

文章编号:1000-2375(2017)05-0546-04

基于部分线性回归的红外光谱多元校正方法

郭露, 彭江涛, 付辉敬

(湖北大学数学与统计学学院, 湖北 武汉 430062)

摘要:对于红外光谱数据而言,光谱-浓度关系常表现为一种复杂的混合线性关系. 本文中提出一种部分线性回归算法,将复杂的光谱-浓度目标回归函数分解为线性和非线性决策函数之和. 具体地,采用一序列的线性和非线性核函数来构建回归模型,分别用于逼近目标函数中的线性和非线性成分. 本文中所提出的方法与偏最小二乘回归算法和正则化最小二乘回归算法在 3 个实例数据集上进行实验对比. 实验结果表明,本文中提出的算法具有更高的预测精度.

关键词:部分线性回归; 红外光谱; 多元校正

中图分类号:X36 文献标志码:A DOI:10.3969/j.issn.1000-2375.2017.05.020

Partially linear regression for multivariate calibration of spectroscopic data

GUO Lu, PENG Jiangtao, FU Huijing

(Faculty of Mathematics and Statistics, Hubei University, Wuhan 430062, China)

Abstract: Spectra-concentrate relation is usually a very complex and mixed linear relation. In this paper, a partially linear regression (PLR) algorithm is proposed for multivariate calibration of spectroscopic data. In PLR, the target regression function is represented as the sum of several linear and nonlinear kernel decision functions, where each single kernel function with specific type and scale can approximate certain component of the target function. The proposed method is compared, in terms of RMSEP, with partial least squares regression (PLS) and regularized least-squares regression (RLS) method on three real spectroscopic data sets. Experimental results demonstrate that the proposed PLR method shows superiority over PLS and the single kernel RLS.

Key words: partially linear regression; infrared spectroscopy; multivariate calibration

多元校正是化学计量学中的一个非常有用的工具. 多元校正能够在光谱和对应的浓度之间建立一个回归模型,揭示物质成分之间的定量关系. 传统的多元校正通常假定回归模型是呈线性关系的,例如多元线性回归 (MLR)、主成分回归 (PCR) 以及偏最小二乘回归 (PLS)^[1-3]. 在这些方法中,PLS 在化学计量学中的应用最为广泛.

PLS 将高维预测变量投射到低维的、不相关的潜在变量集合中,并要求潜在变量与响应之间有最大的协方差. 当变量数量远超过样本数量或者数据中存在共线性预测变量时,PLS 是非常有效的方法^[1-3]. 然而,当数据表现出很强的非线性特征时,传统的线性 PLS 方法不能完全描述光谱与相应的浓度之间的关系,因而会产生较大的误差.

为了更好地描述光谱和浓度之间的非线性关系,正则化最小二乘回归算法 (RLS)^[4] 用核函数来表示决策函数. 由于核函数可完全由训练集中的输入样本决定,选择一个合适的非线性核 (例如高斯径向基核),RLS 就能很好地实现非线性回归. 但是,单核 RLS 的能力是非常有限的,对于复杂的非线性光谱数据,单核 RLS 并不适用. 如果回归函数由多种不同成分组成,例如,既包含线性成分又含有非线性成分,既包含平坦成分又包含陡变成分,此时 RLS 会造成过拟合或者欠拟合现象. 因此,采用多种不同类型的核函数组合比单核更加有效,线性核和非线性核分别能够处理目标函数中的线性部分和非线性部分.

收稿日期:2017-06-01

基金项目:湖北省教育厅中青年人才项目(Q20161003)资助

作者简介:郭露(1992-),女,硕士生;付辉敬,通信作者,讲师,E-mail: fxy0204@126.com

在本文中,我们提出一种部分线性回归算法(PLR),用于多元校正.在PLR中,目标回归函数表示为线性和非线性核决策函数的和,每个核函数能够逼近目标函数中的不同成分.

1 算法

学习理论中回归问题的目的是从样本中学习得到回归函数或者得到其好的逼近.在最小二乘回归问题中,寻找回归函数的最小二乘正则化算法是与 Mercer 核 K 相联系的.设 $K: X \times X \rightarrow R$ 是一连续、对称且正定的函数,称为 Mercer 核^[5].由核 K 生成的再生核希尔伯特空间 H_K 定义为由函数集 $\{K_x: = K(x, \cdot): x \in X\}$ 所张成的闭包,其中内积 $\langle \cdot, \cdot \rangle_{\mathcal{H}_K} = \langle \cdot, \cdot \rangle_K$ 定义为 $\langle K_x, K_{x'} \rangle_K = K(x, x')$,再生性表现为

$$\langle K_x, K_{x'} \rangle_K = K(x, x') \tag{1}$$

与 Mercer 核 K 相联系的回归问题的最小二乘正则化算法定义为:根据一个训练样本集 $z = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$,寻找与 z 相关联的最小二乘优化问题的最小化函数:

$$\min_{f \in \mathcal{H}_K} \left\{ \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \|f\|_K^2 \right\} \tag{2}$$

