_i K(x_i, x) + b \tag{4}$$

传统的SVM中 α 只有一小部分分量不为零(支持向量), 而LS-SVM中 α 的每一个分量与样本的误差 ξ_i 成正比, 所以在LS-SVM中没有支持向量的概念. LS-SVM中常用的核函数 $K(x_i, x)$ 包括线性核 $x_i \cdot x$, 多项式核 $(x_i \cdot x + 1)^d$ 以及高斯径向基核 $e^{-p\|x_i - x\|^2}$ 等.

为解决SVM对于孤立点过分敏感并由此而带来的过拟合问题[6], Lin等将模糊隶属度的概念引入SVM, 模糊化输入样本集, 提出了模糊支持向量机[7](FSVM)的概念. 将这一思想引入LS-SVM, 为LS-SVM中每个样本引入模糊隶属度 μ_i, 模糊化输入样本集 $(x_1, y_1, \mu_1), \cdots, (x_i, y_i, \mu_i), \cdots, (x_l, y_l, \mu_l)$, $0 < \mu_i < 1$. 将式(1)中的目标函数重写为

$$\min \Phi = \frac{1}{2}\|w\|^2 + \frac{1}{2}C\mu_i\sum_{i=1}^{l}\xi_i^2, \tag{5}$$

与LS-SVM函数估计方法一样, 构造Lagrange函数, 最后得到矩阵方程

$$\begin{bmatrix} 0 & 1_v^{\mathrm{T}} \\ 1_v & \Omega + (C\mu_i)^{-1}I \end{bmatrix}\begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix}. \tag{6}$$

求解矩阵方程(6)即可得到模糊最小二乘支持向量机的估计函数. 与式(3)相比, 矩阵方程(6)中多了模糊隶属度 μ_i, 故这一方法被称为模糊最小二乘支持向量机(FLS-SVM).

3 模糊隶属度

在分类问题中, Lin等提出一种根据样本采集的先后顺序确定样本隶属度的模型[7], 该模型认为, 最近得到的样本相对要比其他的样本重要, 其隶属度也大, 但该模型缺乏理论上的依据. Huang提出一种基于孤立点检测的模糊隶属度模型[8], 他将样本集分为两部分: 一部分为孤立点集, 另一部分为主体集. 对于主体集中的样本, 根据样本到其聚类中心的距离确定模糊隶属度; 而对于孤立点集中的样本其模糊隶属度则赋予一个很小的正数. 显然Huang并没有区分孤立点集中的样本, 仅简单地为每个孤立点赋予相同的模糊隶属度. 为了确定模糊隶属度函数的形式, 需要衡量一个样本偏离其所在类总体的程度. 本文采用支持向量数据域描述方法, 将数据样本映射到一个高维的空间, 然后在这个高维空间中寻找其最小包含超球, 并根据样本到超球球心的距离确定其隶属度值.

3.1 支持向量数据域描述

支持向量数据域描述[9]方法可描述为: 给定训练样本集 $X = \{x_1, \cdots, x_i, \cdots, x_l\}$, 其中: $x_i \in R^n$ 为输入空间, $i = 1, 2, \cdots, l$, l 为样本个数. 为了建立样本的数据域描述模型, 需要寻找样本的最小包含超球. 当输入空间中的样本为非球形分布时, 引入映射 $\Phi: R^n \to F$, 将输入空间中的样本映射到一个高维的特征空间 F, 然后求解下面的二次规划:

$$\min W = R^2 + C\sum_{i=1}^{l}\xi_i,$$
$$\text{s.t.}\quad \|\Phi(x_i) - a\|^2 \le R^2 + \xi_i,$$
$$\xi_i \ge 0,\ i = 1, 2, \cdots, l \tag{7}$$

其中: R 为最小包含超球半径, a 为球心, ξ_i 为松弛变量, C 为惩罚因子. 引入Lagrange系数 β_i 和 η, 并进行对变换, 最后得到Wolfe对偶为

