第21卷第9期 控 制 与 决 策 2006年9月
Vol. 21 No. 9 Control and Decision Sep. 2006

文章编号: 1001-0920(2006)09-1073-04

KPCA-LSSVM 建模方法及在钢材淬透性中的应用研究

郭 辉, 刘贺平, 王 玲
(北京科技大学 信息工程学院, 北京 100083)

摘 要: 通过等式约束条件修改普通的支持向量机可以得到最小二乘支持向量机, 不需要再次求解复杂的二次规划问题. 提出了利用核主元分析进行特征提取, 在高维特征空间中计算主元, 降低样本的维数, 然后用最小二乘支持向量机进行建模. 仿真结果表明了该方法的有效性和优越性.

关键词: 核的主元分析; 最小二乘支持向量机; 主元; 特征提取
中图分类号: TP301.5 **文献标识码**: A

Modeling Approach Based on KPCA-LSSVM and Its Application to Steel Harden-ability Prediction

GUO Hui, LIU He-ping, WANG Ling
(School of Information Engineering, University of Science and Technology Beijing, Beijing 100083, China. Correspondent: GUO Hui, E-mail: panl-gh@sina.com)

Abstract: The standard support vector machines (SVM) formulation is modified by considering equality constraints within a form of ridge regression instead of inequality constraints. The solution can be obtainsed from solving a set of linear equations instead of a quadratic programming problem. The kernel principal component analysis (KPCA) is applied to least squares support vector machines (LSSVM) for feature extraction. KPCA calculates principal component in high dimensional feature space. The way reduces dimensions of sample and regression is applied with the LSSVM. Simulation results show that the method proposed is effective and superior.

Key words: Kernel principal component analysis; Least squares support vector machines; Principal component; Feature extraction

1 引 言

Vapnik[1]于1995年提出一种新型统计学习方法——支持向量机(SVM), 它具有完备的统计学习理论基础和出色的学习性能, 已成为机器学习界的研究热点, 并在很多领域都得到了成功的应用. Suykens[2,3]提出最小二乘支持向量机(LSSVM), 这种方法采用最小二乘线性系统作为损失函数, 代替传统SVM 采用的二次规划方法, 并应用到模式识别和非线性函数估计.

核函数方法是把非线性变换后的高维空间的内积运算转换为原始空间中的核函数计算, 从而避免了直接在变换后的高维空间计算, 大大减少了计算量. 目前, 把核函数方法和其他方法混合建模, 并应用在过程工业中成为一种趋势. 王华忠等[4]提出混合PCA-SVM 方法进行软测量建模, 该方法结合了二者的优点, 用主元分析(PCA)提取特征, 然后用SVM 建模, 在对工业过程软测量建模中起到了较好的效果. 冯瑞等[5]提出了用模糊支持向量机分类算法对输入数据进行预处理, 得到多模型模糊隶属度, 再用模糊支持向量机建立多模糊估计器.

在建立模型问题上, 随着样本数目的增大, 所需的计算时间和空间存储资源都会成几何级数增加. 所以在系统模型问题中, 特征提取非常重要, 可以降

收稿日期: 2005-07-08; **修回日期**: 2005-11-23.
基金项目: 国家科技部攻关基金项目(2003EG113016); 北京市教委重点学科共建基金项目.
作者简介: 郭辉(1972—), 男, 长春人, 博士生, 从事机器学习、复杂系统建模等研究; 刘贺平(1951—), 男, 沈阳人, 教授, 博士生导师, 从事人工智能的研究.

低学习问题的复杂性, 提高学习算法的泛化性能, 简化学习模型. 本文提出利用核PCA (KPCA) 对数据进行特征提取, 消除数据的相关性和噪声, 提取包含样本数据信息的主元, 降低样本空间的维数, 这些新特征作为LSSVM 的输入, 并将该方法称为KPCA-LSSVM 建模. 仿真结果表明KPCA 特征提取后, LSSVM 的建模效果优于没有特征提取的效果, 从而验证了KPCA-LSSVM 方法的有效性.

