模糊最小二乘支持向量机回归研究及应用①

孙 政, 潘 丰

(江南大学 轻工过程先进控制教育部重点实验室, 无锡 214122)

摘 要: 针对传统支持向量机对训练样本内的噪声和孤立点比较敏感, 导致建模精度不高的问题, 将模糊集理论引入到最小二乘支持向量机回归中, 建立一种基于数据域描述的模糊最小二乘支持向量机回归的数学模型, 该方法将样本映射到高维空间, 在高维空间中寻找最小包含超球, 然后根据样本到超球心的距离确定模糊隶属度的大小, 通过仿真实验验证, 该算法提高了支持向量机回归的训练精度, 将此模型应用于谷氨酸发酵过程菌体浓度预测, 结果表明此方法的有效性.

关键词: 最小二乘支持向量机回归; 数据域描述; 谷氨酸发酵;

Research and Application of Fuzzy Least Square Support Vector Machine Regression

SUN Zheng, PAN Feng

(Key Laboratory of Advanced Process Control for Light Industry (Ministry of Education), Jiangnan University, Wuxi 214122, China)

Abstract: The traditional SVM is more sensitive to the noise and isolated points in training sample, and have lower modeling accuracy. In this paper, fuzzy set theory is introduced to least squares support vector regression, and then to establish a data domain description fuzzy least squares support vector machine regression. This method will sample mapped into a high dimensional space, and search a minimum enclosing sphere in high dimensional space. Meanwhile, according to the distance from sample to the center of the sphere, the size of fuzzy membership can be determined. A simulation experiment is provided to demonstrate that this algorithm can improve the accuracy of support vector machine regression. This model is applied to predict the concentration of glutamic acid bacteria fermentation process. Results we obtain in simulation show the effectiveness of the proposed approach.

Key words: least squares support vector machine regression; data domain description; glutamic acid fermentation

支持向量机(Support Vector Machine SVM)是 20 世纪 90 年代中期由 Vapnik 等人提出的新的基于统计学习理论的机器学习算法[1], 它有效地解决小样本、非线性和高维模式识别问题, 并在很大程度上克服了“维数灾难”和“过学习”等问题, 使得它一出现就受到广泛的关注, 并在故障诊断、图像识别、孤立点检测[2-4]等领域得到成功的应用, 在生物发酵方面也进行了有益的探索,

谷氨酸发酵过程极其复杂, 反应过程中非线性、时变性和不确定性严重, 菌体浓度、基质浓度等重要变量难以在线测量, 这给生产带来了很大的影响, 解决这一问题主要通过软测量建模, 近年来, Suykens 等人提出了最小二乘支持向量机(Least Squares Support Vector Machine LSSVM)[5],极大地减少了 SVM 中求解约束二次凸规划带来的计算复杂性, 其虽提高了更快的训练速度, 但不能保证解是全局最优解, 而且其精度也有所下降, 文献[6]提出了支持向量域描述(Support Vector Domain Description SVDD), 主要思想为采用将样本映射到一个高维空间, 在这个高维空间中寻找最小包含超球, 然后根据样本到超球心的距离确定模糊隶属度的大小, 本文将其思想引入到最小二乘支持向量机中, 提出一种基于数据域描述的模糊最小二乘支持向量机回归(Fuzzy Least Squares Support Vector Machines Regression FLSSVR).

① 基金项目:国家自然科学基金(61273131);江苏高校优势学科建设工程资助项目(PAPD)
收稿时间:2013-12-04;收到修改稿时间:2013-12-30

1　最小二乘支持向量机回归

对于给定训练集，$T=\{(x_i,y_i),i=1\cdots l\}$ $x_i\in R^N$，$y_i\in R$ 在特征空间建立具有如下式的回归方程:

$$f(x)=\langle\omega,\varphi(x)\rangle+b \tag{1}$$

其中 $\varphi(x)$ 为非线性映射函数，ω 和 b 分别为权值向量和偏置. 采用最小二乘支持向量机进行函数估计，回归问题可以表示为如下约束优化问题:

$$\min_{w,b,e} J(w,e)=\frac{1}{2}w^Tw+\frac{C}{2}\sum_{i=1}^{n}e_i^2 \tag{2}$$

$$s.t.\ \ y_i=w^T\phi(x_i)+b+e_i,i=1,2,...,n. \tag{3}$$

式中，误差变量为 $e_i\in R$，惩罚系数为 $C\in R$. 求解上述优化问题，把约束优化问题变为无约束优化问题，引入 Lagrange 乘子 a，定义 Lagrange 函数形式为:

$$L(w,b,e,a)=J(w,e)-\sum_{i=1}^{n}a_i\{w^T\phi(x_i)+b+e_i-y_i\} \tag{4}$$

根据 KKT 条件有:

$$\frac{\partial L}{\partial w}=0\rightarrow w=\sum_{i=1}^{l}a_i\varphi(x_i) \tag{5}$$

$$\frac{\partial L}{\partial b}=0\rightarrow\sum_{i=1}^{l}a_i=0 \tag{6}$$

$$\frac{\partial L}{\partial e_i}=0\rightarrow a_i=Ce_i \tag{7}$$

$$\frac{\partial L}{\partial a_i}=0\rightarrow w^T\phi(x_i)+b+e_i-y_i=0 \tag{8}$$

