第31卷，第5期
2011年5月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 31, No. 5, pp -
May, 2011

多任务最小二乘支持向量回归机及其在近红外光谱分析技术中的应用研究

徐 硕[1]，乔晓东[1]，朱礼军[1]，安 欣[2]，张录达[3]*

1. 中国科学技术信息研究所信息技术支持中心，北京 100038
2. 对外经济贸易大学国际经济与贸易学院，北京 100029
3. 中国农业大学理学院，北京 100193

摘 要 在近红外光谱定量分析中，许多模型分开考虑各种样品成分含量，失去了样品成分间潜在的联系。针对该问题，文章将建模分析每种样品成分含量的问题看作一个任务，将同时建模分析所有样品成分含量的问题转换为多任务学习问题。在LS-SVR的基础上，提出了多任务LS-SVR(MTLS-SVR)，并给出一种有效的大规模问题求解算法。最后，以高粱样品数据集为实验材料，建立了三种样品成分(蛋白质，赖氨酸及淀粉)的同时定量分析模型。三种样品成分的预测值与实际值的平均相对误差分别为1.52%，3.04%和1.01%，相关系数分别为0.993 1，0.894 0和0.940 6，经分析比较，发现MTLS-SVR模型优于PLS，LS-SVR以及多因变量LS-SVR(MLS-SVR)，从而验证了MTLS-SVR模型的可行性和有效性。

关键词 近红外光谱；化学计量学；多任务LS-SVR

中图分类号：O657.3 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2011)05-00-04

引 言

近红外光谱技术[1]具有操作简单、分析速度快以及测定一次光谱可同时获得样品多种成分含量的独特优点，使其在作物品质分析上得到了广泛应用。目前近红外光谱定量分析采用的化学计量学建模方法比较多，有些已取得了不错的效果，比如偏最小二乘(PLS)[2]、支持向量回归机(SVR)[3]等。与其他方法相比，SVR具有出色的学习及推广能力，但它需要求解一个二次规划问题，非常耗时。而最小二乘SVR(LS-SVR)[4]用等式约束代替不等式约束，只需求解一个线性方程组，大大加快了求解速度，受到了人们越来越多地重视。

近红外光谱包含了样品中所有成分的光谱信息，但目前许多模型分开考虑样品成分含量，失去了样品成分间潜在的联系。近年来，机器学习领域中多任务学习(MTL)研究逐渐成熟。如果将建模分析每种样品成分含量的问题看作一个任务，则可将样品成分含量同时建模分析的问题转换为多任务学习问题，从真正意义上实现了样品多成分含量的同时建模分析。目前大部分MTL模型[5,6]是基于Bayesian的，但Evgeniou等[7]借助Bayesian MTL的思想，设计了一种正则化MTL模型，不过该模型仍需求解一个二次规划问题。本文在Evgeniou等[7]工作的基础上，提出了多任务LS-SVR(MTLS-SVR)模型。类似于LS-SVR，该模型也只需求解一个线性方程组，而且本文也给出了一种有效的大规模问题求解算法。最后建模分析了高粱样品的三种成分(蛋白质，赖氨酸及淀粉)含量，平均相对误差和相关系数指标均显著优于PLS，LS-SVR和MLS-SVR。

1 最小二乘支持向量回归机(LS-SVR)

对于 $\forall n \in \mathbf{N}$，记$\mathbf{N}_n = \{1, 2, \cdots, n\}$。给定训练数据集$\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$，其中 $\boldsymbol{x}_i \in \mathbf{R}^d$，$y_i \in \mathbf{R}$，记 $\boldsymbol{y} = (y_1; y_2; \cdots; y_N)$。则LS-SVR[4]的原始问题可表述为

$$\min_{\boldsymbol{w}\in\mathbf{R}^{n_h}, b\in\mathbf{R}, \boldsymbol{\xi}\in\mathbf{R}^N} J(\boldsymbol{w}, \boldsymbol{\xi}) = \frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} + \gamma\frac{1}{2}\boldsymbol{\xi}^{\mathrm{T}}\boldsymbol{\xi} \quad (1)$$

$$\text{s.t. } y_i = \boldsymbol{w}^{\mathrm{T}}\varphi(\boldsymbol{x}_i) + b + \xi_i,\ i \in \mathbf{N}_N \quad (2)$$

