联合矩阵局部保持投影的近红外光谱特征提取

胡善科[1]，秦玉华[1*]，段如敏[2]，吴丽君[2]，官会丽[3]

1. 青岛科技大学信息科学技术学院，山东 青岛 266061
2. 云南中烟工业有限责任公司技术中心，云南 昆明 650024
3. 中国海洋大学信息科学与工程学院，山东 青岛 266100

摘 要 近红外光谱存在高维、噪声大、重叠和非线性等特性，严重影响建模准确，因此提出了一种基于联合矩阵局部保持投影(JMLPP)的特征提取方法。首先，利用基于聚类的光谱特征选择方法对原始近红外光谱数据进行有效特征提取，按种与分类相关性强的指标将样本分为种不同的聚类方式，依据类内关联性强，类间差异性大的聚类思想，通过调节类内参数、类间参数确定类内阈值与类间阈值，分别对种不同聚类方式筛选光谱特征区间，得到指标特征矩阵，并集操作生成联合矩阵。其次，从两个方面对局部保持投影算法(LPP)进行了改进：引入测地距离构造邻域距离矩阵，较欧式距离更好的表达了高维数据样本点间的拓扑结构；改进了边权矩阵，解决了样本稀疏导致的不确定性，避免了有效信息的丢失。最后，采用改进的LPP算法对联合矩阵进行降维操作，从而得到最优光谱特征子集。为验证JMLPP算法有效性，首先从光谱投影方面将该算法与PCA、LPP算法进行了对比，结果表明JMLPP算法有较好的等级区分能力，投影空间中的烟叶样品分类清晰，明显优于PCA与LPP算法。其次从模型分类准确性方面进行了对比，分别采用全谱段与PCA，LPP和JMLPP降维后的特征建立烟叶等级分类模型，实验结果表明，JMLPP算法建立的分类模型准确率为93.8%，对5种烟叶分级的敏感度分别为95.2%，93.1%，94.2%，92.1%和92.5%，特异度分别为99.3%，98.4%，98.6%，97.5%和97%，模型准确率、敏感度与特异度均明显优于其他3种方法。该算法通过基于聚类的特征提取和改进的局部保持投影算法实现了烟叶分级特征的有效提取，并保留原始数据的局部线性关系，使最终建立的模型具有良好的稳定性和较高的准确性。

关键词 特征提取；联合矩阵；测地线距离；局部保持投影算法；近红外光谱

中图分类号：O657.33 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2020)12-3772-06

引 言

近红外光谱技术具有快速、高效、准确性好，不损坏样品等特点，目前大量用于石油化工、环境科学、食品药品等领域[1]。我国是烟草大国，每年的烟叶收购量庞大，但烟叶质量受各种因素的影响，需首先经过分级处理才能保证原料的合理利用。然而目前烟叶分级主要以人工为主，烟叶分级存在主观性强、效率低、误差大，利用率低等问题[2]。随着近红外光谱技术的发展，近年来，它在烟草自动分类中得到了很好的应用，不仅能获得烟叶颜色的外观特征，而且能反映烟叶的内在质量信息，与人工、图像视觉提取、数学推理等分类技术相比具有天然优势[3]。然而，近红外光谱数据具有高维、频带重叠、噪声大和非线性等特点，高维空间的稀疏性与空空间等现象也严重影响了结果的准确性，针对这些问题，对高维光谱数据进行与建模相关性高的特征提取尤为重要[4]。鲁梦瑶等提出采用隔点采样的方法对光谱数据进行特征提取，从而加快收敛速度，但该方法容易丢失重要特征；何勇等[5]采用主成分分析(principal component analysis，PCA)与神经网络相结合的方法对光谱数据进行降维，并以PCA变换后的变量作为输入参数，但PCA是一种线性降维方法，无法获取数据的非线性结构特征；高全学等[6]提出了改进(local preserving projection，LPP)的非线性降维算法，在特征提取过程中，融合了局部结构和差分信息，但对稀疏数据的效果并不理想。

