基于支持向量机的模型传递方法研究

熊宇虹[1,2]，温志渝[1]，梁玉前[1]，陈 勤[1]，张 波[1]，刘 妤[1]，向贤毅[1]

1. 重庆大学光电工程学院，重庆 400044
2. 南昌大学计算机科学与技术系，江西 南昌 330031

摘 要 模型传递是以数学方法通过在2台不同仪器之间寻求一种变换关系来增强光谱仪数据通用性、可比性的一种基本途径。由于实际测量数据具有非线性特征，加上校正样本集合的有限性，使得解决小样本条件下非线性关系的模型传递问题显得尤为重要。文章在概述支持向量机基本原理的基础上，探讨了支持向量机方法在光谱仪的模型传递问题中的应用，提出了基于支持向量机的分段直接校正方法，最后采用计算机模拟的方式对该方法进行了举例说明，并和人工神经网络方法进行了相应的比较。

关键词 光谱分析；模型传递；支持向量机
中图分类号：TP39 **文献标识码**：A **文章编号**：1000-0593(2007)01-0147-04

引 言

在光谱仪的实际使用过程中，不同仪器由于光路设计、部件选用和装配误差等原因，使得测量数据间存在着一定的差异，导致一台光谱仪上所测量的大量光谱数据不能直接应用在另一台仪器上，使得仪器间的通用性和可比性差。模型传递正是以数学方法通过在2台不同仪器之间寻求一种变换关系来克服解决该问题的一种基本途径。

目前常用的模型传递方法有斜率偏差算法、直接校正或分段直接校正的偏最小二乘法或神经网络法、有限脉冲响应算法等[1]。从实际应用的效果来看，经典的方法对解决线性关系的模型传递问题效果较好，而对解决小样本条件下非线性关系的模型传递问题结果却不尽人意。本文以解决小样本条件下的非线性关系的模型传递问题为目标，探讨了支持向量机方法在模型传递问题中的应用。在概述支持向量机的基本原理的基础上，提出了基于支持向量机的分段直接校正方法，建立了算法的基本流程，最后采用计算机模拟的方式对该方法进行了举例说明，并和人工神经网络方法的结果进行了相应的比较。

1 支持向量机的基本原理

支持向量机是Vapnik等根据统计学习理论提出的一种建立在结构风险最小化原则的基础上，专门研究小样本情况下统计估计和预测的问题，探索在现有有限样本的情况下如何得到最优解的通用学习方法，它体现了兼顾经验风险和置信范围的一种折中的思想，能较好地解决小样本、非线性、高维数等实际问题，其基本思想可用2类线性可分问题进行说明[2-4]。

图1中，○和□分别代表2类样本；H为最优分类线，不仅能将2类样本无错误的分开而且使分类间隔最大；H1，H2分别为过各类中离分类线最近的样本且平行于分类线的直线，它们之间的距离叫做分类间隔。H1，H2上的训练样本点就称作支持向量，相应的确定最优分类线的函数被称为支持向量机。

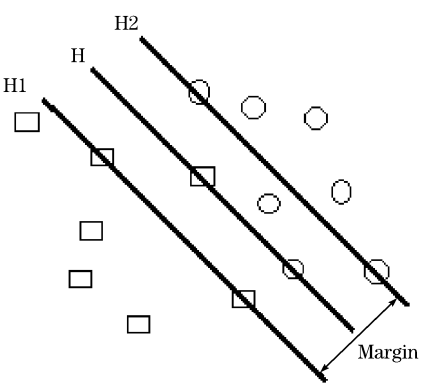

Fig. 1 Linearity separated two classes problem

收稿日期：2005-09-08，**修订日期**：2006-01-11
基金项目：国家自然科学基金项目(60308007)，国家“863”项目(2004AA404023)，科技部国际合作项目(2004DFA00600)和重庆市科委项目(CSTC，2005CF2002)资助
作者简介：熊宇虹，1971年生，重庆大学光电工程学院博士研究生，南昌大学计算机科学与技术系教师 e-mail：xyh341@sohu.com

支持向量机的方法可用于回归分析，其运用思路与在模式识别中类似。首先考虑用线性回归函数 $f(x)=<w\cdot x>+b$ 拟合数据(x_i, y_i)，$x_i\in R^n$，$y_i\in R$，$i=1,\cdots l$ 的问题，并假设所有训练数据都可以在精度 ε 下无误差地用线性函数拟合，即

$$\begin{cases} y_i-w\cdot x_i-b\leqslant\varepsilon \\ w\cdot x_i+b-y_i\leqslant\varepsilon \end{cases} \quad i=1,\cdots l \tag{1}$$