从由式(4)～(8)组成的方程组中消去 e_i，w 后，得到下式:

$$\begin{bmatrix}0 & \mathrm{P}^T\\ \mathrm{P} & \Omega+C^{-1}I\end{bmatrix}\begin{bmatrix}a\\ b\end{bmatrix}=\begin{bmatrix}0\\ y\end{bmatrix} \tag{9}$$

其中，I 为单位阵，$y=(y_1,y_2,...,y_n)^T$

$\mathrm{P}=(1,1,...,1)^T$　$a=(a_1,a_2,...,a_n)^T$

$\Omega=(\phi(x_i)\bullet\phi(x_j))=k(x_i,x_j)$

$i,j=1,2,...,n.$

通过求解(9)可求出 a，b. 设 $H=\Omega+C^{-1}I$ 则

$$\mathrm{P}^Ta=0 \tag{10}$$

$$\mathrm{P}b+Ha=y \tag{11}$$

令 $Q=H^{-1}$，解(10)、(11)得:

$$a=Q\left(y-\frac{\mathrm{P}^TQy}{\mathrm{P}^TQ\mathrm{P}}\right) \tag{12}$$

$$b=\frac{\mathrm{P}^TQy}{\mathrm{P}^TQ\mathrm{P}} \tag{13}$$

求出 a，b 后，从而得到最小二乘支持向量机的回归模型，表达式如下:

$$y(x)=\sum_{i=1}^{n}\alpha_ik(x,x_i)+b \tag{14}$$

2　基于支持向量机数据域的模糊隶属度

在支持向量机数据域方法中，对于给定输入空间的训练集 $X=\{x_1,x_2,\cdots x_l\}$，$x_i\in R^N$，SVDD 方法将输入空间中的样本通过函数 Φ 映射到一个高维特征空间 F，然后在高维特征空间 F 中寻找最小包含超球，因此可以将此归结为一个二次规划问题:

$$\min W_{SD}(\xi_i,R,a)=r^2+C_{SD}\sum_{i=1}^{l}\xi_i \tag{15}$$

$$\text{s.t.}\begin{cases}\|\Phi(x_i)-a\|^2\le r^2+\xi_i\\ \xi_i\ge 0\quad i=1,2,\cdots l\end{cases}$$

其中 r 为特征空间中的最小包含超球的半径，C_{SD} 为惩罚系数，ξ_i 为松弛变量，o 为球心. 引入 Lagrange 乘子 β_i 和 η_i，并求得其对偶方程为:

$$\max\Psi=\sum_{i=1}^{l}K(x_i,x_i)\beta_i-\sum_{i=1}^{l}\sum_{j=1}^{l}\beta_i\beta_jK(x_i,x_j) \tag{16}$$

$0\le\beta_i\le C_{SD}$

其中 $K(x_i,x_j)=\Phi(x_i)\Phi(x_j)$.

对于输入空间中的任意一点 x_i，其映射 $\Phi(x_i)$ 到超球球心的欧式距离定义为:

$$D^2(x_i)=\|\Phi(x_i)-o\|^2 \tag{17}$$

定义 $X_{SV}=\{x_1,\cdots,x_l,\cdots,x_m\}$ 为输入训练集的一个子集，其中 x_k 为非边界支持向量($0<\beta_l<C_{SD}$)，非边界支持向量位于超球面上，所以 $R=D(x_i)$，$x_i\in X_{SV}$. 在求得 R 和 o 之后，就可以得到给定输入数据集的数据域描述，设

$$D_{\max}=\max(D(x_i)|x_i\in X) \tag{18}$$

$$D_{\min}=\min(D(x_i)|x_i\in X)$$

分别为样本到超球球心的最大距离和最小距离. 因此，可得模糊隶属度函数为:

$$\mu_i=\begin{cases}(1-\dfrac{D(x_i)-D_{\min}}{D_{\max}-D_{\min}})^f+\tau & R<D(x_i)\le D_{\max}\\ 1-\dfrac{D(x_i)-D_{\min}}{D_{\max}-D_{\min}} & D_{\max}\le D(x_i)\le R\end{cases} \tag{19}$$

其中 $\tau<1$，是一个足够小的正数，$f\ge 2$. $f=2$ 时，模糊隶属度函数的示意图 1 所示.