针对上述问题，提出了一种基于联合矩阵的局部保持投

收稿日期：2019-11-10，**修订日期**：2020-03-19

基金项目：国家重点研发计划项目(2018YFB1701704)，云南中烟工业有限责任公司项目(2018JC01)资助

作者简介：胡善科，1994年生，青岛科技大学信息科学技术学院硕士研究生 e-mail：1020447071@qq.com

*通讯联系人 e-mail：yuu71@163.com

影(local preserving projection algorithm based on joint matrix, JMLPP)特征提取方法。首先，通过基于聚类的特征提取[7]剔除类内相关度低、类间相关度过高的特征，实现了光谱中噪声信息的剔除。其次，采用改进的 LPP 算法对光谱数据进行降维，解决了冗余特征和非线性结构的影响。此外，在 LPP 算法中引入测地线距离[8]，并对边权矩阵公式进行了改进，解决了样本稀疏带来的不确定性。JMLPP 方法实现了烟叶分级信息的有效提取，提高了烟叶分级准确性。

1　算法与原理

1.1　基于聚类的特征提取

分类是一个复杂过程，其中包括多个指标，需要考虑与分类相关指标的最优特征范围是否相同，很明显多个指标的最优特征范围不可能完全相同，因此不同指标的最优特征范围之和会构成分类的最大且最优特征。假设给定样本集共有 n 类，每类有 M 个 k 维数据，那么第 l 类有 m_l 个样本，第 j 个样本表示为 $X_j^{(l)}=(x_{j1}^{(l)}, x_{j2}^{(l)}, \cdots, x_{jk}^{(l)})$，其中 $j=(1, 2, \cdots, m_l)$，$l=(1, 2, \cdots, n)$。对于分类问题，聚类的思想是把相似性高的对象归为同一类，且不同的类别组之间具有较大的差异性。依据这种思想，每个类别的 m_l 个样本的同维特征聚集性越高，均方差越小，说明这一特征越典型，应该保留；反之，应该剔除。令第 l 类的第 i 个特征的均方差维为 $\bar{\sigma}_i^{(l)}$，根据每类样本的 k 个特征，求出类内阈值 $D^{(l)}$

$$D^{(l)}=\min_{i=1}^{k}(\bar{\sigma}_i^{(l)})+\gamma_1\left[\max_{i=1}^{k}(\bar{\sigma}_i^{(l)})-\min_{i=1}^{k}(\bar{\sigma}_i^{(l)})\right] \tag{1}$$

其中 $\gamma_1\in[0, 1]$。通过调节类内参数 γ_1 找到合适类内阈值之后，筛选出的光谱特征都具有较好的类内聚集性，但有些特征存在每类均值都比较接近的问题，这些特征无法较好的体现分类的作用。依据聚类思想，n 类样本的同维特征均值差异性越大，均方差越大，说明这一特征越典型，应该保留；反之，应该剔除。令 n 类样本的第 i 个特征集的均方差为 $\bar{\sigma}_i$，根据样本的 k 个特征，求出类间阈值 D

$$D=\min_{i=1}^{k}(\bar{\sigma}_i)+\gamma_2\left[\max_{i=1}^{k}(\bar{\sigma}_i)-\min_{i=1}^{k}(\bar{\sigma}_i)\right] \tag{2}$$

其中 $\gamma_2\in[0,1]$。通过调节类间参数 γ_2 找到合适类间阈值之后，筛选出的特征具有较好的离散性。联合类内阈值与类间阈值对光谱数据的处理，最后得到筛选出的指标特征矩阵。