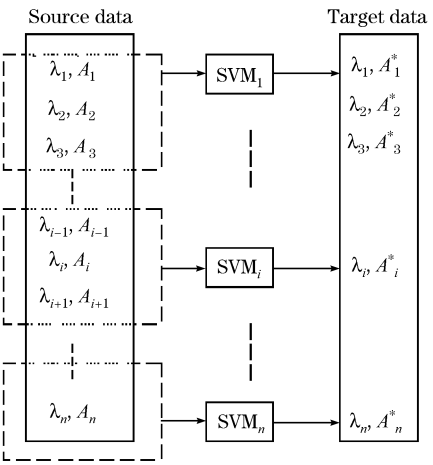

Fig. 2 Algorithm based on support vector machine

考虑到允许拟合误差的情况，引入松弛因子 $\xi_i\geqslant 0$ 和 $\xi_i^*\geqslant 0$，则条件(1)变成

$$\begin{cases} y_i-w\cdot x_i-b\leqslant\varepsilon+\xi_i \\ w\cdot x_i+b-y_i\leqslant\varepsilon+\xi_i^* \end{cases} \quad i=1,\cdots l \tag{2}$$

优化目标变成最小化 $\frac{1}{2}\|w\|^2+C\sum_{i=1}^{l}(\xi_i+\xi_i^*)$，常数 $C>0$ 控制对超出误差 ε 的样本的惩罚程度。采用优化的方法可以得到其对偶问题。在式(3) 的条件下，对 Lagrange 因子 a_i，a_i^* 最大化式(4)所示的目标函数。

$$\sum_{i=1}^{l}(a_i-a_i^*)=0,\ 0\leqslant a_i,\ a_i^*\leqslant C,\ i=1,\cdots l \tag{3}$$

$$W(a_i, a_i^*)=-\varepsilon\sum_{i=1}^{l}(a_i^*+a_i)+\sum_{i=1}^{l}y_i(a_i^*-a_i)-\frac{1}{2}\sum_{i,j=1}^{l}(a_i^*-a_i)(a_j^*-a_i)<x_i\cdot x_j>,\ j=1,\cdots,l \tag{4}$$

得回归函数为

$$f(x)=<w\cdot x>+b=\sum_{i=1}^{l}(a_i-a_i^*)<x_i\cdot x>+b^* \tag{5}$$

这里 a_i，a_i^* 也将只有小部分不为 0，它们对应的样本就是支持向量，一般是在函数变化比较剧烈的位置上的样本，而且这里也是只涉及内积运算，只要用核函数 $K(x_i, x_j)$ 替代(4)、(5) 式中的形如 $<x_i\cdot x_j>$ 的内积运算就可以实现非线性函数拟合，其中，常用的核函数有多项式核函数、径向基函数(RBF)核函数、Sigmoid 核函数等。

2 基于支持向量机的校正算法

模型传递的基本过程是在 2 台仪器上(源机和目标机)同时测定某些样品的光谱，通过这些样品的光谱数据得出 2 台仪器间(源机和目标机)的光谱传递关系，从而根据源机其他的光谱数据和获得的传递关系推测出目标机相应的光谱数据。

对非线性模型传递问题而言，通常采用神经网络方法，其基本的应用方法是以源机的吸光度数据为输入，目标机测得的相应吸光度数据为输出，建立并训练网络，进而对目标机其他未知光谱进行预测。由于神经网络本身是基于样本趋于无限大情况下的学习算法，而且在网络结构设计、参数选择上都存在着一些经验性的因素，稍有不慎，就会产生欠学习或过学习，使得预测效果变差，即使采用减少网络规模的分段直接校正的神经网络方法也还一定程度的存在这种问题。

支持向量机是基于小样本理论的学习方法，已经表现出许多优于神经网络的特性，在许多方面有望替代神经网络，正是基于这种考虑，本文把支持向量机应用到模型传递问题中。为简化计算，采用基于支持向量机的分段直接校正法对目标机相应的光谱数据进行预测，其具体的算法过程如下。

(1)将源机 i 波长附近 $i-k$ 到 $i+k$ 波长段对应的吸光度作为支持向量机的输入向量；目标机 i 波长处的吸光度作为支持向量机的输出值，建立并训练支持向量机。对最开始的前 k 个波长和最后的 k 个波长，可适当采用一些变通的方式进行处理。

(2)将窗口沿波长移动从而得到关于整个波长段的一个支持向量机集合，通过这个支持向量机集合就能把源机的光谱数据转换为目标机的光谱数据。

图 2 是一个窗口范围为 3 的算法示意模型，当时 $i=2$，3，…，$n-1$ 以源机 i 波长附近 $i-1$ 到 $i+1$ 波长段对应的吸光度作为输入向量；目标机 i 波长处的吸光度作为输出值，建立并训练支持向量机；当 $i=1$ 时，以 1 到 3 波长段对应的吸光度作为输入向量，目标机 1 波长处的吸光度作为输出值，当 $i=n$ 时，以 $n-2$ 到 n 波长段对应的吸光度作为输入向量，目标机 n 波长处的吸光度作为输出值，分别建立并训练支持向量机。

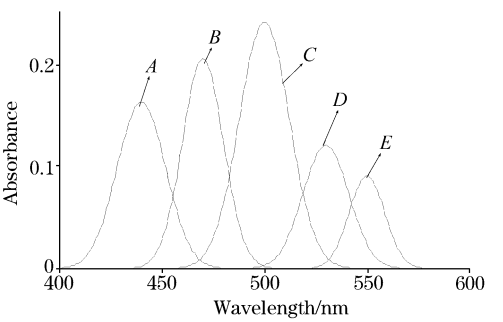

Fig. 3 Single component spectrum(*A*, *B*, *C*, *D*, *E*)