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$$\max Q = \sum_{i=1}^{l} K(x_i, x_i)\beta_i - \sum_{i,j=1}^{l} \beta_i\beta_j K(x_i, x_j),$$
$$\text{s.t. } 0 \le \beta_i \le C,\ i = 1, 2, \cdots, l. \quad (8)$$

求解上述规划问题, 即可得到最优的Lagrange 系数及特征空间中的数据域描述

3.2　基于数据域描述的模糊隶属度函数模型

输入空间中的点x_i在特征空间中映射$\Phi(x_i)$到最小包含超球球心 a 的距离定义为$D^2(x_i) = \|\Phi(x_i) - a\|^2$, 考虑到$a = \sum_{i=1}^{l}\beta_i\Phi(x)_i$, 有

$$D^2(x_i) = \sum_{i,j=1}^{l}\beta_i\beta_j K(x_i, x_j) + K(x_i, x_i) - 2\sum_{j=1}^{l} K(x_j, x_i)\beta_j,\ i, j = 1, 2, \cdots, l. \quad (9)$$

定义$X_{NBSV} = \{x_1, \cdots, x_k, \cdots, x_m\}$为输入空间中的子集, 其中: x_k为样本中非边界支持向量$(0 < \beta_k < C)$, m为非边界支持向量的个数. 特征空间中最小包含超球半径满足$R = D(x_i)\,|\,x_i \in X_{NBSV}$, 当$R$和 a 确定后, 便可得到给定数据集的数据域描述. 定义

$$D_{max} = \max(D(x_i)\,|\,x_i \in X),$$
$$D_{min} = \min(D(x_i)\,|\,x_i \in X), \quad (10)$$

分别为样本到最小包含超球球心最大、最小距离. 定义模糊隶属度函数如下:

$$\mu_i = \begin{cases} \left(1 - \dfrac{D(x_i) - D_{min}}{D_{max} - D_{min}}\right)^f + \sigma, \\ R < D(x_i) \le D_{max}; \\ 1 - \dfrac{D(x_i) - D_{min}}{D_{max} - D_{min}}, \\ D_{min} \le D(x_i) \le R. \end{cases} \quad (11)$$

其中: $\sigma < 1$, 为足够小的正实数; $f \ge 2$. 当$f = 2$时, 模糊隶属函数μ_i随$D(x_i)$变化的曲线如图 1 所示

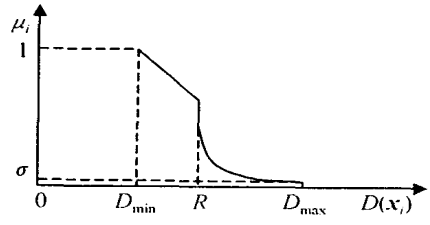

图 1　模糊隶属度随样本到特征空间超球球心距离变化曲线

对于输入空间中的点x_i, 其在特征空间中的映射$\Phi(x_i)$到最小包含超球球心 a 的距离满足$D_{min} \le D(x_i) \le D_{max}$. 当$D_{min} \le D(x_i) \le R$时, 表示$x_i$满足数据域描述, $\Phi(x_i)$位于超球内或球面上, 其模糊隶属度随着$\Phi(x_i)$的增大而线性减少; 当$R < D(x_i) \le D_{max}$时, 样本x_i偏离数据域描述, 也就是偏离其所在的类总体, $\Phi(x_i)$位于超球之外, 其模糊隶属度是$D(x_i)$的二次函数, 随着$D(x_i)$的增大, 其模糊隶属度迅速减小. 当$D(x_i)$接近于D_{max}时, 其隶属度已接近于一个非常小的正实数σ, 这样可以减少这些点的影响

4　仿真研究

重油催化裂化装置(FCCU)是石油二次加工中的关键装置之一, FCCU 主要由反应再生、分馏和吸收稳定 3 个子系统组成. 其中分馏子系统的目的是将反应再生子系统中生成的粗汽油按照不同的馏程进行分割, 得到不同的产品, 轻柴油是其中的产品之一, 由于缺少在线的质量检测仪表, 为控制其质量, 必须对轻柴油的质量指标凝固点进行在线预测

