

基于最小二乘支持向量机对偶优化问题的核偏最小二乘

郭辉 刘贺平 王玲

北京科技大学信息工程学院, 北京 100083

摘要 提出了一种基于对偶优化的核最小二乘(KPLS)方法,把KPLS用最小二乘支持向量机的形式表示.推导了KPLS对偶优化形式的公式,且使其具有最小二乘支持向量机的风格.在初始空间中构造优化问题,应用核技术在特征空间中解对偶问题,这种解与非线性的KPLS具有相似性.实验验证了这种方法的效果,表明了该方法的有效性和优越性.

关键词 优化问题; 偏最小二乘; 最小二乘支持向量机; 核偏最小二乘

分类号 TP301.5

Vapnik在1995年提出一种新型统计学习方法—支持向量机(support vector machines, SVM),支持向量机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,如人脸检测、手写体数字识别、文本自动分类等^[1-2].近年, Suykens提出最小二乘支持向量机方法(LS-SVM)^[3-4],这种方法采用最小二乘线性系统作为损失函数,求解过程变成求解一组等式方程,求解速度相对加快,并应用到模式识别和非线性函数估计中,取得了较好的效果.

偏最小二乘(PLS)是一种新型的多元统计数据分析方法,近十几年来,它的理论、方法和应用都得到了迅速地发展,能较好地处理基于传统最小二乘回归方法难以解决的问题^[7-8].偏最小二乘回归集多元线性回归分析、典型相关分析和主成分分析的基本功能为一体,将建模类型的预测分析方法与非模型式的数据内涵分析方法有机地结合起来:一方面通过数据分析寻找因变量和自变量之间的函数关系,建立模型进行预测;另一方面,通过数据分析简化数据结构,观察变量间的相互关系^[9].核函数方法是当前一个十分活跃的研究领域,这个方法就是用非线性变换 $\phi(\cdot)$ 将 m 维矢量 x 映射到高维特征空间,在高维特征空间

中设计的线性学习算法.通过引入核函数,非线性PLS可以发展为核偏最小二乘(KPLS).KPLS首先把初始输入映射到高维特征空间,然后在高维特征空间中计算主元^[5],但是这种核方法不含正则化表示.

本文提出了一种简单可行的方法,把KPLS用最小二乘支持向量机的形式表示.推导出KPLS对偶优化形式的公式,并且具有最小二乘支持向量机的风格.在初始空间中构造优化问题,应用核技术在特征空间中解对偶问题,这种解与非线性的KPLS具有相似性.过实验验证了这种方法的可行性,得到较好的效果.

1 偏最小二乘

偏最小二乘回归适用于多自变量和多因变量系统.该方法最早应用在化学领域,由Woldh和Albano等人于1983年首次提出.近年来在理论、方法和应用上都取得了很大的发展.基本做法是:首先在自变量集中提取第一成分 v_1 (v_1 是 $x_1, x_2, \dots, x_m$ 的线性组合,其中 x 是输入向量,且尽可能多地提取原自变量集中的变异信息);同时在因变量集中也提取第一成分 w_1 , w_1 是输出变量,并要求 v_1 与 w_1 相关程度达最大.然后建立因变量 $y_1, y_2, \dots, y_p$ 与 v_1 的回归,如果回归方程已达到满意的精度,则算法终止;否则继续对第二成分的提取,直到能达到满意的精度为止.

偏最小二乘是在连续线性组合 (v, x) 与 (w, y) 之间寻找具有最大的协方差.可以用下面的方

收稿日期: 2005-04-25 修回日期: 2005-09-22

基金项目: 国家“863计划”资助项目(No. 2002 AA412010-10)和国家科技部攻关项目(No. 2003EG113016)

作者简介: 郭辉(1973—)男,博士研究生;刘贺平(1951—),男,教授,博士

程表示^[6-8]:

$$\begin{aligned} \max_{\mathbf{v}, \mathbf{w}} \quad & \text{cov}(\mathbf{v}^T \mathbf{x}, \mathbf{w}^T \mathbf{y}) = \mathbf{v}^T \mathbf{C}_{xy} \mathbf{w} \\ \text{s. t.} \quad & \|\mathbf{v}\| = 1 = \|\mathbf{w}\|. \end{aligned} \tag{1}$$

其中 $\mathbf{C}_{xy} = \mathbf{x}^T \mathbf{y}$, 是 $p \times p$ 的样本协方差矩阵.

