

NIR光谱的 Isomap-PLS 非线性建模方法

杨辉华^{1,2}, 覃 锋¹, 王义明², 罗国安²

1. 桂林电子科技大学计算机与控制学院, 广西 桂林 541004

2. 清华大学中药现代化研究中心, 北京 100084

摘 要 针对样品的近红外(NIR)光谱与其物理化学性质之间存在的非线性关系,提出了一种结合等距映射(Isomap)和偏最小二乘(PLS)的非线性建模新方法。Isomap是一种新的非线性降维方法,属于流形学习方法,能有效地发现高维数据中的本真低维结构。Isomap-PLS建模方法首先用Isomap对高维NIR光谱数据作非线性降维,再用PLS降维并建立校正模型。将Isomap-PLS建模方法分别应用于两个公开的NIR光谱标准数据集,并与PLS单独建模进行比较。结果表明,在两个数据集上,用Isomap-PLS方法建立的校正模型比单独用PLS算法建立的校正模型具有更小的交叉验证均方根误差(RMSECV);对某些性质数据,Isomap-PLS模型比PLS模型的RMSECV值要小2~5倍。因此,Isomap能够有效反映NIR光谱中存在的非线性结构,Isomap-PLS比PLS具有更好的建模与预测能力。

关键词 近红外光谱;等距映射;非线性降维;偏最小二乘

中图分类号: O657.3 **文献标识码:** A **DOI:** 10.3964/j.issn.1000-0593(2009)02-0322-05

引 言

近红外(near infrared, NIR)光谱数据是由不同波长(或波数)下的吸光度组成,在数学上可以视为一个向量,其吸光度点的个数即为向量维数。NIR光谱仪原理不同或者设置不同,光谱向量的维数将不一样,但通常在数百至2 000之间。一方面光谱测量基本服从 Lambert-Beer 定律导致光谱数据存在严重的多重共线问题,另一方面由于光谱测量中的非单色光、杂散光、温度等多种因素导致光谱偏离该定律,从而使待测理化性质与光谱之间形成非线性关系^[1,2]。同时,光谱数据除包含有用信息外,常包含一定水平的噪声。因此,为了建立一个有预测能力的NIR校正模型,必须使用能从光谱向量中提取有用信息、并准确反映性质与光谱之间线性或非线性关系的化学计量学算法。

降维是一种有效消除噪声并提取有用信息的方法,PCA和PLS等线性降维方法已成为广泛使用的常规方法^[3-6]。针对NIR光谱的非线性,ANN^[7-9]、SVR^[10-12]等非线性回归方法也被引入到NIR光谱处理中。文献^[13]则提出一种集成非线性降维和半监督学习的NIR光谱建模方法——拉普拉斯正则化最小二乘回归。对NIR光谱进行非线性降维,少有报道,仅见利用局部线性嵌入(LL E)降维^[14]。

等距映射(Isomap)^[15]是Tenenbaum等提出的一种新的非线性降维方法,属于流形学习方法。由于Isomap能有效发现高维数据中的低维结构,近年来受到广泛关注,在人脸和手势的实验中获得了很好的效果^[16-18]。

结合Isomap和PLS,本文提出一种NIR光谱建模的新方法——先用Isomap对NIR光谱数据作非线性降维,再用PLS作线性降维并建立校正模型。结果表明,用Isomap-PLS方法建立的校正模型在预测准确性方面均优于PLS。

1 Isomap 算法

多维尺度变换(MDS)^[19]是一种线性降维方法,它构造的欧式距离矩阵不能反映流形样本点之间的非线性关系。为保持数据点的内在几何性质(两点间的测地距离)不变,Isomap算法在MDS的基础上,使用样本点之间的测地距离代替欧式距离。

Isomap算法首先使用最近邻图中的最短路径得到近似的测地线距离,并用该距离代替欧式距离输入到MDS中,进而发现嵌入在高维空间的低维坐标。Isomap算法描述如下。