分类方式可能有 N 种，可得到 N 个指标特征矩阵，考虑到分级的准确性，对得到的 N 个指标特征矩阵进行并集操作得到联合矩阵。选取与烟叶分级相关性高的成熟度与部位指标进行分类，从光谱矩阵中分别选出与成熟度和部位相关性高的特征，从而得到两个特征矩阵，并集产生一个联合矩阵。通过联合矩阵运算可减少“维度灾难”问题，剔除与分类无关的噪声信息，提高计算精度，但仍存在光谱数据冗余、非线性等特点。

1.2　改进的局部保持投影算法

局部保持投影(LPP)算法[9]是由何小飞教授于 2003 年提出，LPP 是一种线性降维和非线性降维相结合的降维算法。与 PCA 算法相比，LPP 算法能够保留全局信息，在线性降维的同时也保留局部非线性特征。LPP 生成的表现映射可看作 LE (laplacian eigenmap)[10]的线性近似，保留了数据的局部信息，应用在高光谱数据和图像识别等领域[11]。

给定 m 个在欧式空间 R^N 的 N 维数据样本 $X=\{x_1, x_2, \cdots, x_m\}$，$x_j\in R^N$，$(j=1, 2, \cdots, m)$，LPP 通过生成最近局部邻域图，获得样本数据的 k 近邻域。LPP 的目标是将高维空间非线性流行数据 X 投影到低维空间特征映射矩阵 $\boldsymbol{Y}$，找到最优转换矩阵 $\boldsymbol{Z}$，其本质是 Laplacian Eigenmap 的线性逼近，如式式(3)

$$y_j=\boldsymbol{Z}^{\mathrm{T}}x_j \tag{3}$$

优化目标函数后为

$$\min\sum_{j,\,i}(y_j-y_i)^2\boldsymbol{W}_{ji} \tag{4}$$

LPP 算法为了保证映射后矩阵能最大程度保存数据局部结构属性，使距离较近的样本 x_j，x_i 经过映射后仍保持较近距离，引入相似性度量矩阵 $\boldsymbol{W}_{ji}$

$$\boldsymbol{W}_{ji}=\begin{cases}\exp\left(\dfrac{\|x_j-x_i\|^2}{\delta}\right)\\ 0\end{cases} \tag{5}$$

其中 x_j 和 x_i 互为 k 邻域内的点，δ 是一个常数，$\boldsymbol{W}$ 为实对称矩阵。

对优化目标函数进行变化

$$\sum_{j,\,i}(z^{\mathrm{T}}x_j-z^{\mathrm{T}}x_i)\boldsymbol{W}_{ji}=z^{\mathrm{T}}\boldsymbol{X}(\boldsymbol{D}-\boldsymbol{W})\boldsymbol{X}^{\mathrm{T}}z=z^{T}\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}z \tag{6}$$

式(6)中，$\boldsymbol{D}$ 是对角矩阵，即 $\boldsymbol{D}_{jj}=\sum_i\boldsymbol{W}_{ji}$，$\boldsymbol{L}$ 为 Laplacian 矩阵，$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W}$，为防止 0 向解，添加约束条件

$$\text{s. t.}\quad z^{\mathrm{T}}\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^{\mathrm{T}}z=1 \tag{7}$$

则最小化目标函数为

$$\operatorname{argmin}z^{\mathrm{T}}\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}z \tag{8}$$

即求解下式广义矩阵特征值

$$\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}z=\lambda\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^{\mathrm{T}}z \tag{9}$$

矩阵 $\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}^{\mathrm{T}}$，$\boldsymbol{X}\boldsymbol{L}\boldsymbol{X}^{\mathrm{T}}$ 是对称且半正定的，式(9)得到前 h 个最小特征值的特征向量 z_1，z_2，$\cdots$，z_h 构成最优转换矩阵 $\boldsymbol{W}=(w_1, w_2, \cdots, w_z)$。