其中 φ：$\mathbf{R}^d \to \mathbf{R}^{n_h}$ 为输入空间到某一高维(可能为无穷维)特征空间 H 的映射，n_h 表示特征空间 H 的维度，$\boldsymbol{\xi} = (\xi_1; \xi_2;$

收稿日期：2010-10-24，**修订日期**：2011-03-09

基金项目："十一五"国家科技支撑计划(2006BAH03B03)，中国科学技术信息研究所重点工作项目(2009KP01-3-2)和中央高校基本科研业务费专项资金(2009-2-05)资助

作者简介：徐 硕，1979年生，中国科学技术信息研究所博士后 e-mail：xush@istic.ac.cn

*通讯联系人 e-mail：zhangld@cau.edu.cn

…；ξ_N)为由松弛变量组成的向量。

通过 Lagrange 函数，问题(1)～(2)的求解可转换为以下线性方程组的求解

$$\begin{bmatrix}0 & \boldsymbol{e}^{\mathrm{T}}\\ \boldsymbol{e} & \boldsymbol{K}+\gamma^{-1}\boldsymbol{I}\end{bmatrix}\begin{bmatrix}b\\ \boldsymbol{\alpha}\end{bmatrix}=\begin{bmatrix}0\\ \boldsymbol{y}\end{bmatrix} \tag{3}$$

其中 $\boldsymbol{\alpha}=(\alpha_1;\alpha_2;\cdots;\alpha_N)$为 Lagrange 乘子向量，$\boldsymbol{e}=(1;1;\cdots;1)$，$\boldsymbol{I}$ 为单位矩阵，$K_{i,j}=K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\varphi(\boldsymbol{x}_i)^{\mathrm{T}}\varphi(\boldsymbol{x}_j)(i,j\in\mathbf{N}_N)$，$K(\cdot,\cdot)$为满足 Mercer 定理[3]的核函数。

记线性方程组(3)的解为 $\boldsymbol{\alpha}^*=(\alpha_1^*;\alpha_2^*;\cdots;\alpha^*N)$，$b^*$，则决策函数为

$$\begin{aligned}f(\boldsymbol{x})&=\boldsymbol{w}^{*\mathrm{T}}\varphi(\boldsymbol{x})+b^*=\sum_{i=1}^{N}\alpha_i^*\varphi(\boldsymbol{x}_i)^{\mathrm{T}}\varphi(\boldsymbol{x})+b^*\\&=\sum_{i=1}^{N}\alpha_i^*K(\boldsymbol{x}_i,\boldsymbol{x})+b^*\end{aligned} \tag{4}$$

2 多任务 LS-SVR(MTLS-SVR)

假设有 $M(\geqslant 1)$ 个任务，对于任务 $m\in\mathbf{N}_M$，给定 N_m 个训练数据$\{\boldsymbol{x}_{m,i},y_{m,i}\}_{i=1}^{N_m}$，其中 $\boldsymbol{x}_{m,i}\in\mathbf{R}^d$，$y_{m,i}\in\mathbf{R}$，这样就有 $N=\Sigma_m N_m$ 个训练数据。为方便，记 $\boldsymbol{y}=(y_1;y_2;\cdots;y_M)$，其中 $\boldsymbol{y}_m=(y_{m,1};y_{m,2};\cdots;y_{m,N_m})$。对于近红外光谱技术，所有样品成分对应的光谱数据是一样的，即对所有任务，输入向量 $\boldsymbol{x}_{m,i}\in\mathbf{R}^d(i\in\mathbf{N}_{N_m})$ 是相同的，但本文提出的方法不限于此，即它具有更广的适用性。

2.1 原始问题及推导过程

对于任务 $m\in\mathbf{N}_M$，记需要建模的回归函数为 $f_m(\boldsymbol{x})=\boldsymbol{w}_m^{\mathrm{T}}\varphi(\boldsymbol{x})+b_m$。为引入样品成分间潜在的联系，假设所有权重向量 $\boldsymbol{w}_m$ 在某个均值向量 $\boldsymbol{w}_0$ 周围波动，波动的幅度用向量 $\boldsymbol{v}_m$ 来描述，即 $\boldsymbol{w}_m=\boldsymbol{w}_0+\boldsymbol{v}_m$。如果任务间联系比较强，$\boldsymbol{v}_m\to 0$，否则 $\boldsymbol{w}_0\to 0$。类似于正则化 MTL[7]，MTLS-SVR 的原始问题可表述为