2 核的主成分分析

基于核函数的PCA 方法不直接计算特征向量, 而是将其转化为求核矩阵的特征向量和特征值, 避免了在特征空间求特征向量, 而数据在特征向量上的投影转换为求核函数的线性组合, 这大大简化了计算量.

首先将样本 $x_k (k = 1, \cdots, n, x_k \in R^N)$ 映射到特征空间 $\Phi(x_k)$, 计算协方差矩阵[6] 如下:

$$C = \frac{1}{n}\sum_{j=1}^{n}\Phi(x_j)\Phi(x_j)^{\mathrm{T}}. \tag{1}$$

然后通过解特征值问题计算主成分, 可以找到 $\lambda > 0$ 和 $V \neq 0$ 满足

$$\lambda V = CV = \frac{1}{n}\sum_{j=1}^{n}(\Phi(x_j)V)\Phi(x_j). \tag{2}$$

进一步, 从式(2) 可以看出, 所有特征值非零的特征矢量必然映射数据的张集上, 可以表示为

$$V = \sum_{i=1}^{n}\alpha_i\Phi(x_i). \tag{3}$$

将式(2) 左乘 $\Phi(x_k)$ 变为

$$\lambda(\Phi(x_k)V) = \Phi(x_k)CV, \quad k = 1, 2, \cdots, n. \tag{4}$$

定义一个 $n \times n$ 矩阵 K_{ij}, 有

$$K_{ij} = K(x_i, x_j) = \Phi(x_i)\Phi(x_j). \tag{5}$$

计算展开系数 α_i 的特征值问题仅取决于核函数, 即

$$n\lambda\alpha = K\alpha \tag{6}$$

其中 α 表示 $\alpha_1, \cdots, \alpha_n$ 中的一个列向量.

得到的解 (λ_k, α^k) 需要利用 $\lambda_k(\alpha^k\alpha^k) = 1$ 进行归一化处理. 接下来要提取一个测试样本 x 的特征, 只需将映射样本 $\Phi(x)$ 投影到 V^k 上[6], 有

$$s(i) = V^k \cdot \Phi(x) = \sum_{i=1}^{m}\alpha_i^k(\Phi(x_i)\Phi(x)) = \sum_{i=1}^{m}\alpha_i^k K(x_i, x), \quad i = 1, \cdots, m. \tag{7}$$

从以上方程可以看出, KPCA 提取的最大主元个数是 n, 如果前几个特征向量就能反映全部特征, 那么样本的主元数目可以减少, 往往提取的主元数目 $m < n$.

3 基于 KPCA-LSSVM 的建模方法

用KPCA 特征提取后, 得到 m 个主元, 训练数据的样本可以表示为 $(s_1, y_1), (s_2, y_2), \cdots, (s_m, y_m)$, 其中: y_i 是目标值, s_i 是提取后的输入向量. LSSVM 在优化目标中的损失函数为误差 ξ 的二次项, 使得约束条件变成了等式约束, 优化问题可以描述为求解如下问题[3]:

$$\min_{\omega, b, e} J(\omega, e) = \frac{1}{2}\omega^{\mathrm{T}}\omega + \gamma\frac{1}{2}\sum_{i=1}^{m}\xi_i^2,$$

$$\text{s.t. } y_i = \Phi(s_i)\cdot\omega + b + \xi_i, \quad i = 1, \cdots, m.$$

其中: $\Phi(\cdot): R^n \to R^{n_h}$ 是核函数(与 KPCA 中为同一核函数), 权矢量 $\omega \in R^{n_h}$, 误差变量 $\xi_i \in R$, b 是偏差量, γ 是可调参数. 核函数可以将原始空间中的样本映射为高维特征空间中的一个向量, 以解决线性不可分的问题, 用拉格朗日法求解此优化问题[3]

$$L(\omega, b, e; \alpha) = \frac{1}{2}\omega^{\mathrm{T}}\omega + \gamma\frac{1}{2}\sum_{i=1}^{m}\xi_i^2 - \sum_{i=1}^{m}\alpha_i(\omega^{\mathrm{T}}\Phi(s_i) + b + \xi_i - y_i).$$