引入拉格朗日乘子, 得到拉格朗日方程, 解优化方程可以得到:

$$\mathbf{C}_{xy} \mathbf{w} = \lambda \mathbf{v}, \quad \mathbf{C}_{yx} \mathbf{v} = \lambda \mathbf{w} \tag{2}$$

其中 $\mathbf{C}$ 是正定的、对称的, $\mathbf{v}$ 和 $\mathbf{w}$ 是最大特征值对应的具有最大方差的特征向量. 从维度减少的观点看, 偏最小二乘可以看作是以最小二乘方式搜索最适度空间. 偏最小二乘的最小二乘损失函数表示为:

$$\begin{aligned} J(\mathbf{v}, \mathbf{w}) = \sum_{i=1}^p (& \|\mathbf{x}_i - \mathbf{v} \mathbf{v}^T \mathbf{x}_i\|^2 + \|\mathbf{v}^T \mathbf{x}_i - \\ & \mathbf{w}^T \mathbf{y}_i\|^2 + \|\mathbf{y}_i - \mathbf{w} \mathbf{w}^T \mathbf{y}_i\|^2) \end{aligned} \tag{3}$$

可见, 偏最小二乘是在于优化系数向量 $\mathbf{v}$, $\mathbf{w}$, 同时分别在样本输入 $\mathbf{x}$ 空间、输出 $\mathbf{y}$ 空间搜索数据点的最大映射, 并且保持二者具有最大协方差矩阵.

2 核偏最小二乘的对偶优化问题

核函数方法用于偏最小二乘, 提出了基于核的偏最小二乘方法(KPLS), 从而将原本用于线性相关分析的PLS方法扩展到了非线性相关分析的领域^[4]. 通过简化式(3), 把数据映射到特征空间中, 并且加入正则化参数 γ , 得到下面的最小二乘优化公式:

$$\max J_{\text{PLS}}(\mathbf{v}, \mathbf{e}) = \gamma \sum_{i=1}^n e_i r_i - \frac{1}{2} \mathbf{v}^T \mathbf{v} - \frac{1}{2} \mathbf{w}^T \mathbf{w}$$

约束条件:

$$e_i = \mathbf{v}^T \phi(\mathbf{x}_i), \quad r_i = \mathbf{w}^T \phi(\mathbf{y}_i), \quad i = 1, \dots, p \tag{4}$$

其中正则化参数 $\gamma > 0$, e_i 和 r_i 是误差项. 给出的优化问题具有相应的对偶公式, 现在引进拉格朗日乘子 α_i, β_i , 根据目标函数及约束条件建立拉格朗日方程为:

$$\begin{aligned} L(\mathbf{v}, \mathbf{w}, e_i, r_i, \alpha_i, \beta_i) = & \gamma \sum_{i=1}^p e_i r_i - \frac{1}{2} \mathbf{v}^T \mathbf{v} - \\ & \frac{1}{2} \mathbf{w}^T \mathbf{w} - \sum_{i=1}^n \alpha_i [e_i - \mathbf{v}^T \phi(\mathbf{x}_i)] - \\ & \sum_{i=1}^n \beta_i [r_i - \mathbf{w}^T \phi(\mathbf{y}_i)] \end{aligned} \tag{5}$$

利用拉格朗日优化方法, 极值应满足下面条件:

$$\frac{\partial L}{\partial \mathbf{v}} = 0 \rightarrow \mathbf{v} = \sum_{i=1}^p \alpha_i \phi(\mathbf{x}_i) \tag{6}$$

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \rightarrow \mathbf{w} = \sum_{i=1}^p \beta_i \phi(\mathbf{y}_i) \tag{7}$$

$$\frac{\partial L}{\partial e_i} = 0 \rightarrow \alpha_i = \gamma r_i \quad i = 1, \dots, p \tag{8}$$

$$\frac{\partial L}{\partial r_i} = 0 \rightarrow \beta_i = \gamma e_i \quad i = 1, \dots, p \tag{9}$$

$$\frac{\partial L}{\partial \alpha_i} = 0 \rightarrow e_i = \mathbf{v}^T \phi(\mathbf{x}_i) \quad i = 1, \dots, p \tag{10}$$

$$\frac{\partial L}{\partial \beta_i} = 0 \rightarrow r_i = \mathbf{w}^T \phi(\mathbf{y}_i) \quad i = 1, \dots, p \tag{11}$$

在消去变量 $\mathbf{e}, \mathbf{r}, \mathbf{v}, \mathbf{w}$, 并且定义 $\lambda = \frac{1}{\gamma}$, 可以得到下面的对偶问题:

$$\begin{bmatrix} 0 & \Omega_{c,2} \\ \Omega_{c,1} & 0 \end{bmatrix} \begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix} = \lambda \begin{bmatrix} \mathbf{I} & 0 \\ 0 & \mathbf{I} \end{bmatrix} \begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix} \tag{12}$$

其中,

$$\Omega_{c,1_{i,j}} = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j), \quad \Omega_{c,2_{i,j}} = \phi(\mathbf{y}_i)^T \phi(\mathbf{y}_j).$$