$$\begin{aligned}\min_{\boldsymbol{w}_0\in\mathbf{R}^{n_h},\boldsymbol{v}_m\in\mathbf{R}^{n_h},b\in\mathbf{R}^M,\boldsymbol{\xi}=\mathbf{R}^N} J(\boldsymbol{w}_0,\boldsymbol{v}_m,\boldsymbol{\xi})&=\frac{1}{2}\boldsymbol{w}_0^{\mathrm{T}}\boldsymbol{w}_0+\\ \frac{1}{2}\frac{\lambda}{M}\sum_{m=1}^{M}\boldsymbol{v}_m^{\mathrm{T}}\boldsymbol{v}_m+\gamma\frac{1}{2}\sum_{m=1}^{M}\boldsymbol{\xi}_m^{\mathrm{T}}\boldsymbol{\xi}_m&\end{aligned} \tag{5}$$

$$\text{s.t. } y_{m,i}=(\boldsymbol{w}_0,\boldsymbol{v}_m)^{\mathrm{T}}\varphi(\boldsymbol{x}_{m,i})+b_m+\xi_{m,i},m\in\mathbf{N}_M,\ i\in\mathbf{N}_{N_m} \tag{6}$$

其中 $\boldsymbol{b}=(b_1;b_2;\cdots;b_M)$为偏置向量，$\boldsymbol{\xi}=(\xi_1;\xi_2;\cdots;\xi_M)$，$\boldsymbol{\xi}_m=(\xi_{m,1};\xi_{m,2};\cdots;\xi_{m,N_m})(m\in\mathbf{N}_M)$为由松弛变量组成的向量。

问题(5)～(6)的 Lagrange 函数为

$$\begin{aligned}L(\boldsymbol{w}_0,\boldsymbol{v}_m,\boldsymbol{b},\boldsymbol{\xi},\boldsymbol{\alpha})=J(\boldsymbol{w}_0,\boldsymbol{v}_m,\boldsymbol{\xi})-\sum_{m=1}^{M}\sum_{i=1}^{N_m}\alpha_{m,i}\times\\ \{(\boldsymbol{w}_0+\boldsymbol{v}_m)^{\mathrm{T}}\varphi(\boldsymbol{x}_{m,i})+b_m+\xi_{m,i}-y_{m,i}\}\end{aligned} \tag{7}$$

其中 $\boldsymbol{\alpha}=(\boldsymbol{\alpha}_1;\boldsymbol{\alpha}_2;\cdots;\boldsymbol{\alpha}_M)$，$\boldsymbol{\alpha}_m=(\alpha_{m,1};\alpha_{m,2};\cdots;\alpha_{m,N_m})(m\in\mathbf{N}_M)$表示 Lagrange 乘子向量，则对应的 KKT 条件如下

$$\frac{\partial L}{\partial\boldsymbol{w}_0}=0\Rightarrow\boldsymbol{w}_0=\sum_{m=1}^{M}\sum_{i=1}^{N_m}\alpha_{m,i}\varphi(\boldsymbol{x}_{m,i})$$

$$\frac{\partial L}{\partial\boldsymbol{v}_m}=0\Rightarrow\boldsymbol{v}_m=\frac{M}{\lambda}\sum_{i=1}^{N_m}\alpha_{m,i}\varphi(\boldsymbol{x}_{m,i})$$

$$\frac{\partial L}{\partial b_m}=0\Rightarrow\sum_{i=1}^{N_m}\alpha_{m,i}=0,\quad m\in\mathbf{N}_M$$

$$\frac{\partial L}{\partial\xi_{m,i}}=0\Rightarrow\alpha_{m,i}=\gamma\xi_{m,i},\quad m\in\mathbf{N}_M,i\in\mathbf{N}_{N_m}$$

$$\begin{aligned}\frac{\partial L}{\partial\alpha_{m,i}}=0\Rightarrow(\boldsymbol{w}_0+\boldsymbol{v}_m)^{\mathrm{T}}\varphi(\boldsymbol{x}_{m,i})+b_m+\\ \xi_{m,i}-y_{m,i}=0,\quad m\in\mathbf{N}_M,i\in\mathbf{N}_{N_m}\end{aligned} \tag{8}$$