其中, $\lambda \geq 0$ 是正则项参数.根据表示理论^[6],问题(2)的解可表示为:

$$f(x) = \sum_{i=1}^n \alpha_i K(x, x_i) \tag{3}$$

同时, $\alpha = (\alpha_1, \dots, \alpha_n)^T$ 也是适定线性问题(4)的唯一解.

$$(n\lambda I + K[x])\alpha = y \tag{4}$$

在(4)式中, $K[x]$ 是 $n \times n$ 的矩阵,第 (i, j) 个元素为 $K(x_i, x_j)$,以及 $y = (y_1, y_2, \dots, y_n)^T$.问题(2)的正则项满足

$$\lambda \|f\|_K^2 = \lambda \sum_{i,j=1}^n \alpha_i \alpha_j K(x_i, x_j) \tag{5}$$

从表示理论[6]可以看出,RLS 算法通过对 n 个基函数 $K(x, x_i)$ 的加权求和去逼近回归函数,这些基函数有着相同的结构.在由高斯径向基核 K_{RBF} 生成的 RKHS 中,所有基函数都有相同的结构.例如 $K_{\text{RBF}}(x, x_i) = \exp(-\|x - x_i\|_2^2 / \sigma^2)$,核的宽度 σ 是相同的,只有每个函数的中心 x_i 不同.但是,当回归函数包含比较复杂的成分时,用单核生成的 RKHS 去逼近目标函数是远不够的,一个自然的推广方法就是采用多核.因此,我们提出一种基于多核的部分线性回归方法(PLR).

在 PLR 中,目标回归函数表示为多个单核决策函数的和,每个核函数都有特定的类型和大小,能够逼近目标函数中的特定成分.线性核函数和非线性核函数分别逼近目标函数中的线性部分和非线性部分,尺寸不同的核逼近目标函数中不同频率偏差的部分.假定 K_1, K_2, \dots, K_l 是 l 个尺寸、类型(包括线性核和非线性核)不全相同的 Mercer 核. \mathcal{H}_{\oplus} 是由这 l 个核生成的 RKHS 和空间, \mathcal{H}_{\oplus} 中的函数可表示为

$f = \sum_{i=1}^l f_i, f_i \in \mathcal{H}_{K_i}$,从而得到相应的优化问题:

$$\min_{f_i \in \mathcal{H}_{K_i}} \left\{ \frac{1}{n} \sum_{i=1}^n (y_i - \sum_{t=1}^l f_t(x_i))^2 + \sum_{t=1}^l \lambda_t \|f_t\|_{K_t}^2 \right\} \tag{6}$$

由表示定理,问题(6)的解可表示为: $f_t(x) = \sum_{i=1}^n \alpha_i^t K_t(x, x_i)$ (7)

其中, $t = 1, 2, \dots, l (l \geq 2)$ 是核函数的数量, $\alpha^t = (\alpha_1^t, \dots, \alpha_n^t)^T$ 可由下面的线性问题(8)求解得到:

$$\begin{bmatrix} n\lambda_1 I + K_1[x] & \cdots & K_l[x] \\ \vdots & \cdots & \vdots \\ K_1[x] & \cdots & n\lambda_l I + K_l[x] \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_l \end{bmatrix} = \begin{bmatrix} y \\ \vdots \\ y \end{bmatrix} \tag{8}$$

在(8)式中, $K_t[x]$ 是 $n \times n$ 的矩阵,第 (i, j) 个元素为 $K_t(x_i, x_j)$,以及 $y = (y_1, y_2, \dots, y_n)^T$.

对于光谱回归问题,本文中考虑由 2 个核函数生成的 RKHS 和空间来构建回归模型,具体包括一个线性核函数 $K_L(x, z) = x^T z$, 和一个高斯径向基核函数 $K_R(x, z) = \exp(-\|x - z\|_2^2 / \sigma^2)$, 代入前文提出的 PLR 方法,可以得到最终的解

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} n\lambda_1 I + K_L & K_R \\ K_L & n\lambda_2 I + K_R \end{bmatrix}^{-1} \begin{bmatrix} y \\ y \end{bmatrix} \quad (24)$$

2 实验

2.1 数据集 选取 3 个公共数据集来进行测试分析.