LPP 算法在保持全局非线性结构的同时进行局部线性降维，但烟叶光谱数据具有高冗余、高噪声、重叠、离散性大等特点，且 LPP 算法单纯依据欧式距离构造邻域图，无法表达样本点间真实的拓扑结构，对烟叶近红外光谱数据的处理存在一定不足。本文对 LPP 算法作了如下改进：用测地线距离代替欧式距离，根据 Dijkstra 算法得到的最小距离构造邻域图，并改进边权矩阵。利用贪心算法得到样本中某一点距离较近的前 k 个顶点，作为 k 近邻域。

设构造的邻域图为：$G=\{V, E, W\}$，其中 V 为样本顶点集合，E 是边集合，W 是边权矩阵，设测地线距离为 $d_G(x_j, x_i)$，则改进后的边权矩阵为

$$\boldsymbol{W}_{ji}=\begin{cases}\exp\left(\dfrac{d_G(x_j, x_i)}{\delta}\right)\\ 0\end{cases} \tag{10}$$

在离散性大的高维流形数据中，测地线距离可以较好的表达两点之间的实际距离，使样本点整体分布趋于均匀，相

较于欧式距离具有明显优势，提高了 LPP 的降维效果。

1.3 基于联合矩阵的局部保持投影特征提取方法

基于联合矩阵的局部保持投影(JMLPP)特征提取方法具体步骤如下：

(1)按 N 种与分类相关性强的指标将样本分为 N 种不同的分类方式，每种分类方式筛选 k 个特征进行基于聚类的特征选择。

(2)基于聚类的特征选择需要挑选类内关联性强，类间差异性大的特征。通过调节类内参数 γ_1、类间参数 γ_2 确定类内阈值 $D^{(l)}$ 与类间阈值 D，分别对 N 种不同聚类方式筛选光谱特征区间得到 N 个指标特征矩阵 M_1，M_2，…，M_N，并集操作生成联合矩阵 $\boldsymbol{M}$。

(3)将联合矩阵 $\boldsymbol{M}$ 采用改进的 LPP 算法进行降维操作，得到去噪、去冗余的数据特征子集 $Y=\{y_1, y_2, \cdots, y_m\}$。

2 实验部分

2.1 样品制备

来自某烟草企业提供的包括 B2V，B1F，C4F，C1L，X2L 五个不同等级共 650 个烟叶样品，其中每个等级各 130 个。将样品放置在 60 ℃的烘箱中干燥 2 h，磨粉过 60 目筛，密封平衡 8 h 后进行光谱采集。

2.2 烟叶光谱采集与预处理

使用赛默飞世尔公司 Antaris Ⅱ近红外光谱仪，采用漫反射方式，扫描范围为 3 800～10 000 cm^{-1}，分辨率为 8 cm^{-1}，室温保持在 18～22 ℃，每个样品取 15 g 压实后置于光谱仪中扫描 3 次，计算其平均值作为最终光谱。

为了消除基线漂移和噪声的影响，需要对采集到的光谱数据进行预处理，经比较本文采用一阶导数和 Savitzky Golay平滑[12]。

3 结果与讨论

3.1 聚类参数、的确定与特征提取

因影响烟叶分级的关键指标包括成熟度与部位，分别从 650 个样品中按成熟度与部位选取部分特征明显的烟叶样品进行基于聚类的特征提取。其中按成熟度分为成熟、尚熟与假熟，共选取了 420 个样品；按部位分为上部、中部与下部，共选取了 450 个样品。具体样品信息划分如表 1 所示。

表 1 聚类特征提取实验样品划分

Table 1 Sample partition of cluster feature extraction experiment

成熟度	训练样本	测试样本	部位	训练样本	测试样本
成熟	120	40	上部	120	40
尚熟	98	32	中部	120	40
假熟	98	32	下部	98	32

首先利用基于聚类的特征提取方法分别从成熟度和部位指标筛选与烟叶分级相关的特征。根据文献[10]与实验分析，类内参数 γ_1、类间参数 γ_2 的取值分别在 0.9～1，0～0.01 之间细化搜索得到最佳取值。图 1 和图 2 分别为 γ_1 和 γ_2 按部位和成熟度聚类的搜索结果。