类似于 LS-SVR[4]，式(8)可表示为线性方程组

$$\begin{bmatrix}0 & \boldsymbol{A}^{\mathrm{T}}\\ \boldsymbol{A} & \boldsymbol{\Omega}+\gamma^{-1}\boldsymbol{I}_N+(M/\lambda)\boldsymbol{B}\end{bmatrix}\begin{bmatrix}\boldsymbol{b}\\ \boldsymbol{\alpha}\end{bmatrix}=\begin{bmatrix}0\\ \boldsymbol{y}\end{bmatrix} \tag{9}$$

其中 $\boldsymbol{I}_N$ 为 N 阶单位矩阵；$\boldsymbol{A}=\mathrm{blockdiag}\{\boldsymbol{e}_{N_1},\boldsymbol{e}_{N_2},\cdots,\boldsymbol{e}_{N_M}\}$，$\boldsymbol{e}_{N_m}(m\in\mathbf{N}_M)$为包含 N_m 个元素的全 1 列向量；$\boldsymbol{B}=\mathrm{blockdiag}\{\boldsymbol{\Omega}_1,\boldsymbol{\Omega}_2,\cdots,\boldsymbol{\Omega}_M\}$，$\boldsymbol{\Omega}_m=\boldsymbol{Z}_m^{\mathrm{T}}\boldsymbol{Z}_m$，$\boldsymbol{Z}_m=(\Phi(\boldsymbol{x}_{m,1}),\varphi(\boldsymbol{x}_{m,2}),\cdots,\varphi(\boldsymbol{x}_{m,N_m}))(m\in\mathbf{N}_M)$；$\boldsymbol{\Omega}=\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z}$，$\boldsymbol{Z}=(\boldsymbol{Z}_1,\boldsymbol{Z}_2,\cdots,\boldsymbol{Z}_M)$。

记线性方程组(9)的解为 $\boldsymbol{\alpha}^*=(\boldsymbol{\alpha}_1^*;\boldsymbol{\alpha}_2^*;\cdots;\boldsymbol{\alpha}_M^*)$，$\boldsymbol{\alpha}_m=(\alpha_{m,1}^*;\alpha_{m,2}^*;\cdots;\alpha_{m,N_m}^*)(m\in\mathbf{N}_M)$，$\boldsymbol{b}^*=(b_1^*;b_2^*;\cdots;b_M^*)$，则任务 $m\in\mathbf{N}_M$ 的决策函数为

$$\begin{aligned}f_m(\boldsymbol{x})&=(\boldsymbol{w}_0^*+\boldsymbol{v}_0^*)^{\mathrm{T}}\varphi(\boldsymbol{x})+b_m^*=\sum_{m'=1}^{M}\sum_{i=1}^{N_{m'}}\alpha_{m',i}^*\varphi(\boldsymbol{x}_{m',i})^{\mathrm{T}}\varphi(\boldsymbol{x})+\\ &\frac{M}{\lambda}\sum_{i=1}^{N_m}\alpha_{m,i}^*\varphi(\boldsymbol{x}_{m,i})^{\mathrm{T}}\varphi(\boldsymbol{x})+b_m^*=\sum_{m'=1}^{M}\sum_{i=1}^{N_{m'}}\alpha_{m',i}^*K(\boldsymbol{x}_{m',i},\boldsymbol{x})\\ &+\frac{M}{\lambda}\sum_{i=1}^{N_m}\alpha_{m,i}^*K(\boldsymbol{x}_{m,i},\boldsymbol{x})+b_m^*\end{aligned} \tag{10}$$

2.2 大规模问题求解算法

线性方程组(9)包含 $N+M$ 个方程，而且涉及到的矩阵通常为稠密型的，目前有许多求解线性方程组的成熟方法[8-10]。研究表明[11]，HS(hestenes-stiefel)共轭梯度法在求解大规模线性方程组具有一定优势，但该方法要求 $\boldsymbol{Ax}=\boldsymbol{B}$ 中的矩阵 $\boldsymbol{A}$ 必须是正定对称阵，但是式(9)中的矩阵 $\boldsymbol{A}$ 是对称阵却不是正定阵。为利用该方法，类似于 Suykens 等[4]，可等价变换式(9)。为方便，记式(9)为

$$\begin{bmatrix}0 & \boldsymbol{A}^{\mathrm{T}}\\ \boldsymbol{A} & \boldsymbol{H}\end{bmatrix}\begin{bmatrix}\boldsymbol{\zeta}_1\\ \boldsymbol{\zeta}_2\end{bmatrix}=\begin{bmatrix}\boldsymbol{d}_1\\ \boldsymbol{d}_2\end{bmatrix} \tag{11}$$