数据集 1(Corn) 包含 80 份玉米样品近红外光谱数据, 由 m5 光谱仪采集测量得到^[7]. 采集波长范围为 1 100-2 498 nm, 采样间隔为 2 nm, 共 700 个通道扫描. 每个样本包含水分、油脂、蛋白质和淀粉这 4 个浓度响应, 浓度的参考范围分别为 9.38 ~ 10.99%、3.09 ~ 3.83%、7.65 ~ 9.71% 和 62.83 ~ 66.47%. 取出 Corn 数据集中的 60 个样本作为训练样本, 其余 20 个样本为测试样本.

数据集 2(Tablet 2002) 包含 655 份药片的近红外光谱数据, 每组样本包括 291 个特征量和 1 个浓度响应^[8]. 数据集分为训练样本(460 个光谱)和测试样本(155 个光谱). 药片的采集波长范围为 1 050 ~ 1 630 nm, 测定的参考范围为 152 ~ 239 mg.

数据集 3(Meat) 是 240 份肉类样本的近红外吸收光谱数据, 采集的波长范围是 850 ~ 1 050 nm^[9]. 每个样本包括 100 通道的吸收光谱, 以及水分、脂肪和蛋白质这 3 个浓度响应. 取出该数据集中的 172 组作为训练样本, 其余 43 组为测试样本.

2.2 算法评价

本文中提出的部分线性回归算法(PLR)将与偏最小二乘回归(PLS)和正则化最小二乘回归算法(RLS)进行比较, 从而验证 PLR 算法的预测能力. 我们采取交叉验证的方法来寻找不同算法的最优参数. 具体地, 对于 PLS, 潜在变量数(主成分数目)的选择需要在最小化预测残量误差平方和(PRESS)与降低模型复杂度之间寻求一个折中. Haaland 和 Thomas 采用 F 检验的方法来检测交叉验证 PRESS 值得变化情况, 并尽量选择简单的模型^[6]. 对于 RLS, 我们采用网格寻优方法, 确定径向基核的大小 σ 和正则项参数 λ , σ 在呈指数变化的范围中选择 $\sigma = [2^{-3}, 2^{-2}, \dots, 2^3]$, λ 的选择范围为 $\lambda = [10^{-10}, 10^{-9}, \dots, 10^0]$. 对于 PLR, 我们需要寻找 3 个最优参数, 分别是径向基核的大小 $\sigma = [2^{-3}, 2^{-2}, \dots, 2^3]$, 非线性正则项参数 $\lambda_1 = [10^{-10}, 10^{-9}, \dots, 10^0]$, 线性正则项参数 $\lambda_2 = [10^{-10}, 10^{-9}, \dots, 10^0]$. 我们采用简单的两步策略来寻找最优参数, 先固定 λ_1 和 σ , 通过最小化验证误差来确定最优 $\lambda_2 = \lambda_2^*$, 然后再寻找最优 $\lambda_1 = \lambda_1^*$ 和最优 $\sigma = \sigma^*$.

对于不同算法, 均采用均方误差根(RMSEP)来衡量其预测性能. RMSEP 衡量测试集样本的预测值与实际值之间的差异程度, 定义为:

$$\text{RMSEP} = \sqrt{\frac{1}{m} \sum_{i=1}^m (y_i - \hat{y}_i)^2} \quad (25)$$

其中, m 是测试集的样本数, y_i 是实际测量的浓度响应, 以及 \hat{y}_i 是从样本数据学习到的预测浓度响应.

3 结果分析

3.1 Corn 数据集

Corn 数据集包含水分、油脂、蛋白质和淀粉这 4 种成分的校正问题. 对训练数据进行数据归一化, 采取交叉验证的方法对各个算法寻找最优参数, 得到的最优参数结果见表(1)中的第一行. 利用最优参数建立回归模型, 代入测试数据得到预测结果见表(2). 结果显示, 对于 Corn 数据集 4 种成分, PLR 方法预测结果的均方误差 RESEP 均低于 PLS 和 RLS. 总体而言, PLR 算法的预测能力更强.

表 1 不同算法的最优参数——不同数据集

数据集	PLS	RLS	PLR
Corn	lr = 12 $\sigma = 4$	$\lambda = 1e-8$ $\sigma' = 4$	$\lambda_1 = 1e-10$ $\lambda_2 = 1e-8$
Tablet 2002	lr = 6 $\sigma = 8$	$\lambda = 1e-4$ $\sigma' = 8$	$\lambda_1 = 1e-5$ $\lambda_2 = 1e-4$
Meat	lr = 11 $\sigma = 4$	$\lambda = 1e-8$ $\sigma' = 4$	$\lambda_1 = 1e-8$ $\lambda_2 = 1e-7$

表 2 不同算法的预测结果——Corn

	PLS	RLS	PLR
水分	0.010 6	0.011 0	0.006 5
油脂	0.056 1	0.074 8	0.027 6
蛋白质	0.102 8	0.126 3	0.094 0
淀粉	0.230 7	0.122 5	0.117 4

3.2 Tablet 2002 数据集

对训练数据进行数据归一化, 采取交叉验证的方法对各个算法寻找最优参数, 得到的最优参数结果见表(1)中的第二行. 利用最优参数建立回归模型, 代入测试数据得到预测结果.