在特征空间中用非线性主元构造线性模型, 定义 $k(\mathbf{x}, \mathbf{y}) = \Phi(\mathbf{x})^T \Phi(\mathbf{y})$, 从上面推导的优化条件, 可以得到 KPLS 回归模型:

$$\begin{aligned} f(\mathbf{x}) = & \mathbf{w}^T \phi(\mathbf{x}) + b = \\ & \sum_{j=1}^s \mathbf{w}_j \left[\sum_{i=1}^n \alpha_{ij}^{(n)} k(\mathbf{x}_i, \mathbf{x}) \right] + b \\ & \sum_{i=1}^s \alpha_i' k(\mathbf{x}_i, \mathbf{x}_j) + b \end{aligned} \tag{13}$$

s 是成分个数, 其中 $\alpha_i' = \sum_{j=1}^s \mathbf{w}_j \alpha_{ij}^{(m)}$.

3 实验与分析

设计两个实验进一步分析核偏最小二乘的对偶优化表示问题: Mackey-Glass 时间序列问题和标准数据集测试.

3.1 Mackey-Glass 时间序列分析

混沌是自然界与人类社会普遍存在的运动形式, 对于混沌动力学系统的研究已成为动力系统研究的中心内容之一, 并常常用来检验非线性系统模型的性能. 在实验中对 Mackey-Glass 时间序列进行预测. Mackey-Glass 时间序列由差分延迟方程产生, 该方程定义如下:

$$\frac{dx(t)}{dt} = -bx(t) + \frac{ax(t-t_d)}{1+x^{10}(t-t_d)} \quad \text{for } t_d > \tau \tag{14}$$

其中: $a = 0.2$, $b = 0.1$, t_d 是延迟时间, t 是当前时间. 当 $\tau > 17$ 时, 动力学系统变为混沌状态, 因此选择 $\tau = 30$. 实验任务是使用已知 x 在 t 点的值预测将来 x 在 $t + \tau$ 的值. 从 $x(501) \sim x(700)$ 抽取 200 个数据构成训练数据集, 从 $x(701) \sim x(1000)$ 取 300 个数据构成测试数据集. 为了使

最小二乘支持向量机选取最优参数, 用交叉验证方法, 得到优化的参数值, 采用高斯核函数, 通过 5 折交叉检验, 得到 $\sigma=17.2$, $\gamma=15$. 采用高斯函数作为 KPLS 的核函数, 这是由于在实验中这种核函数效果最好. 用标准均方误差(NMSE) 作为评价准则, 定义如下:

$$MSE=\sqrt{\frac{1}{k}\sum_{i=1}^k(x_i-\hat{x}_i)^2}.$$

其中 x_i , $\hat{x}_i$ 分别为第 n 个时间序列实际值和预测值. 用 S-KPLS 表示对偶优化的 KPLS. 从图 1 和图 2 来看对偶优化的 KPLS 的 MSE 误差小于 KPLS 的 MSE, 且变化趋于平缓, 模型精度较好, 具有很好的泛化性能.

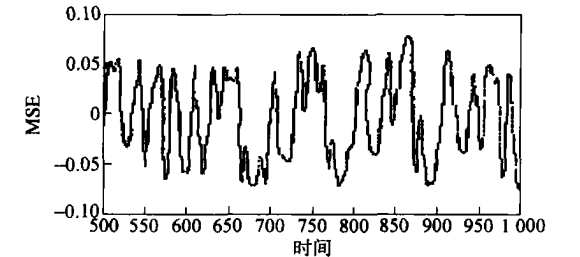


图 1 KPLS 的均方误差
Fig. 1 Mean square error of KPLS

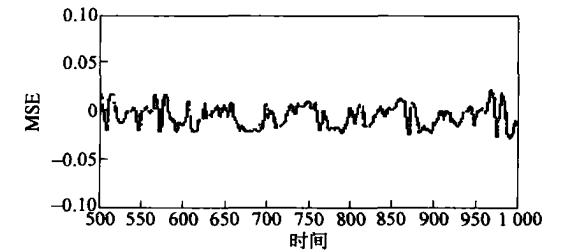


图 2 对偶优化的 KPLS(S-KPLS) 的均方误差
Fig.2 Mean square error of dual optimization KPLS

3. 2 标准数据集

采用 UCI 机器学习数据库中的数据集进行实验, 它提供了大量的、可被重复验证和对比的、目前公认的真实数据. 实验数据集的描述和测试方法如表 1 所示. 在所有的数据集中均采用高斯核函数, 最小二乘支持向量机选取最优参数, 通过 5 折交叉检验, 得到核宽度参数 $\sigma=9.2$, $\gamma=12$. 用与上面相同的 MSE 作为评价准则, 比较了最小二乘支持向量机、KPLS、对偶优化的 KPLS 三种方法的 MSE. 表 2 为得到的结果, 其中对偶优化的 KPLS 方法用 S-KPLS 表示. 从表 2 中可以看出, 最小二乘支持向量机与对偶优化 KPLS 方法的均方误差(MSE) 很接近, 而 S-KPLS 的均方

误差(MSE) 小于 KPLS, 说明这种对偶优化表示能得到比较好的结果.