其中 $\boldsymbol{H}=\boldsymbol{\Omega}+\gamma^{-1}\boldsymbol{I}_N+(M/\lambda)\boldsymbol{B}$，$\boldsymbol{\zeta}_1=\boldsymbol{b}$，$\boldsymbol{\zeta}_2=\boldsymbol{\alpha}$，$\boldsymbol{d}_1=0$，$\boldsymbol{d}_2=\boldsymbol{y}$。因为矩阵 $\boldsymbol{H}$ 是正定对称阵，故式(11)可等价变换为

$$\begin{bmatrix}\boldsymbol{S} & 0\\ 0 & \boldsymbol{H}\end{bmatrix}=\begin{bmatrix}\boldsymbol{\zeta}_1\\ \boldsymbol{H}^{-1}\boldsymbol{A}\boldsymbol{\zeta}_1+\boldsymbol{\zeta}_2\end{bmatrix}=\begin{bmatrix}-\boldsymbol{d}_1+\boldsymbol{A}^{\mathrm{T}}\boldsymbol{H}^{-1}\boldsymbol{d}_2\\ \boldsymbol{d}_2\end{bmatrix} \tag{12}$$

其中 $\boldsymbol{S}=\boldsymbol{A}^{\mathrm{T}}\boldsymbol{H}^{-1}\boldsymbol{A}$，而且容易证明矩阵 $\boldsymbol{S}$ 是可逆的。从而可得算法如下：

(1) 利用 HS 共轭梯度法求解 $\boldsymbol{H\eta}=\boldsymbol{A}$ 和 $\boldsymbol{Hv}=\boldsymbol{d}_2$，并记最优解分别为 $\boldsymbol{\eta}^*$ 和 $\boldsymbol{v}^*$；

(2) 计算 $\boldsymbol{S}=\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\eta}^*$；

(3) 线性方程组(9)的解为：$\boldsymbol{b}=\boldsymbol{\zeta}_1=\boldsymbol{S}^{-1}\boldsymbol{\eta}^{*\mathrm{T}}\boldsymbol{d}_2$，$\boldsymbol{\alpha}=\boldsymbol{\zeta}_2=\boldsymbol{v}^*-\boldsymbol{\eta}^*\boldsymbol{\zeta}_1$。

3 实验方法及实验数据

128 个高粱样品由中国农业科学研究院品种资源所提

供，样品用带有 1.0 mm 空径筛网的粉碎机磨碎，其在 19 个波长点处的近红外漫反射光谱由 LA450 近红外光谱仪测定。该仪器由 BRAN+LUEBBE(德国)公司生产，具有从 1 445～2 348 nm 之间的 19 个分隔的滤光片。并对每个样品分别测定了蛋白质、赖氨酸以及淀粉三种成分的含量，其中蛋白质含量采用国家标准 GB290521982 半微量凯氏定氮法测定；赖氨酸含量的测定采用国家标准 GB480121984 谷物籽粒赖氨酸测定(染料法)；淀粉含量采用国家标准 BG500621985 谷物籽粒淀粉测定法测定。这样，对应于三种样品成分含量，需要同时建模三个任务。

4　实验结果及分析

本文从 128 个高粱样品中随机选取 96 个样品组成训练集，余下的 32 个样品构成测试集。为说明 MTLS-SVR 模型的优势，以 LS-SVR[4]、多因变量 LS-SVR(MLS-SVR)[12]和 PLS[2]三种模型做参照，平均相对误差 δ 和相关系数 R 作为评价指标。核函数采用径向基(RBF)核：$K(\boldsymbol{x},\boldsymbol{z})=\exp(-p\|\boldsymbol{x}-\boldsymbol{z}\|^2)$，$p>0$。