可以看出，按部位分组时，类内参数 $\gamma_1=0.95$，类间参数 $\gamma_2=0.000\ 4$ 时识别率较好，提取的光谱数据特征为 983 个。按成熟度分组时，类内参数 $\gamma_1=0.95$，类间参数 $\gamma_2=0.001\ 4$ 时识别率较好，提取的光谱数据特征为 892 个。为保证信息提取的完整性，本文将两个特征子集进行并集操作生成一个联合矩阵，联合矩阵的光谱特征从 1 560 减少到 1 102 个，较全光谱数据减少了 28.9%。

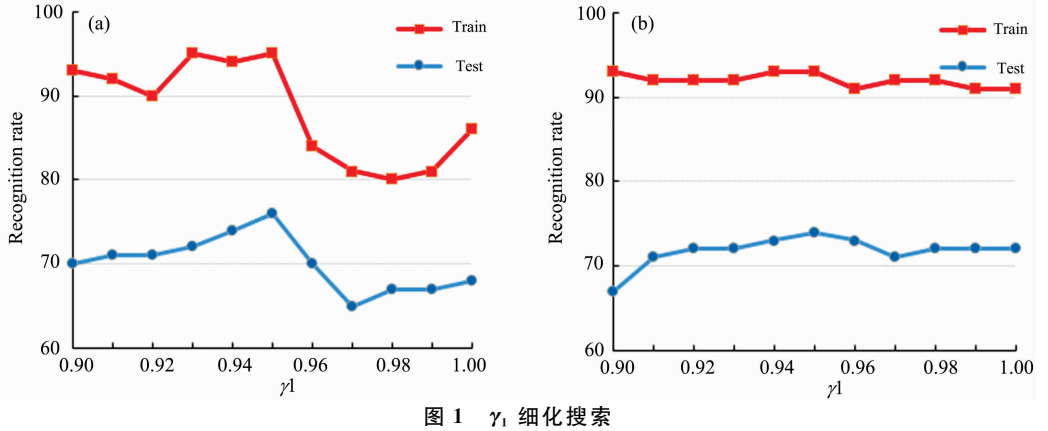

图 1 γ_1 细化搜索

(a)：γ_1 部位分组；(b)：γ_1 成熟度分组

Fig. 1 Refined search of γ_1

(a)：γ_1 grouped by location；(b)：γ_1 grouped by maturity

3.2 降维投影分析

特征选择可消除对分级无关的噪声特征，但筛选出的光谱数据仍存在冗余、非线性特征，这将对烟叶分级的准确性产生影响，因此采用改进的 LPP 方法对提取的特征进行进一步降维处理，从而消除冗余特征的影响。图 3—图 5 为 JMLPP 与 PCA，LPP 的投影对比。