其中 $\alpha_i (i = 1, \cdots, m)$ 是拉格朗日乘子. 根据优化条件

$$\frac{\partial L}{\partial \omega} = 0 \to \omega = \sum_{i=1}^{m}\alpha_i\Phi(s_i), \tag{8}$$

$$\frac{\partial L}{\partial b} = 0 \to \sum_{i=1}^{m}\alpha_i = 0, \tag{9}$$

$$\frac{\partial L}{\partial \xi_i} = 0 \to \alpha_i = \gamma\xi_i, \tag{10}$$

$$\frac{\partial L}{\partial \alpha_i} = 0 \to \omega^{\mathrm{T}}\Phi(s_i) + b + \xi_i - y_i = 0. \tag{11}$$

优化问题转化为求解线性问题

$$\begin{bmatrix} 0 & 1_v^{\mathrm{T}} \\ 1_v & \Omega + \frac{1}{\gamma}I \end{bmatrix}\begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix}. \tag{12}$$

其中

$$s = [s_1, \cdots, s_m], y = [y_1, \cdots, y_m], \alpha = [\alpha_1, \cdots, \alpha_n], \Omega_{ik} = \Phi(x_i)\Phi(x_k), \quad i, k = 1, \cdots, m.$$

定义核函数 $K(s_i, s_j) = \Phi(s_i)\Phi(s_j)$ 是满足Mercer 条件的对称函数. 最后的回归估计为

$$y(s) = \sum_{i=1}^{m}\alpha_i K(s, s_i) + b. \tag{13}$$

4 KPCA-LSSVM 方法的钢材淬透性预测建模

4.1 淬透性的原理

钢的淬透性是指钢在淬火冷却时, 获得马氏体

组织深度的能力[7]. 淬透性的概念是人们在发展淬火技术的过程中逐步形成的, 淬火的目的是通过一定速度的冷却使得钢件获得一定的表面硬度和一定深度的高硬度层, 以保证零件所需的使用性能

影响淬透性的因素很多, 包括钢的化学成分、奥氏体温度(冷却温度)、晶粒度的大小等, 其中主要的是化学成分. 以往计算淬透性采用多重回归的方法, 但受到钢种的约束, 计算精度有限, 这就极大地限制了模型的应用. 随着各种智能计算技术的发展, 陆续有学者将神经网络方法应用于淬透性的预报中, 从而使建立的模型摆脱机理的约束. 其中比较有代表性的工作包括: Vermenlen[8] 用 4 000 个样本训练神经网络, 输入为钢的化学成分和奥氏体化温度, 输出为 Jominy 硬度曲线, 由此建立的模型平均硬度值误差是 2; Dobrzanski[9] 在建立神经网络模型时, 先把钢种按化学成分进行分类, 然后分别对每一类建立模型, 输入为钢的 5 项化学成分含量, 输出为 Jominy 硬度曲线, 隐层数是 30, 网络结构是 5 × 30 × 15, 训练结果优于以上基于机理模型方法的结果. 本文采用 KPCA-LSSVM 方法预测建模, 首先用 KPCA 提取输入数据的主元, 然后用 LSSVM 进行建模

4.2　训练样本的获取和处理

建模采用的是某钢厂实际生产中搜集的数据, 采样时间从 2001 年 8 月 14 日到 2004 年 9 月 13 日. 原始数据保存在两张表中: 一张表中存放的是各炉钢中的 20 种化学成分的含量(%), 表结构是: 实验日期(syrq)、炉号(lh)、实际建模中用到的化学成分(C, Si, Mn, P, S, Mo, Ni), 每一炉测量一组化学成分, 共有 211 条数据; 另一张表中存放的是各炉的淬透性值(只包括 3 个距离端淬的值), 表结构是: 炉号(lh)、序号(xh)、淬透性值(J4#, J7#, J12#), 每一炉测量淬透性值 1～3 次, 共有 286 条数据. 为了建立回归模型, 对于一炉有多个淬透性值的数据, 只取出第一次测量的数据, 删除其他数据, 这样表中只剩下 211 条数据; 然后根据炉号将两张表做连接, 共得到 151 条数据; 最后, 由于 2001 年只有一条数据, 为了避免可能给生产工艺变化带来的影响, 将 2001 年的数据删除