3　实例分析

为了简化程序，方便讨论，采用计算机模拟的方式产生各种所需的原始数据。假设以一组五组分(A，B，C，D，E)混合样品体系为例进行分析(见图 3)，首先利用高斯函数分别得出 5 个组分的单位光谱，进而运用均匀设计的方法设计多个不同浓度组成的量测样品，利用多组分加和性原理，就得出了各个样品的吸光度曲线，以此作为样品的真实测量数据，也认为是样品的源机的测量数据；以源机的测量数据为基础，通过一定的非线性变换就得到了目标机的测量数据。

在本文中模拟产生了 8 个样品，其各成分的浓度组成关系如表 1 所示，以表 1 中，1，2，3，4 行所示的 4 个样品作为校正样本集合，5，6，7，8 行所示的 4 个样品作为检验样本集合，从机光谱数据可按式子 $y=x+0.5x^2$ 生成，其中 x 为主机某波长点的吸光度，y 为目标机相应波长点的吸光度。以波长为 455 nm、窗口大小为 3 时的数据进行分析。

源机的校正样本光谱数据和检验样本光谱数据如表 2 所示，目标机数据可根据 $y=x+0.5x^2$ 求出。利用上述数据，分别运用神经网络(ANN)和支持向量机方法(SVM)利用校正样本集进行模型训练，然后用检验样本集对模型进行检验。对支持向量机方法，经比较发现采用二阶多项式核函数时结果较为理想。表 3 为神经网络(ANN)和支持向量机(SVM)模型对检验样本集的预测误差表，从中可以得出，支持向量机方法优于神经网络方法(注：采用的神经网络模型是经过了优化设计的)。

Table 1　Table of samples' concentration

No	A	B	C	D	E
1	1	2	4	5	7
2	2	4	8	1	5
3	3	6	3	6	3
4	4	8	7	2	1
5	5	1	2	7	8
6	6	3	6	3	6
7	7	5	1	8	4
8	8	7	5	4	2

Table 2　Spectral source date

Class	Calibration samples				Test samples			
No	1	2	3	4	1	2	3	4
454 nm	0.194 7	0.389 4	0.582 3	0.777 0	0.459 4	0.654 1	0.846 9	1.041 6
455 nm	0.206 1	0.412 2	0.615 8	0.821 9	0.429 5	0.635 6	0.839 1	1.045 3
456 nm	0.219 6	0.439 1	0.655 3	0.874 9	0.402 0	0.621 5	0.837 7	1.057 3

4　结束语

模型传递是增强光谱仪数据通用性和可比性的一种基本方法，模型传递问题是光谱仪产品化过程中应该解决的重要问题之一。如何较好地解决小样本条件下非线性关系的模型传递问题是实际运用中的难点，本文探讨了支持向量机方法在模型传递中的应用，提出了基于支持向量机的分段直接校正方法，并对该方法进行了实例分析，结果表明该方法和人工神经网络方法相比具有预报误差更小的特点。

Table 3　Comparison of predicted error

No	1	2	3	4	Average
ANN/%	1.498 9	1.030 0	0.860 7	1.390 6	1.195 1
SVM/%	0.180 9	0.107 1	0.087 2	0.064 7	0.110 0

参 考 文 献

[1] CHU Xiao-li, YUAN Hong-fu, LU Wan-zhen(褚小立, 袁洪福, 陆婉珍). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2001, 21(6): 881.
[2] ZHANG Xue-gong(张学工). Acta Automatica Sinica(自动化学报), 2000, (1): 32.
[3] XIONG Yu-hong(熊宇虹). Acta Photonica Sinica(光子学报), 2005, 34(10): 1514.
[4] Vapnik Vladimir N, ZHANG Xue-gong(瓦普尼克, 张学工). The Nature of Statistical Learning Theory(统计学习理论的本质). Beijing: Tsinghua University Press(北京: 清华大学出版社), 2000. 21.

Model Transfer Method Based on Support Vector Machine

XIONG Yu-hong[1,2], WEN Zhi-yu[1], LIANG Yu-qian[1], CHEN Qin[1], ZHANG Bo[1], LIU Yu[1], XIANG Xian-yi[1]
1. College of Optoelectronic Engineering, Chongqing University, Chongqing 400044, China
2. Department of Computer Science and Technology, Nanchang University, Nanchang 330031, China

Abstract The model transfer is a basic method to build up universal and comparable performance of spectrometer data by seeking a mathematical transformation relation among different spectrometers. Because of nonlinear effect and small calibration sample set in fact, it is important to solve the problem of model transfer under the condition of nonlinear effect in evidence and small sample set. This paper summarizes support vector machines theory, puts forward the method of model transfer based on support vector machine and piecewise direct standardization, and makes use of computer simulation method, giving a example to explain the method and compare it with artificial neural network in the end.

Keywords Spectral analysis; Model transfer; Support vector machine

(Received Sep. 8, 2005; accepted Jan. 11, 2006)