将提出的模糊最小二乘支持向量机方法用于某炼油厂催化裂化分馏塔轻柴油凝固点的估计. 首先进行二次变量的选择. 根据工艺分析, 影响轻柴油凝点的主要因素有轻柴油的抽出温度, 抽出层气相温度, 一中循环量, 一中抽出温度以及一中返塔温度. 将这 5 个变量作为估计轻柴油凝固点的辅助变量. 从现场收集数据, 经过预处理和归一化等步骤后, 得到 200 个样本用于建模训练, 250 个样本用于测试

用LS-SVM 进行训练, 核函数为高斯径向基核, 对于参数C和p的选择采用两层网格搜索的策略. 首先设定C和p的候选集为比较松散的网格$(2^{-5}, 2^{-3}, \cdots, 2^{15})$和$(2^{-9}, 2^{-7}, \cdots, 2^{11})$, 以网格中的节点为参数样本进行 10 倍样本交叉检验, 得到最大的交叉检验精度所对应的网格节点为$(2^5, 2^3)$; 然后在其一定的范围类内构造比较细的网格$(2^4, \cdots, 2^{4.75}, 2^5, 2^{5.25}, \cdots, 2^6)$和$(2^2, \cdots, 2^{2.75}, 2^3, 2^{3.25}, \cdots, 2^4)$, 再次以网格中的节点为参数样本进行 10 倍样本交叉检验, 最后得到C和p的值为$(2^{5.75}, 2^{3.5})$. 采用LS-SVM 训练后, 用得到的模型对测试集进行测试, 预测结果如图 2 所示, 预测均方差为 0.115 7.

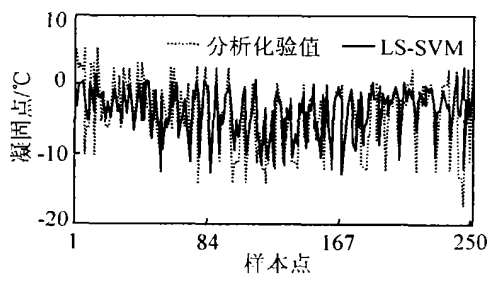

图 2　采用LS-SVM 的轻柴油凝固点预测值曲线

为了利用FLS-SVM 在数据集$(x_1, y_1), \cdots, (x_i, y_i), \cdots, (x_l, y_l)$上进行训练, 首先需要确定输入样本$\{x_1, \cdots, x_i, \cdots, x_l\}$的最小包含超球半径$R$以及$D_{min}$和$D_{max}$. 在$\{x_1, \cdots, x_i, \cdots, x_l\}$上采用 SVDD 进行训

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

练, 核函数为高斯径向基核, $p = 1, C = 0.1$, 得到$R = 20.4, D_{max} = 20.64, D_{min} = 20.3$. 选取$f = 2, \sigma = 0.001$, 模糊隶属度函数为

$$\mu_i = \begin{cases} 8.81D^2(x_i) + 363.62D(x_i) + 3751.88, \\ \quad 20.3 \leq D(x_i) \leq 20.4; \\ -2.97D(x_i) + 61.25, \\ \quad 20.3 \leq D(x_i) \leq 20.4 \end{cases} \tag{12}$$

得到模糊隶属度函数之后再用 FLS-SVM 进行训练, 采用与 LS-SVM 相同的核函数和模型参数$(2^{5.75}, 2^{3.5})$, 训练完后对 250 个样本进行预测, 结果如图 3 所示, 预测均方差为 0.0722.

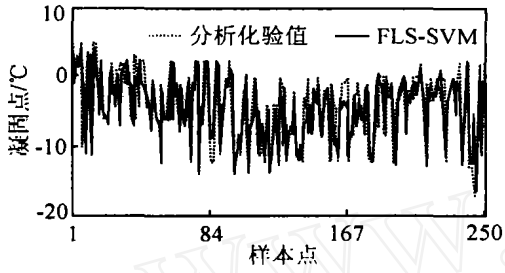

图 3 采用 FLS-SVM 的轻柴油凝固点预测值曲线

由图 2 和图 3 可以看出, FLS-SVM 的拟合效果要好于 LS-SVM, 且预测均方差显著减小.