为了消除量纲的影响或者突出某些属性的作用, 常将数据做标准化处理, 常见的标准化方法有最小-最大标准化和 Z-Score 标准化, 为避免孤立点带来的影响, 选择后一种方法. Z-Score 标准化的计算公式为

$$\overline{x_i} = \frac{1}{n}\sum_{j=1}^{n} x_i^j,\ \overline{x_i^j} = \frac{x_i^j - \overline{x_i}}{\sigma_i}, \tag{14}$$

$$\sigma_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_i^j - \overline{x_i})^2. \tag{15}$$

其中: $\overline{x_i}$ 和 σ_i 分别是属性的平均值和标准差; x_i^j 和 $\overline{x_i^j}$ 分别是属性的原始数据和标准化之后的数据; $j = 1, \cdots, n$ 是维数. 如果训练数据进行了 Z-Score 标准化运算, 在模型中将会记录下相关的信息, 在模型应用时, 会对测试数据进行同样的处理

4.3　结果与讨论

将所有的数据按照 N(训练集)∶N(测试集)≈3∶1 的标准进行划分, 既能保证提供足够的数据用来模型的学习, 又能验证模型的效果. 另外, 为了验证模型的健壮性和适应性, 对训练集数据采取随机采样的方法, 即每次从全体数据中随机地选择 3/4 作为训练集, 其余的数据作为测试集

本文采用 KPCA-LSSVM 方法进行淬透性建模, 并与 LSSVM、神经网络建模进行比较, KPCA-LSSVM 方法和 LSSVM 均选用 RBF 核函数, 并且核函数参数都通过 10 重交叉验证得到, 用 VC#.net 开发的工业过程建模平台中编程实现算法. 其中 KPCA-LSSVM 方法参数为: 核宽度 Gamma = 0.075, γ = 12; LSSVM 参数为: 核宽度 Gamma = 0.028, γ = 7.

图 1 表示 KPCA-LSSVM 方法淬透性建模示意图. 分别训练各自的模型, 这样得到 3 个模型: J4# 7, J7# 9 和 J12# 7. KPCA-LSSVM 方法中, KPCA 提取的主元数目为 25.

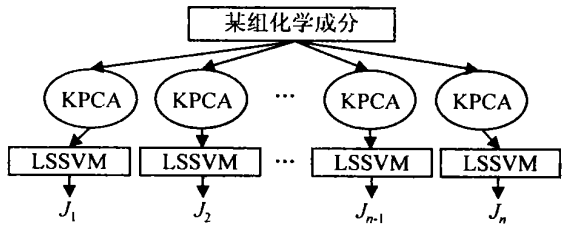

图 1　KPCA-LSSVM 方法淬透性模型预测示意图

表 1 为 KPCA-LSSVM, LSSVM 和神经网络(NN)模型预测误差 E 的统计信息. 这里的误差 E 是指(真实值 - 模型预报值)的绝对值, 范围划分为 6 个等级. 对于 J4# 7, KPCA-LSSVM 模型有 15 个预测值的 E 落在 0～1 区间内, 没有 E > 3 的预测值; LSSVM 模型有 5 个预测值的 E 落在 0～1 区间内, 有多达 7 个预测值的 E > 3; 而 NN 模型只有 3 个预测值的 E 落在 0～1 区间内. 对于 J7# 9, KPCA-LSSVM 模型有 8 个 E 落在 0～1 之间, 3 个值大于 3, LSSVM 模型有 10 个 E 落在 0～1 之间, 5 个 E 值大于 3, 而 NN 模型有 5 个 E 落在 0～1 区间内. 对于 J12# 7, KPCA-LSSVM 模型有 13 个 E

落在0～1之间,有2个值大于3,LSSVM模型有11个E落在0～1之间,有3个E值大于3,而NN模型有7个E落在0～1区间内,有7个E值大于3.总体来看,KPCA-LSSVM的模型预测效果最好,优于其他两种预测模型