4.1　参数选择

参数(γ，p，λ)对 LS-SVR，MLS-SVR 和 MTLS-SVR 模型的性能影响比较大。为了选择一组合适的参数，本文采用两步格网搜索选参法[13,14]：第一步粗格网搜索，参数 γ，p 和 λ 按对数增长的方式分别从集合$\{2^{-5},2^{-3},\cdots,2^{15}\}$，$\{2^{-15},2^{-13},\cdots,2^{3}\}$及$\{2^{-10},2^{-8},\cdots,2^{10}\}$中取值，根据总体平均相对误差最小的原则采用留一法确定合适的参数 γ^*，p^* 和 λ^*。如果某个参数处于搜索空间的边界，将相应空间按相同的乘积步长($2^{\pm 2}$)增加一个元素，直到没有进一步的性能改进为止；第二步细格网搜索，首先对 $c\in\{\gamma,p,\lambda\}$构造新的搜索空间：$\{2^{-1.75}\times c^*,2^{-1.5}\times c^*,\cdots,2^{1.75}\times c^*\}$，然后重复上一步，将该步确定的值作为最优参数值，见表 1。对于 MLS-SVR 还需确定一个最优权重向量，本文采用了安欣等[12]给出的方法。对于 PLS 模型，本文将输入矩阵的行秩与列秩的最小值作为主成分的个数。

Table 1　Optimal parameters for LS-SVR, MLS-SVR and MTLS-SVR on the broomcorn data set

	LS-SVR		MLS-SVR		MTLS-SVR		
	γ	p	γ	p	γ	p	λ
蛋白质	$2^{27.5}$	$2^{-16.25}$	$2^{26.25}$	$2^{-16.25}$			
赖氨酸	27.75	2^{-1}	$2^{17.5}$	$2^{-9.5}$	2^{20}	$2^{-9.75}$	2^{-1}
淀粉	$2^{17.25}$	$2^{-7.25}$	$2^{16.5}$	$2^{-7.25}$			

4.2　结果分析

根据表 1 确定的最优参数，分别利用 PLS，LS-SVR，MLS-SVR 和 MTLS-SVR 建立定量分析模型，确定高粱样品各成分含量的预测值并计算相应的评价指标，见表 2。容易看出，MTLS-SVR 的效果最好，MLS-SVR 与 LS-SVR 的性能相当，PLS 的性能最差。本文认为原因有三个方面：(1) MTLS-SVR 同时考虑了各个样品成分的特性($\boldsymbol{v}_m$)和共性($\boldsymbol{w}_0$)，而 LS-SVR 只建模了各个样品成分的特性；(2)MLS-SVR 尽管为各个样品成分设置了不同的权重，但对偶问题仍需分开求解，因而也只建模了样品成分的特性，而且由于样品成分权重的引入，使得参数优化变得更加困难；(3)本文采用的 PLS 可看作原始空间中的一种线性模型[15]，而近红外光谱数据与样品成分含量间有时存在非线性关系。

Table 2　Comparisons of the predicted results with PLS, LS-SVR, MLS-SVR and MTLS-SVR quantitative analysis models

	δ/%				R			
	PLS	LS-SVR	MLS-SVR	MTLS-SVR	PLS	LS-SVR	MLS-SVR	MTLS-SVR
蛋白质	1.77	1.52	1.52	1.52	0.991 7	0.992 2	0.992 3	0.993 1
赖氨酸	5.75	4.98	4.91	3.04	0.731 7	0.762 7	0.772 5	0.894 0
淀粉	1.36	1.11	1.11	1.01	0.884 2	0.922 3	0.922 3	0.940 6
Avg.	2.96	2.54	2.51	1.85	0.869 2	0.892 4	0.895 7	0.942 6

5　结论及讨论

受正则化 MTL 的启发，本文在 LS-SVR 的基础上设计了一种新的 MTL 模型—MTLS-SVR。将同时建模分析多种样品成分含量的问题转换为 MTL 的问题，从而可充分利用各样品成分含量间潜在的联系，从真正意义上实现了样品多成分含量的同时建模分析。针对样品成分较多的情形，给出了一种大规模问题求解算法。以高粱样品数据集为实验材料，通过与 PLS、LS-SVR 以及 MLS-SVR 的对比分析，验证了 MTLS-SVR 的有效性和高效性，同时为发展化学计量学提供了一种新的定量分析建模方法。

References

[1] YAN Yan-lu, ZHAO Long-lian, HAN Dong-hai, et al(严衍禄，赵龙莲，韩东海，等). Foundation of Near Infrared Spectral Analysis and its Applications(近红外光谱分析基础与应用). Beijing: China Light Industry Press(北京：中国轻工业出版社), 2005.