表 1 数据集描述
Table 1 Data sets description

数据集合	尺寸	维数	训练集	测试集
波士顿房屋	506	13	471	35
计算机属性	8 192	21	7 500	692
身体肥胖	253	14	200	53
空间群	3 108	7	2 000	1 108
嘧啶胺	75	27	70	5
鲍鱼	4 177	9	3 000	1 177
三嗪基	187	36	100	87

表 2 三种方法的均方误差的比较
Table 2 Comparison of normalized mean square error among the three methods

数据集合	LSSVM	KPLS	S-KPLS
波士顿房屋	9. 1	11. 0	8. 7
计算机属性	29. 2	38. 4	29. 8
身体肥胖	3. 11	5. 96	2. 05
空间群	12. 5	26. 0	12. 9
嘧啶胺	2. 24	4. 57	2. 15
鲍鱼	18. 4	35. 0	18. 4
三嗪基	1. 40	5. 66	1. 28
误差平均值	10. 85	16. 49	10. 91

4 结论

本文提出了一种基于对偶优化的 KPLS 方法, 推导出的公式具有最小二乘支持向量机的风格, 并通过实验验证了这种方法. 从实验中可以看出, 最小二乘支持向量机与对偶优化的 KPLS 方法的均方误差(MSE) 很接近, 而与单独采用 KPLS 方法相比, 能得到较好的效果.

参 考 文 献

[1] Vapnik V N. 统计学习理论的本质. 张学工, 译. 北京: 清华大学出版社, 2000

[2] Burges C J C. A tutorial on support vector machines for pattern recognition. *Data Min Knowl Discovery*, 1998, 2(2): 121

[3] Johan S A K. Nonlinear modeling and support vector machines //IEEE Instrumentation and Measurement Technology Conference. Budapest, 2001

[4] Suykens A K, Vandewalle. Least squares support vector machine classifiers. *Neural Process Lett*, 1999, 9(3): 293

[5] Trejo R. Kernel partial least squares regression in reproducing

kernel Hilbert space. **J Mach Learning Res**, 2001, 2: 97

[6] RÄannar B, Lindgren J, Geladi S, et al. A PLS kernel algorithm for data sets with many variables and fewer objects. **Chemometrics Intell Lab Syst**, 1994, 8: 111

[7] HÄoskuldsson A. PLS regression methods. **J Chemometrics**, 1988, 2: 211

[8] Barker B E, Rayens I M. Partial least squares for discrimination. **J Chemometrics**, 2003, 17: 166

Kernel partial least squares based on least squares support vector machine primal-dual optimization problem

GUO Hui, LIU Heping, WANG Ling

Information Engineering School, University of Science and Technology Beijing, Beijing 100083, China

ABSTRACT A kernel partial least squares (KPLS) method based on dual optimization was proposed, which was expressed by least squares support vector machine. The KPLS formulae in the form of dual optimization were deduced, which had the style of least squares support vector machine. The optimization problem was constructed in a prime space, the dual problem was solved in a eigenspace by the kernel skill and the solutions were the same as nonlinear KPLS. The model was illustrated with some examples. The results show that the proposed method is effective and superior.

KEY WORDS optimization problem; partial least squares (PLS); least squares support vector machines (LS-SVM); kernel partial least squares (KPLS)

Microstructure of flow pattern defects in boron-doped Czochralski-grown silicon

LIU Caichi¹⁾, HAO Qiuyan¹⁾, ZHANG Jianfeng¹⁾, TENG Xiaoyun¹⁾, Sun Shilong¹⁾, Qigang Zhou²⁾, WANG Jing²⁾, XIAO Qinghua²⁾

1) Institute of Information Functional Materials, Hebei University of Technology, Tianjin 300130, China
2) General Research Institute of Non-ferrous Metals, Beijing 100088, China

ABSTRACT The morphology and microstructure of flow pattern defects (FPDs) in lightly boron-doped Czochralski-grown silicon (Cz-Si) crystals were investigated using optical microscopy and atomic force microscopy. The experimental results showed that the morphology of FPDs was parabola-like with several steps. Single-type and dual-type voids were found on the tip of FPDs and two heaves exist on the left and right sides of the void. All the results have proved that FPDs were void-type defects. These results are very useful to investigate FPDs in Cz-Si wafers further and explain the annihilation of FPDs during high-temperature annealing.

KEY WORDS flow pattern defects; grow n-in defects; atomic force microscopy; Czochralski-grown silicon

[摘自 Rare Metals, 2006, 25(4) : 389]