输入样本 $X = \{x_i\}_{i=1}^n$, $x_i \in R^p$, 样本本真维数 q , 邻域参数 k 。

收稿日期: 2007-08-10, 修订日期: 2007-11-20

基金项目: 广西科学基金项目(桂科青 0542037), 国家自然科学基金项目(30860381)和国家“973”中医理论专项项目(2005CB523503)资助

作者简介: 杨辉华, 1972年生, 桂林电子科技大学计算机与控制学院教授 e-mail: yanghuihua@tsinghua.edu.cn

输出低维嵌入 $T = \{t_i\}_{i=1}^n, t_i \in R^q \quad (q < p)$ 。

算法, 第一步, 构建 k -邻域图 G 。计算每个样本点 x_i 同其余样本点 x_j 之间的欧式距离 d_{ij}^E 。当 x_j 是距最近的 k 个点中的一个时, 则认为 x_i 与 x_j 是相邻的, 即图 G 有边 E_{ij} , 并设边 E_{ij} 的权为 d_{ij}^E 。

第二步, 计算最短路径。当图 G 有边 E_{ij} 时, 设最短路径 $d_{ij}^G = d_{ij}^E$; 否则设 $d_{ij}^G = \infty$ 。在图 G 上, 根据迪杰斯特拉 (Dijkstra) 算法或弗洛伊德 (Floyd) 算法求最短路径距离矩阵 D^G 。

第三步, 计算 q 维嵌入。将 MDS 应用于最短路径矩阵 D^G 。

(1) $S = (s_{ij}) = ((D_{ij}^G)^2)$, $H = (h_{ij}) = (\delta_{ij} - 1/n)$, 其中 $\delta_{ij} = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}$ 。

(2) $\tau^G = -HS/2$, 记 τ^G 的最大 q 个特征值 $\lambda_1, \lambda_2, \dots, \lambda_q$ 对应的特征向量 $u_1, u_2, \dots, u_q$ 所构成的矩阵为 $U = [u_1, u_2, \dots, u_q]$, 则 $T = \text{diag}(\lambda_1^{1/2}, \lambda_2^{1/2}, \dots, \lambda_q^{1/2}) U^T$ 即是 q 维嵌入结果。

2 实 验

2.1 数据集描述

2.1.1 数据集 1

采用文献^[20, 21]的 NIR 数据集。该数据集来自一个乙醇、水和异丙醇的三元混合物, 共 22 个样本。各样本按一定比配制, 使得各组分的百分比在 0%~100% 之间均匀变化。使用 HP 8453 型 UV/Vis 分光光度计, 波长范围 580~1 091 nm, 间隔 1 nm, 维数为 512。每张光谱为 20 次测量光谱的平均值。分别采集各样本在 30, 40, 50, 60, 70 (温度控制精度 ± 0.2 内) 下的光谱。其中, 30 时所有样本的光谱图如图 1 所示。

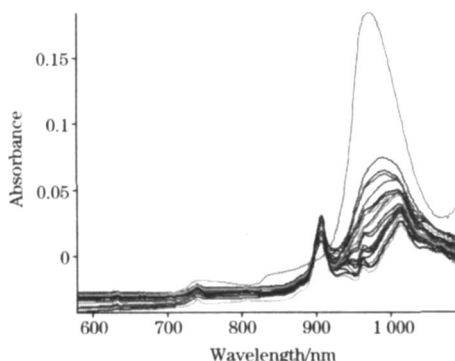


Fig 1 Spectra of 19 mixtures of ethanol, water and isopropanol, the spectra of the pure compounds are recorded in a 1 cm cuvette at 30 degrees Celsius

2.1.2 数据集 2

采用文献^[22]的玉米数据集。该数据集包括 80 个玉米样本, 分别在 m5, mp5, mp6 三台仪器上测试样本的 NIR 光谱, 波长范围 1 100~2 498 nm, 间隔 2 nm, 数据均为 700