表1 误差E的统计信息

		J4#7	J7#9	J12#7
0～1	KPCA-LSSVM	15	8	13
	LSSVM	5	10	11
	NN	3	5	7
1～2	KPCA-LSSVM	7	11	6
	LSSVM	6	9	10
	NN	5	4	7
2～3	KPCA-LSSVM	3	3	4
	LSSVM	7	1	1
	NN	7	6	4
3～4	KPCA-LSSVM	0	2	0
	LSSVM	2	4	1
	NN	6	7	6
4～5	KPCA-LSSVM	0	1	1
	LSSVM	3	0	1
	NN	1	2	1
>5	KPCA-LSSVM	0	0	1
	LSSVM	2	1	1
	NN	3	1	0

准确的表征模型性能的指标是ASE(asymptote atandard error),即渐近标准误差计算如下:

$$\mathrm{ASE}=\frac{\text{误差平方和}}{\text{样本总个数}},$$

ASE越小,说明拟合程度越高.表2列出了KPCA-LSSVM,LSSVM和NN模型预测的ASE计算结果,从表2中可以看出,KPCA-LSSVM模型在3点处的ASE均小于LSSVM和NN模型,尤其在J4#7点,而且随着淬火端距离的逐渐增加,相应的ASE值是逐渐增大的,但是NN的ASE变化比KPCA-LSSVM和LSSVM的ASE大,而KPCA-LSSVM方法随淬火端的距离变化时,误差变化比较缓慢,受淬火端的距离影响较小.

图2直观地显示了KPCA-LSSVM方法在J4#7点的预测效果

表2 ASE计算结果模型

ASE	J4#7	J7#9	J12#9
NN	7.35	11.73	14.28
LSSVM	5.21	8.49	10.17
KPCA-LSSVM	1.73	3.12	3.35

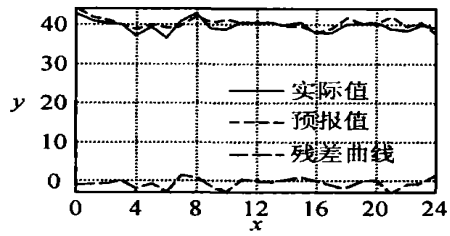

图2 J4#7的KPCA-LSSVM模型预测效果

5 结 论

本文提出了一种KPCA-LSSVM的建模方法,将核PCA算法与最小二乘支持向量机结合起来,用核PCA提取的主元数量作为最小二乘支持向量机训练数据,有效地降低了样本的维数,提高了训练的泛化能力.文中将该方法应用于钢材淬透性预测建模中,仿真结果表明了该方法的可行性和有效性,并且与单独采用最小二乘支持向量机建模、神经网络建模相比,该方法具有良好的建模效果

参考文献(References)

[1] Vapnik V N. *Statistical Learning Theory* [M]. New York, 1998.

[2] Suykens J A K. Nonlinear Modeling and Support Vector Machines [A]. *IEEE Instrumentation and Measurement Technology Conf* [C]. Budapest, 2001: 108-119.

[3] Suykens J A K, Vandewalle J. Least Squares Support Vector Machine Classifiers [J]. *Neural Processing Letters*, 1999, 9(3): 293-300.

[4] Wang H Z, Yu J S. Soft Sensor Modelling and Application Based on PCA-SVM [J]. *Process Automation Instrumentation*, 2004, 25(2): 30-32.

[5] Feng R, Shen W, Zhang Y Z, et al. Mutiple Modeling Approach Using Fuzzy Support Vector Machines [J]. *Control and Decision*, 2003, 18(6): 646-650.

[6] Scholkopf B, Smola A J, Müller K-R. Nonlinear Component Analysis as a Kernel Eigenvalue Problem [J]. *Neural Computing*, 1998, 9(3): 1299-1319.

[7] Wu J X. *Application of Steel Harden-ability* [M]. Beijing: Mechanism Industry Publishing Company, 1994: 15-50.

[8] Vermeulen W G, Van der Wolk, Weijer de A P, et al. Prediction of Jominy Hardness Profiles of Steels Using Artificial Neural Networks [J]. *J of Materials Engineering and Performance*, 1996, 5(1): 57-63.

[9] Dobrzanski L A, Sitek W. Application of a Neural Network in Modeling of Harden Ability of Constructional Steels [J]. *J of Materials Processing Technology*, 1998, 78(3): 59-66.