[2] Abdi H. Partial Least Squares (PLS) Regression. Encyclopedia for Research Methods for the Social Sciences, Lewis-Beck M, Bryman A,

Futing T, eds. Sage, Thousand Oaks, CA, pp. 792.
[3] Vapnik V N. The Nature of Statistical Learning Theory, 2nd Edition. New York: Springer Verlag, 1999.
[4] Suykens J A K, Gestel T V, Brabanter J D, et al. Least Squares Support Vector Machines. Singapore: World Scientific Pub. Co., 2002.
[5] Bakker B, Heskes T. Journal of Machine Learning Research, 2003, 4(May): 83.
[6] Heskes T. Empirical Bayes for Learning to Learn. Proceedings of the 17th International Conference on Machine Learning (ICML), San Francisco, CA, USA, 2000. 367.
[7] Evgeniou T, Pontil M. Regularized Multi-Task Learning. Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), Seattle, WA, USA, 2004. 109.
[8] Golub G H, Van Loan C F. Matrix Computations, 3rd Edition. Baltimore and London: Johns Hopkins University Press, 1996.
[9] Press W H, Teukolsky S A, Vetterling W T, et al. Numerical Recipes in C: The Art of Scientific Computing. New York: Cambridge University Press, 1992.
[10] Saad Y. Iterative Methods for Sparse Linear Systems, 2nd Edition. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, USA, 2003.
[11] Hamers B, Suykens J A K, Moor B D. A Comparison of Iterative Methods for Least Squares Vector Machine Classifiers. Internal Report 01-110, 2001. ESAT-SISTA, K. U. Leuven, Leuven, Belgium.
[12] AN Xin, XU Shuo, ZHANG Lu-da, et al(安 欣, 徐 硕, 张录达, 等). Spectroscopy and Spectral Analysis (光谱学与光谱分析), 2009, 29(1): 127.
[13] Hsu C-W, Chang C-C, Lin C-J. A Practical Guide to Support Vector Classification. Available [online]: http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf.
[14] Xu S, Ma F J, Tao L. Learn from the Information Contained in the False Splice Sites as well as in the True Splice Sites using SVM. Proceedings of the International Conference on Intelligent Systems and Knowledge Engineering (ISKE), Chengdu, China, 2007. 1360.
[15] Rosipal R, Trejo L J. Journal of Machine Learning Research, 2001, 2(Dec): 97.

Multi-Task Least-Squares Support Vector Regression Machines with Applications in NIR Spectral Analysis

XU Shuo[1], QIAO Xiao-dong[1], ZHU Li-jun[1], AN Xin[2], ZHANG Lu-da[3]*
1. Information Technology Supporting Center, Institute of Scientific and Information of China, Beijing 100038, China
2. School of International Trade and Economics, University of International Business and Economics, Beijing 100029, China
3. College of Science, China Agricultural University, Beijing 100193, China

Abstract In near infrared spectral quantitative analysis, many models consider separately each composition when modeling sample compositions' content, this disregarding the underlying relatedness among sample compositions. To address this problem, the paper views modeling each sample composition's content as a task, thus one can transform the problem that models simultaneously all sample compositions' content to a multi-task learning problem. On the basis of the LS-SVR, a multi-task LS-SVR (MTLS-SVR) model is proposed. Furthermore, an efficient large-scale algorithm is given. The broomcorn samples are taken as experimental material, and corresponding quantitative analysis models are constructed for three sample compositions' content (protein, lysine and starch) with LS-SVR, PLS, multiple dependent variables LS-SVR (MLS-SVR) and MTLS-SVR. For the MTLS-SVR model, the average relative errors between actual values and predicted ones for the three sample compositions' content are 1.52%, 3.04% and 1.01%, respectively, and the correlation coefficients are 0.993 1, 0.894 0 and 0.940 6, respectively. Experimental results show MTLS-SVR model outperforms significantly the three others, which verifies the feasibility and efficiency of the MTLS-SVR model.

Keywords Near infrared spectrum; Chemometrics; Multi-task LS-SVR

(Received Oct. 24, 2010; accepted Mar. 9, 2011)

* Corresponding author