维。玉米样本的湿度、油、蛋白质、淀粉四个性质用标准方法测定。其中 m5 仪器测得的光谱图如图 2 所示。

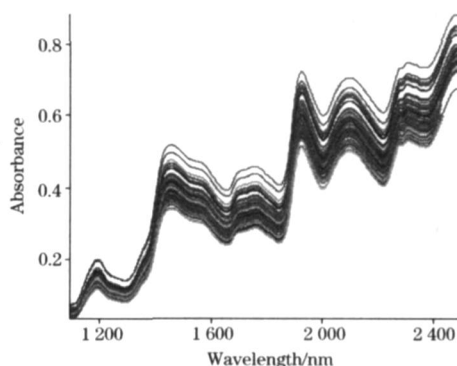


Fig 2 NIR spectra of 80 corn samples measured by NIR instrument m5

2.2 数据处理

将 Isomap-PLS 建模方法应用于 NIR 光谱数据集, 并与 PLS 单独建模型进行比较。在模型评价上以留一法交叉验证均方根误差 (RMSECV) 最小为最优, 其中 RMSE 为

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

其中 $\hat{y}_i$ 为样本的性质的预测值, y_i 为样本的性质的真值, n 为样本个数。

Isomap 算法采用 Wittman 开发的程序包^[23], 其他程序如 Isomap 参数调整算法、PLS 算法等在 MatlabTM 7.1 上自编实现。

3 结果与讨论

3.1 Isomap 参数的影响

Isomap 算法有 2 个可调的参数 k 和 d , 其中 k 为邻域参数, d 为样本本真维数。在不同参数下, Isomap-PLS 得到不同的建模结果。本文采用参数网格搜索法并用留一法验证进行参数优化。

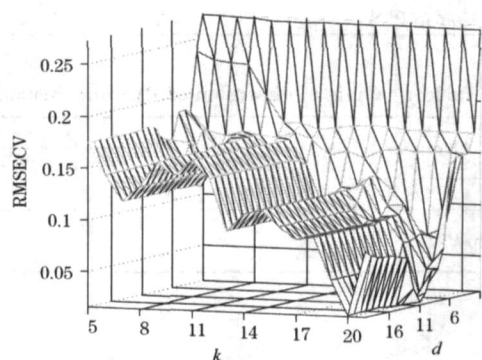


Fig 3 RMSECV varies with parameter k and d

如, 对数据集 1 在 70 下的光谱矩阵, 分别选取 k 从 5 到 20, d 从 1 到 20 作 Isomap 降维, 再用 PLS 降维并回归建立乙醇浓度的校正模型, 并用留一法 RMSECV 来评价模型。

如图 3 所示, 当 $d = 11$ 和 $k = 20$ 时, 具有最小的 RMSECV 值, RMSECV = 0. 015 9 和 $R^2 = 0. 996\ 4$ 。

3. 2 数据集 1 的建模结果比较

对数据集 1 中各温度下的样本, 分别用 Isomap-PLS 和 PLS 算法建立校正模型, 结果如表 1 所示。可见, Isomap-PLS 比 PLS 建模效果好; 在某些情况下 Isomap-PLS 建模比 PLS 建模其 RMSECV 值小 2~5 倍。

进一步考察在最优参数条件下, PLS 和 Isomap-PLS 对 NIR 光谱降维的情况, 结果如表 2 所示。由表 2 可见, 高达 700 维的 NIR 原始光谱存在严重的信息冗余, 经 PLS 降维后有效维数一般在 5~10 之间。而原始光谱经 Isomap 降维后其维数一般较 PLS 稍大, 但进一步经 PLS 降维后维数有所降低, 与单独应用 PLS 相当。由于事实上 Isomap-PLS 较 PLS 具有更好的建模效果, 这说明 NIR 光谱可能存在一定的非线性流形结构, 且 Isomap 能够有效提取这一信息。