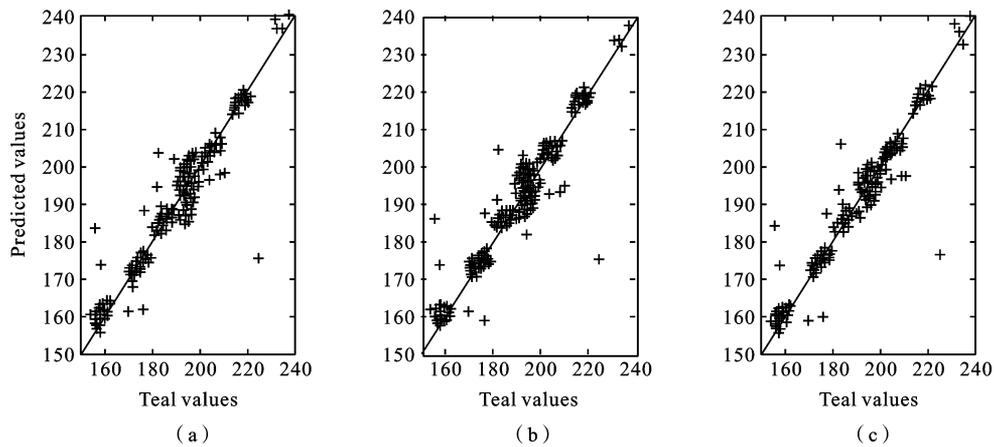


图1 真实值与预测值的对比图

(a) PLS; (b) RLS; (c) PLR

3种方法预测结果的均方误差分别为4.693 7、4.581 3和4.126 5.可以看出,本文中算法可使预测精度得到提高,性能优于PLS和RLS算法.为了更好地看出各个算法的拟合效果,图1中显示各算法的预测值与真实浓度值之间的拟合效果图.从图中可以看出,PLR算法具有更好的拟合精度.

3.3 Meat数据集 同样地,Meat数据集的最优参数结果见表1中的最后一行.利用最优参数建立回归模型,针对测试数据得到预测结果见表3.从结果可以看出,对于水分、脂肪、和蛋白质这3种成分,PLR方法的预测结果均优于PLS和RLS.

表3 不同算法的预测结果——Meat

	PLS	RLS	PLR
水分	2.191 8	1.338 1	0.044 6
脂肪	2.541 3	1.854 4	0.287 6
蛋白质	0.696 6	1.218 3	0.528 9

4 结论

针对复杂光谱数据的多元校正问题,本文中提出一种部分线性回归算法(PLR),其决策函数被表示为多核组合形式.由于多核(多类型核、多尺度核)决策函数具有更强的预测能力,能够逼近光谱回归函数中的不同成分,本文中所提出的PLR算法在3个公共数据集上都展现出了比传统算法(如偏最小二乘回归和正则化最小二乘回归)更优的预测性能.

5 参考文献

- [1] Wold B S, Ruhe A, Wold H, et al. III, The collinearity problem in linear regression: The partial least squares approach to generalized inverses[J]. Siam Journal on Scientific & Statistical Computing, 2013, 5(3):735-743.
- [2] Wold H. Soft modelling by latent variables: the nonlinear iterative partial least squares approach[M]. Perspectives in Probability and Statistics, London:Academic Press:1975, 520-540.
- [3] Haaland D M, Thomas E V, Chem A. Partial Least-Squares Methods for Spectral Analyses. 1. Relation to Other Quantitative Calibration Methods and the Extraction of Quantitative Information[J]. Analytical Chemistry, 1988, 60(11):1193-1202.
- [4] Xu Y L, Chen D R, Li H X, et al. Least square regularized regression in sum space. [J]. IEEE Transactions on Neural Networks & Learning Systems, 2013, 24(4):635-646.
- [5] Aronszajn A. Theory of reproducing kernels. Trans Am Math Soc[J]. Transactions of the American Mathematical Society, 1950, 68(3):337-404.
- [6] Cucker F, Smale S. On the mathematical foundations of learning[J]. Bulletin of the American Mathematical Society, 2001, 39(1):332.

(责任编辑 赵燕)