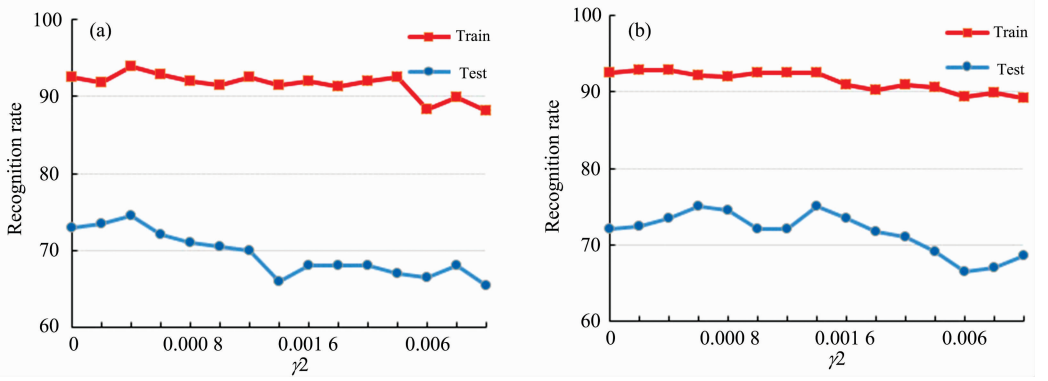

图 2　γ_2 细化搜索

(a)：γ_2 部位分组；(b)：γ_2 成熟度分组

Fig. 2　Refined search of γ_2

(a)：γ_2 grouped by location；(b)：γ_2 grouped by maturity

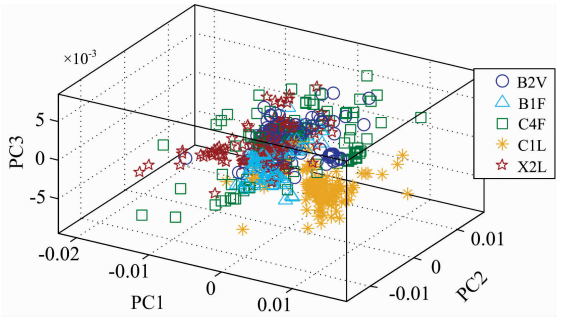

图 3　PCA 投影图

Fig. 3　PCA projection plot

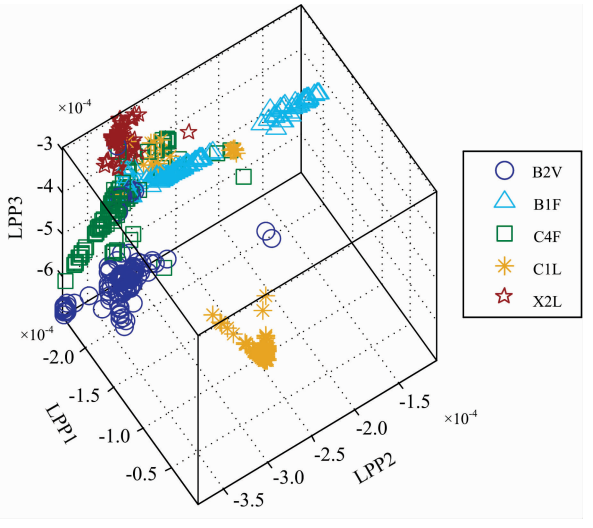

图 4　LPP 投影图

Fig. 4　LPP projection plot

可以看出，PCA 投影空间中样品混合现象比较严重，各等级边界模糊，难以实现烟叶等级的区分。LPP 投影空间中的烟叶等级分类效果好于 PCA，但仍存在较多样品区分模糊问题。而 JMLPP 投影空间中的烟叶样品分类清晰，效果明显好于 PCA 与 LPP，说明该方法有较好的等级区分能力。

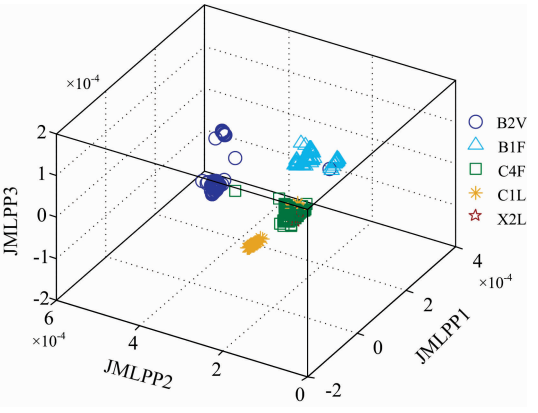

图 5　JMLPP 投影图

Fig. 5　JMLPP projection plot

3.3　分类结果对比分析

选取 75%的样本做为训练集，25%的样本做为测试集，分别采用全谱段与 PCA，LPP 和 JMLPP 降维后的特征建立烟叶等级分类模型。几种降维方法选取前 6 个成分做为输入指标，采用 SVM 做为分类器。表 2 为几种方法下郴同等级烟叶分类准确性对比，为防止偶然性，准确率取 5 次实验结果的平均值。