3. 3 数据集 2 的建模结果比较

对数据集 2 的四个性质分别用 Isomap-PLS 和 PLS 算法建立校正模型, 其 RMSECV 值、降维的结果分别如表 3 和表 4 所示。从表 3 可见, Isomap-PLS 算法的 RMSECV 比 PLS 算法都稍有减小, 但差别很小。这说明在数据集 2 中, NIR 光谱与性质之间主要是线性关系, 非线性不显著, 导致 Isomap-PLS 优势不明显。

4 结 论

Isomap 能够对 NIR 光谱进行流形降维, 提取有用信息, 结合 PLS 可实现 NIR 光谱的非线性建模。实验结果表明 Isomap-PLS 算法比单独的 PLS 算法建模效果更好, 特别是当性质数据与 NIR 光谱存在非线性关系时预测误差显著减小。本研究为将流形学习算法引入 NIR 光谱建模提供了一种有益尝试。

Table 1 RMSECV of Isomap-PLS in comparison with PLS on datasets measured at 30, 40, 50, 60 and 70

Algorithms	30			40			50			60			70		
	E	W	I	E	W	I	E	W	I	E	W	I	E	W	I
PLS	0. 029 6	0. 030 6	0. 053 7	0. 026 1	0. 031 1	0. 030 3	0. 021 1	0. 013 8	0. 014 2	0. 070 4	0. 033 8	0. 035 0	0. 077 3	0. 031 2	0. 044 8
Isomap-PLS	0. 026 1	0. 010 0	0. 027 0	0. 026 1	0. 015 5	0. 018 9	0. 019 5	0. 007 2	0. 013 4	0. 016 3	0. 009 1	0. 013 0	0. 015 9	0. 007 4	0. 012 6

Note: E: Ethanol; W: Water, I: Isopropanol

Table 2 Optimal dimension of PLS and Isomap-PLS on datasets measured at 30, 40, 50, 60 and 70

Algorithms	30			40			50			60			70		
	E	W	I	E	W	I	E	W	I	E	W	I	E	W	I
PLS	5	10	5	7	8	10	10	10	9	9	9	8	8	9	8
Isomap-PLS	Isomap	9	10	10	18	8	12	11	11	11	7	8	8	11	11
	PLS	8	10	10	7	8	10	9	9	9	7	8	8	9	10

Note: E: ethanol; W: Water, I: Isopropanol

Table 3 RMSECV of Isomap-PLS in comparison with PLS on datasets measured at NIR instrument m5, mp5 and mp6

Algorithms	m5				mp5				mp6			
	Moisture	Oil	Protein	Starch	Moisture	Oil	Protein	Starch	Moisture	Oil	Protein	Starch
PLS	0. 018 3	0. 062 2	0. 123 3	0. 257 1	0. 129 7	0. 092 7	0. 136 2	0. 349 9	0. 143 1	0. 093 1	0. 138 5	0. 337 8
Isomap-PLS	0. 018 2	0. 061 9	0. 123 1	0. 256 6	0. 125 7	0. 085 4	0. 133 9	0. 345 4	0. 143 0	0. 089 4	0. 131 6	0. 333 8

Table 4 Optimal dimension of PLS and Isomap-PLS on datasets measured at NIR instrument m5, mp5 and mp6

Algorithms	m5				mp5				mp6			
	Moisture	Oil	Protein	Starch	Moisture	Oil	Protein	Starch	Moisture	Oil	Protein	Starch
PLS	10	10	10	10	9	8	10	10	10	10	10	9
Isomap-PLS	Isomap	22	35	30	29	26	14	18	13	27	20	12
	PLS	10	10	10	10	10	9	10	10	10	8	10