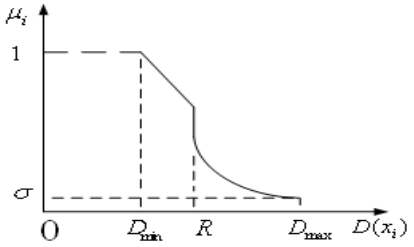

图 1 支持向量机数据域描述的模糊隶属度函数

3 模糊最小二乘支持向量机回归

Lin 等学者提出了模糊支持向量机方法[7]. 引入模糊隶属度 μ_i，原本的输入样本集就变为 $T=\{(x_i,y_i,\mu_i),i=1\cdots l\}$，则最小二乘支持向量回归优化问题变为:

$$\min_{w,b,e} J(w,e)=\frac{1}{2}w^Tw+\frac{C}{2}\mu_i\sum_{i=1}^{n}e_i^2 \tag{20}$$

则最后得到的矩阵方程:

$$\begin{bmatrix}0 & e1^T\\ e1 & \Omega+\mu_iC^{-1}I\end{bmatrix}\begin{bmatrix}a\\ b\end{bmatrix}=\begin{bmatrix}0\\ y\end{bmatrix} \tag{21}$$

从而 $Q=H^{-1}=\left(\Omega+\mu_iC^{-1}I\right)^{-1}$. 与(9)相比, (21)引入了模糊隶属度 μ_i 函数, 因此得到模糊最小二乘支持向量机回归(FLSSVR).

4 仿真及应用

为了说明 FLSSVR 算法的正确性和有效性, 针对回归问题给过一个数值仿真实例和在谷氨酸发酵过程菌体浓度预测的仿真, 并与模糊支持向量机回归(FSVR)算法进行比较.

4.1 $\sin c$ 函数仿真

在有白噪声的情况下, $\sin c$ 函数的输入 x 和输出 y 之间的关系为:

$$y=\frac{\sin x}{x}+e \tag{22}$$

其中, 均值0, 方差为0.01, e 为高斯白噪声. 从 x 的定义区间等间距的取 180 个点, 并且取 20 个孤立的点组成训练集. FSVR 和 FLSSVR 的参数取: $C=10$, 核函数为 RBF 函数, $\sigma=30$, 训练结果比较如图 2 所示.

训练误差比较如表1所示. 均方误差计算式如下:

$$MSE=\frac{1}{n}\sum_{i=1}^{n}(y_i-f(x_i))^2 \tag{23}$$

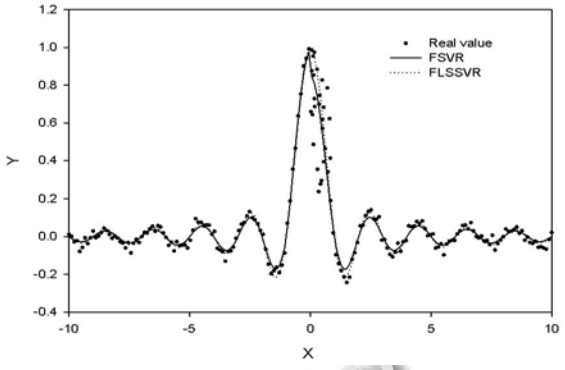

图 2 训练结果比较

表 1 训练均方差比较

方法	MSE
FSVR	1.07E-03
FLSSVR	7.94E-04

从表1结果可以看出, FLSSVR相比与FSVR,降低了孤立点和噪声的干扰, 训练误差值较小, 提高了模型精度. 因此, FLSSVR更加适合实际过程的建模.

4.2 谷氨酸菌体浓度预测

4.2.1 谷氨酸菌体浓度输入变量的选择

通过对谷氨酸发酵工艺和机理分析[8], 菌体的生长过程与溶氧浓度、氨水消耗量摄氧量、二氧化碳生成量对菌体的生长及产物的产量和性质有较大的促进和抑制作用, 从而直接影响着谷氨酸发酵过程的其他一些参数的变化, 并影响到产物生成量的变化; 本文以当前时刻的采样时间、溶氧浓度、谷氨酸浓度、菌体浓度作为输入变量建模[9], 以下一时刻的谷氨酸浓度作为模型的输出.

4.2.2 FLSSVR 中参数的选取

惩罚因子、核函数及其参数和隶属度的选取在模糊支持向量机具体使用中对分类精度有重要影响.

对于 FLSSVR 参数的选取, 核函数为径向基核函数, =0.7, 其余参数值为=200, =0.001.

对于 FOSVR 参数的选取, 核函数为径向基核函数, =0.7, =20, =0.001; SVDD 算法核函数选用径向基函数, , , 经过训练后得到, , ; 模糊隶属度函数, , 因此可得

根据模糊隶属度函数, 对所有的样本分配模糊隶属度.