表 2　烟叶分级结果对比

Table 2　Comparison of tobacco leaf grading results%

	B2V	B1F	C4F	C1L	X2L	总体准确率
全谱段	58.9	62.8	65.0	70.2	68.6	65.1
PCA	63.1	69.5	64.2	71.1	69.6	67.5
LPP	79.8	80.4	75.9	73.6	81.8	78.3
JMLPP	96.7	95.7	89.1	94.9	92.6	93.8

由表 2 可以看出，对于每个等级烟叶的分类准确率，全谱段做为输入特征效果最差，主要由于高维光谱中存在较多噪声和冗余信息，无法实现烟叶分级信息的有效提取，影响了分类的准确性。JMLPP 方法烟叶总体分类的准确率为

93.8%，每个等级的分类准确性都明显高于其他方法，说明该方法能较好的对烟叶分级信息进行提取，这与前面投影分析结果一致。

敏感度与特异度可以分别衡量算法对于正例与负例的识别能力，表 3 为几种分级算法模型对 5 种等级烟叶分类的敏感度与特异度对比。

表 3 烟叶分级算法敏感度与特异度对比

Table 3 Comparison of sensitivity and specificity of tobacco leaf classification algorithms

	敏感度/%					特异度/%				
	B2V	B1F	C4F	C1L	X2L	B2V	B1F	C4F	C1L	X2L
全谱段	67.1	66.4	63.2	65.1	64.3	88.3	87.2	89.6	87.2	88.5
PCA	67.3	70.1	67.6	66.2	65.2	89.3	90.2	91.3	88.6	89.4
LPP	80.5	79.8	76.3	78.1	77.4	96.1	94.8	95.3	94.6	92.8
JMLPP	95.2	93.1	94.2	92.1	92.5	99.3	98.4	98.6	97.5	97.0

可以看出，JMLPP 算法的敏感度、对烟叶等级的识别错误率明显好于其他几种方法，进一步说明 JMLPP 方法具有较好的鲁棒性。

4 结 论

基于联合矩阵局部保持投影算法较好的解决了光谱数据高维、重叠、高噪声的问题。该方法通过聚类实现了与分类相关性强的多个特征子集的提取，并集后得到联合矩阵，有效降低了光谱数据维度，减少了噪声干扰。通过对 LPP 算法的改进，解决了高维数据欧氏距离度量不准确的问题，提高了降维效果。实验结果表明，JMLPP 方法对于烟叶等级判定具有更好的准确率与鲁棒性，可以作为烟叶分级的一种新方法。下一步，需要提高算法效率，拓宽算法的应用范围。

References

[1] KONG Qing-qing, DING Xiang-qian, GONG Hui-li(孔清清，丁香乾，宫会丽). Laser & Optoelectronics Progress(激光与光电子学进展), 2018, 55(1): 013006.

[2] YAO Xue-lian, HE Fu-qiang, PING An, et al(姚学练，贺福强，平 安，等). Tobacco Science & Technology(烟草科技), 2009, 42(11): 197.

[3] ZHANG-Ying, HE Li-yuan(章 英，贺立源). Transactions of the Chinese Society of Agricultural Engineering(农业工程学报), 2011, 27(4): 350.

[4] Yun Yonghuan, Li Hongdong, Deng Baichuan, et al. Trends in Analytical Chemistry, 2019, 113: 102.

[5] HE Yong, LI Xiao-li(何 勇，李晓丽). Journal of Infrared and Millimeter Waves(红外与毫米波学报), 2006, 25(3): 192

[6] GAO Quan-xue, XIE De-yan, XU Hui(高全学，谢德燕，徐 辉，等). Acta Automatica Sinica(自动化学报), 2010, 36(8): 1107.

[7] ZHAO Hai-dong, SHEN Jin-yuan, LIU Run-jie, et al(赵海东，申金媛，刘润杰，等). Infrared Technology(红外技术), 2013, 35(10): 659.