参 考 文 献

- [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 Application(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. 1.
- [2] LU Wan-zhen, YUAN Hong-fu, XU Guang-tong, et al(陆婉珍, 袁洪福, 徐广通, 等). Modern Near Infrared Spectroscopy Analysis Technology (现代近红外分析技术). Beijing: China Petrochemical Press(北京: 中国石化出版社), 2001. 10.
- [3] WANG Hui-wen, WU Zai-bin, MENG Jie(王惠文, 吴载斌, 孟洁). Partial Least-Squares Regression Linear and Nonlinear Methods (偏最小二乘回归的线性与非线性方法). Beijing: National Defense Industry Press(北京: 国防工业出版社), 2006. 9.
- [4] Manabu K, Hiromu O, Shiji H, et al. Computers and Chemical Engineering, 2001, 25(7): 1103.
- [5] Wold S, Rrhe A, Wold H. SIAMJ Journal of Science Statistics Computing, 1984, 5(3): 735.
- [6] LUO Guo-an, YANG Hui-hua, WANG Yong, et al(罗国安, 杨辉华, 王勇, 等). In: LU Wan-zhen, YUAN Hong-fu, CHU Xiao-li, et al(陆婉珍, 袁洪福, 褚小立, 等). Near Infrared Spectroscopy Technology in Modern China - Proceedings of the 1st Chinese Conference on Near Infrared Spectroscopy. Beijing: China Petrochemical Press(北京: 中国石化出版社), 2006. 47.
- [7] ZHAO Chen, QU Hai-bin, CHENG Yi-yu(赵琛, 瞿海斌, 程翼宇). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2004, 24(1): 50.
- [8] WANG Feng-xia, ZHANG Zhuo-yong, WANG Ya-min(王凤霞, 张卓勇, 王亚敏). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2006, 26(5): 908.
- [9] Pascal C, Serge W, Michel U. Analytica Chimica Acta, 2007, 591(2): 219.
- [10] ZHANG Lu-da, JIN Ze-chen, SHEN Xiao-nan, et al(张录达, 金泽宸, 沈晓南, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2005, 25(9): 1400.
- [11] WU Rong-hui, SHAO Xue-guang(吴荣晖, 邵学广). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2006, 26(4): 617.
- [12] Ustun B, Melssen W J, Oudenhuijzen M, et al. Analytica Chimica Acta, 2005, 544(1): 292.
- [13] YANG Hui-hua, QIN Feng, LINAG Qiong-lin, et al. Chinese Chemical Letters, 2007, 18(7): 852.
- [14] YANG Hui-hua, QIN Feng, WANG Yong, et al(杨辉华, 覃锋, 王勇, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2007, 27(10): 1955.
- [15] Tenenbaum J B, Silva V D, Langford J C. Science, 2000, 290(22): 2319.
- [16] XU A, GUO P. Lecture Notes in Computer Science, 2006, 3972: 486.
- [17] ZHAO Lian-wei, LUO Si-wei, ZHAO Yan-chang, et al(赵连伟, 罗四维, 赵艳敞, 等). Journal of Software(软件学报), 2005, 16(8): 1423.
- [18] ZHAN De-chuan, ZHOU Zhi-hua(詹德川, 周志华). Journal of Computer Research and Development(计算机研究与发展), 2005, 42(9): 1533.
- [19] Cox T, Cox M. Multidimensional Scaling. London: Chapman and Hall, 2001.
- [20] http://www.bdagroup.nl/downloads/bda_downloads.html.
- [21] Wulfert F, Kok W T, Smilde A K. Anal. Chem. 1998, 70: 1761.
- [22] <http://www.models.kvl.dk/research/data/>.
- [23] <http://www.math.umn.edu/~wittman/mani/index.html>.

Isomap-PLS Nonlinear Modeling Method for Near Infrared Spectroscopy

YANG Hui-hua^{1,2}, QIN Feng¹, WANG Yi-ming², LUO Guo-an²

1. College of Computer and Control, Guilin University of Electronic Technology, Guilin 541004, China

2. Modern Research Center of Traditional Chinese Medicine, Tsinghua University, Beijing 100084, China

Abstract For modeling the nonlinear relationship existing between samples' near infrared (NIR) spectra and their chemical or physical properties, a novel modeling method was put forward in the present paper, which builds model by combining Isomap and partial least squares (PLS). Isomap is a newly proposed nonlinear dimension reduction algorithm, and belongs to the algorithm family of manifold learning, which is a new branch of machine learning. Isomap is based on multidimensional scaling (MDS) algorithm; however, it replaces the Euclidean distance in MDS with an approximated geodesic distance, so it can effectively find out the intrinsic low dimensional structure from high dimensional data. By combining Isomap and PLS, referred to as Isomap-PLS, a novel nonlinear modeling method for NIR spectra analysis was proposed. In this method, Isomap was used to extract nonlinear information from high dimensional NIR spectra while keeping the invariance of geometric property, and then PLS