5 结 语

本文将模糊隶属度概念引入 LS-SVM 中, 提出了一种基于支持向量数据域描述的模糊隶属度函数模型. 首先得到训练集中样本的数据域描述模型, 然后根据样本偏离数据域的程度赋予不同的隶属度. 该方法提高了 LS-SVM 的抗噪声能力, 尤其适合于未能完全揭示输入样本特性的情况. 将提出的方法运用于催化裂化分馏塔轻柴油凝固点的软测量建模, 仿真结果表明, 提出的模糊隶属度函数模型可有效地提高 LS-SVM 的预测精度.

参考文献(References)

[1] 徐敏, 俞金寿. 软测量技术[J]. *石油化工自动化*, 1998, 10(2): 1-3.
(Xu M, Yu J S. Technology of soft sensor [J]. *Automation in Petro-chemical Industry*, 1998, 10(2): 1-3.)

[2] Vapnik V. *Statistical learning theory* [M]. New York: Wiley Springer, 1998: 146-175.

[3] Suykens J A K, Vandewalle J. Least squares support vector machine classifiers [J]. *Neural Processing Letter*, 1999, 9(3): 293-300.

[4] 阎威武, 邵惠鹤. 支持向量机和最小二乘支持向量机的比较及应用研究[J]. *控制与决策*, 2003, 18(3): 358-360.
(Yan W W, Shao H H. Application of support vector machines and least squares support vector machines to heart disease diagnoses[J]. *Control and Decision*, 2003, 18(3): 358-360.)

[5] Lin C J. Formulations of support vector machines: A note from an optimization point of view [J]. *Neural Computation*, 2001, 13(2): 307-317.

[6] Zhang X G. Using class-center vectors to build support vector machines [A]. *Neural Networks for Signal Processing IX - Proc of the* 1999 *IEEE Workshop* [C]. Wisconsin: IEEE Inc, 1999: 33-37.

[7] Lin C F, Wang S D. Fuzzy support vector machines [J]. *IEEE Trans on Neural Networks*, 2002, 13(3): 466-471.

[8] Huang H P, Liu Y H. Fuzzy support vector machines for pattern recognition and data mining [J]. *Int J of Fuzzy Systems*, 2002, 4(3): 3-12.

[9] Tax D M J, Duin R P W. Data domain description by support vectors[A]. *Proc of 8th European Symposium on Artificial Neural Networks*[C]. Brussels: Facto D, 1999: 251-256.

(上接第 620 页)

[6] 郑有才, 蔡希尧. 元数据驱动的可通用通信软件的设计[J]. *西安电子科技大学学报*, 1998, 25(6): 778-781.
(Zhen Y C, Cai X Y. The implementation of metadata-driven reusable communication software [J]. *J of Xidian University*, 1998, 25(6): 778-781.)

[7] Bharat Jayaraman. Semantics of EqL [J]. *IEEE Trans on Software Engineering*, 1988, 14(4): 472-480.

[8] Kewley Robert Hargreaves Jr. Computational intelligence for support of military tactical decision making [D]. Troy: Rensselaer Polytechnic Institute, 2000.

[9] Amit Jayant Patel. Obstal: A language with objects, subtyping, and classes [D]. Stamford: Stanford: University, 2001.

[10] Subrahmanyam P A, Singh K J, Guy Story, et al. Quality assurance in scripting[J]. *IEEE Multimedia*, 1995, 2(2): 50-59.

[11] Jun Miura, Motokuni Ito, Yoshiaki Shirai. A three-level control architecture for autonomous vehicle driving in a dynamic and uncertain traffic environment [A]. *Proc of IEEE Conf on Intelligent Transportation Systems* [C]. Boston: IEEE Press, 1997: 706-711.

© 1994-2010 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net