[8] Tenenbaum J B, de Silva V, Langford J C. Science, 2000, 290(5500): 2319.

[9] He Xiaofei, Niyogi P. Cambridge, Locality Preserving Projections, NIPS'03: Proceedings of the 16th International Conference on Neural Information Processing Systems, 2003. 153.

[10] Belkin M, Niyogi P. Neural Computation, 2003, 15(6): 1373.

[11] Zhai Y, Zhang L, Wang N, et al. IEEE Geoscience & Remote Sensing Letters, 2017, 13(8): 1059.

[12] HU Yong, WANG Yu-heng, LIU Wei, et al(胡 涌，王宇恒，刘 伟，等). Journal of China Agricultural University(中国农业大学学报), 2018, 23(3): 106.

Research on Feature Extraction of Near-Infrared Spectroscopy Based on Joint Matrix Local Preserving Projection

HU Shan-ke[1], QIN Yu-hua[1*], DUAN Ru-min[2], WU Li-jun[2], GONG Hui-li[3]

1. College of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266061, China
2. Technical Research Center, China Tobacco Yunnan Industrial Co., Ltd., Kunming 650024, China
3. College of Information Science and Engineering, China Ocean University, Qingdao 266100, China

Abstract Aiming at the problem that the high-dimensional, high-noise, overlap and nonlinear features of the near-infrared spectrum seriously affect the modeling accuracy, a feature extraction method based on joint matrix local preservation projection (JMLPP) is proposed in this paper. First, the cluster-based spectral feature selection is used for effective features extraction. According to kinds of indicators with a strong correlation of classification, the samples are divided into kinds of different clustering modes. Based on the idea of strong intra-class correlation and great inter-class difference, the intra-class threshold and the inter-class threshold are determined by adjusting the intra-class parameter and the inter-class parameter. The spectral feature regions are selected according to kinds of different clustering modes, and feature matrices are obtained, whereas a joint matrix is generated by the union operation. Cluster-based feature extraction eliminates features with low intra-class correlation and high correlation between classes, and realizes the elimination of noise information in the spectrum. Secondly, the local preservation projection algorithm (LPP) is improved in this paper from two aspects: the geodesic distance is introduced to construct the neighborhood distance matrix, and the topology between the high-dimensional sample data is better expressed than the Euclidean distance. Meanwhile, the edge weight matrix is also improved, which solves the uncertainty caused by sample sparseness and avoids the loss of effective information. Finally, the improved LPP algorithm is used to reduce the dimensionality of the joint matrix, and the optimal spectral feature subset of the low-dimensional mapping is obtained. In order to verify the effectiveness of the JMLPP algorithm, this paper first compares the JMLPP with PCA and LPP from the perspective of spectral projection. The results show that JMLPP has better classification ability, and the tobacco samples in the projection space are clearly classified, and the effect is obviously better than PCA and LPP. In addition, the results of the model classification are also compared. The classification models were established by using the full spectra and dimension reduction features of the PCA, LPP and JMLPP. The experimental results show that the accuracy of the classification model established by JMLPP algorithm is 93.8%. The sensitivity of the five categories of tobacco grading classification are 95.2%, 93.1%, 94.2%, 92.1%, 92.5%, and the specificities are 99.3%, 98.4%, 98.6%, 97.5%, and 97%, respectively. The accuracy, sensitivity and specificity of the model are significantly higher than the other three methods. The JMLPP algorithm effectively extracts useful information of classification based on cluster-based feature extraction and local preserving projection algorithm, and maintains the local linear relationship of the original data. The stability and accuracy of model are desirable.

Keywords Feature extraction; Joint matrix; Geodesic distance; Local preservation projection algorithm; Near-infrared spectroscopy

(Received Nov. 10, 2019; accepted Mar. 19, 2020)

* Corresponding author