was adopted to remove linear information redundancy and build a calibration model. The parameters of the Isomap, i. e. the number of the nearest neighbor k and output dimension d , can affect the performance of the method. In this paper, a grid search approach was used for parameter optimization. The Isomap-PLS modeling method was applied to two public benchmark NIR datasets, and the modeling results were compared with that of PLS. The results demonstrated that in both datasets, each model built with Isomap-PLS had a smaller rooted mean square error of cross-validation (RMSECV) than the corresponding model built with PLS. Moreover, for some properties, the RMSECV of Isomap-PLS was significantly reduced by a factor of 2-5 compared with that of PLS. It can be concluded that by taking the virtue that Isomap can reflect the intrinsic nonlinear structure of NIR spectra, Isomap-PLS can effectively model the nonlinear correlations between spectra and physicochemical properties of the samples, and so it gains more power in calibration and prediction than PLS.

Keywords Near infrared spectroscopy; Isomap; Nonlinear dimension reduction; PLS

(Received Aug. 10, 2007; accepted Nov. 20, 2007)

《光谱学与光谱分析》2009 年征订启事

欢迎投稿 欢迎订阅

《光谱学与光谱分析》1981 年创刊, 国内统一刊号: CN 11-2200/O4, 国际标准刊号: ISSN 1000-0593, CODEN 码: GYGFED, 国内外公开发刊, 大 16 开本, 288 页, 月刊; 是中国科协主管, 中国光学学会主办, 钢铁研究总院、中国科学院物理研究所、北京大学、清华大学共同承办的学术性刊物。北京大学出版社出版, 每期售价 35.00 元, 全年 420 元; 国内邮发代码 82-68, 国外发行代码 M905。刊登主要内容: 激光光谱测量、红外、拉曼、紫外、可见光谱、发射光谱、吸收光谱、X 射线荧光光谱、激光显微光谱、光谱化学分析、国内外光谱化学分析领域内的最新研究成果、开创性研究论文、学科发展前沿和最新进展、综合评述、研究简报、问题讨论、书刊评述。

《光谱学与光谱分析》适用于冶金、地质、机械、环境保护、国防、天文、医药、农林、化学化工、商检等各领域的科学研究单位、高等院校、制造厂家、从事光谱学与光谱分析的研究人员、高校有关专业的师生、管理干部。

《光谱学与光谱分析》为我国首批自然科学核心期刊, 中国科协优秀科技期刊, 中国科协择优支持基础性、高科技学术期刊, 中国科技论文统计源刊, “中国科学引文数据库”, “中国物理文摘”, “中国学术期刊文摘”, 同时被国内外的 CSCD, SCI, AA, CA, Ei, PKI, MEDLINE 等文献机构收录。根据国家科技部信息研究所发布信息, 中国科技期刊物理类影响因子及引文量《光谱学与光谱分析》都居前几位。欢迎国内外厂商在《光谱学与光谱分析》发布广告 (广告经营许可证: 京海工商广字第 8094 号)。

《光谱学与光谱分析》的主编为黄本立院士。

欢迎新老客户到全国各地邮局订阅, 若有漏订者可直接与光谱学与光谱分析期刊社联系。

联系地址: 北京市海淀区学院南路 76 号, 光谱学与光谱分析期刊社 邮政编码: 100081

联系电话: 010-62181070, 62182998

电子信箱: chngpxygpfx@vip.sina.com; 修改稿专用邮箱: gp2008@vip.sina.com

网 址: <http://www.gpxygpfx.com>