4.3 实验仿真

为了验证改进算法的性能, 选取 100 个试验样本

数据作为训练集，发酵所用的 Corynebacterium Glutamicum S9114 种子体由江南大学教育部工业生物技术重点实验室提供，进行训练和测试. 根据设定的参数，采用FSVR和FLSSVR离线训练模型. 另外再用10批数据测试训练好的模型,仿真结果如图 3 所示，底物浓度估计结果的均方差如表 2 所示.

表 2　FSVR 和 FLSSVR 训练均方差

	FSVR	FLSSVR
MSE	1.521 E-03	9.754 E-04

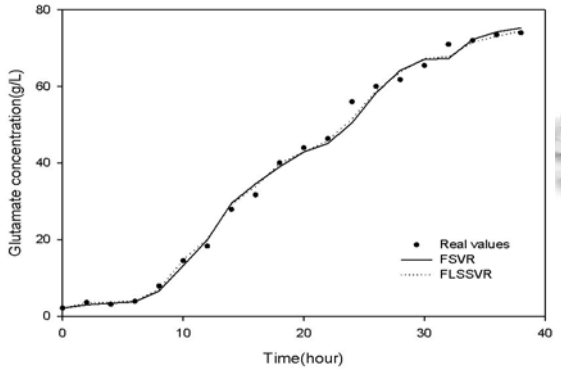

图 3　基于 FSVR 和 FLSSVR 的谷氨酸浓度估计结果

由图 3 可以看出基于 FLSSVR 的谷氨酸模型与测试值表现出良好的拟合效果.

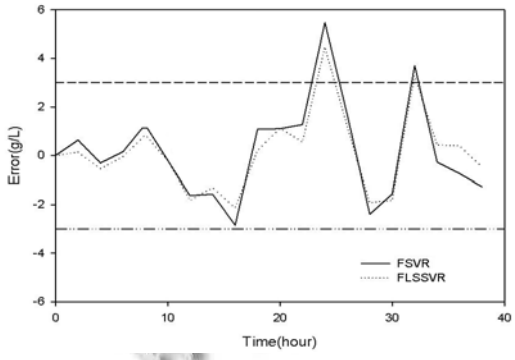

图 4　基于 FSVR 和 FLSSVR 的谷氨酸浓度预测误差

由图 4 和表 2 可以看出，基于 FLSSVR 的谷氨酸模型在精度上明显高于 FSVR，这主要是因为对 FLSSVR 对原有 FSVR 算法中固定惩罚因子做出改变，对每个样本定义不同的模糊隶属度，分配了不同的惩罚因子，从而抑制了边缘数据和噪声对算法的影响，使得建模精度得到提高.

5　结论

本文将支持向量描述的模糊隶属度概念引入 LSSVM 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，得到 FLSSVR. 首先得到训练集中样本的数据域描述模型，然后根据样本到超球心的距离确定模糊隶属度的大小. 该方法减少了由于数据样本中孤立点的存在而带来的过拟合现象，提高了支持向量机的抗噪声能力,将提出的方法运用于谷氨酸发酵过程菌体浓度预测的软测量建模，仿真结果表明,提出的FLSSVR可有效地提高了支持向量机的预测精度. 但本文未对最小二乘支持向量机的稀疏性问题进行探讨，这将是以后的工作重点，以期能够保证最小二乘支持向量机具有良好的鲁棒性.

参考文献

1 Vapnik VN.张学工译.统计学习理论的本质.北京:清华大学出版社,2000.

2 彭光金,司海涛,俞集辉,等.改进的支持向量机算法及其应用.计算机工程与应用,2011,47(18):218–221.

3 Shi GR. The use of support vector machine for oil and gas identification in low porosity and low permeability reservoirs. International Journal of Mathematical Modeling and Numerical Optimization, 2009, 1(1/2): 75–87.

4 Floris E, Achim S. Forecasting respiratory motion with accurate online support vector regression. International Journal of Computer Assisted Radiology and Surgery, 2010, 4(5): 439–447.

5 Suykens JAK, Vandewalle J. Least squares support vector machine classifiers. Neural Processing Letter, 1999, 9(3): 293–300.

6 Tax DMJ, Duin RPW. Data domain description by support vectors. Proc of 8th European Symposium on Artificial Neural Networks. Brussels: Facto D. 1999. 251–256.

7 Lin CF, Wang SD. Fuzzy support vector machines. IEEE Trans. on Neural Networks, 2002, 13: 464–471.

8 于信令,于军.氨基酸发酵工程的新进展.发酵科技通讯,2007(2):41-42.

9 刘国海,张东娟,梅从立.基于 IRLS-ELM 生物发酵在线软测量建模方法研究.东南大学学报,2011,41(S):433–44.