ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 31,No. 6,Nov. 2016,pp. 1097－1105
DOI:10.16337/j.1004-9037.2016.06.003

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

光谱数据的特征挖掘降维方法

戴琼海　张　晶　李菲菲　范静涛

(清华大学自动化系,北京,100084)

摘　要:“去繁存精”的光谱数据解耦方法可去除高维光谱数据的大量冗余,提炼其特征谱段,是光谱仪器得以广泛应用的重要基础。应用各异性和光谱特征优选方法普适性所构成的矛盾,在一定程度上制约了光谱仪器的应用。本文提出了序列前向选择(Sequential forward selection,SFS)的光谱特征自适应数据挖掘方法,生成最优变量组合作为支持向量机(Support vector machine,SVM)分类模型的输入,在对光谱数据降维的同时,实现了高精度的数据分类。本文方法可有效解决大量光谱数据的多类分类问题,并在红木分类中得到了实际验证和应用,为破解因光谱特征峰高度混叠而难以进行主观经验特征选择的困境提供了新思路。

关键词: 光谱数据;特征挖掘;序列前向选择;数据降维

中图分类号: O433.4　　**文献标志码:** A

Dimension Reduction of Spectral Data Based on Feature Mining

Dai Qionghai, Zhang Jing, Li Feifei, Fan Jingtao

(Department of Automation, Tsinghua University, Beijing, 100084)

Abstract: The method of spectral data analysis, which can remove a lot of redundancy of high-dimensional spectral data and extract its characteristic spectrum, is an important foundation for the widespread application of spectral instruments. The contradiction of the applicability of the heterogeneity and spectral characteristics of the method of universal selection, to a certain extent, restricts the application of spectral instruments, need to be resolved. In this paper, a sequential forward selection (SFS) spectral feature adaptive data mining method is proposed to generate the optimal combination of variables as support vector machine (SVM) classification model input, to achieve the spectral data reduction and obtain a high-precision data classification. This method can effectively solve the problem of multi-class classification of a large number of spectral data, which is proved and applied in the classification of mahogany. It provides a new way to solve the difficulty of subjective experience feature selection in height-aliasing of spectral peaks.

Key words: spectral data; feature mining; sequential forward selection; dimension reduction

引　　言

光谱信息能够反映物质的结构和成分,是物质的 DNA。在光谱数据采集方面,文献[1,2]提出了一

基金项目:国家自然科学基金委国家重大科研仪器设备研制专项(61327902)资助项目;国家自然科学基金面上(61271450)资助项目。

收稿日期:2016-11-25;**修订日期:**2016-11-29

种基于棱镜-掩膜的新型双通道光谱仪(Prism-mask imaging spectrmeter,PMIS),核心思想是在不牺牲时间分辨率的条件下直接采集光谱的视频信息,通过红绿蓝和位置空间的两个高斯函数来实现光谱融合,最终获得在光谱(5 nm@550 nm)、空间(1 024×768)和时间(5～15 帧/s)三个维度上的高分辨率信息获取。针对高光谱视频采集方法重构精度受限问题,文献[3]提出一种编码压缩高光谱成像方法,通过空间光谱编码光学相机设计,完整的高光谱字典学习和稀疏约束计算重建,得到了鲁棒的非线性稀疏重建方法,并从具有更高性能的编码投影中恢复高光谱图像。为解决低光子效率与有限的光谱范围和高成本之间的矛盾,文献[4]提出用单桶检测器进行多光谱成像,得到了高灵敏度、宽光谱范围、低成本、小尺寸和质量轻的探测器,利用检测器的快速响应,场景的 3D 空间光谱信息被复用到密集的 1D 测量序列中,然后在单像素成像方案下解算复用,最终捕获从 450～650 nm 范围内的 64 像素×64 像素×10 波长带的多光谱数据,采集时间为 1 min,该成像方案对于各种低光照和高空应用具有巨大的潜力,可以用于制造便携式多光谱成像器件。近年来,清华大学实现了高性能建模、高效率编码调制,并对时-空-光三维数据进行了联合优化采样,获得了高时空分辨率光谱视频的重建[5-10]。通过这些多样的采集设备获取光谱数据后,需要配合在光谱维度对数据进行自适应筛选,最终实现高性能、微型化和低成本的智能检测设备。本文的光谱数据解耦方法基于各类光谱采集仪器,对未经化学分离的混合物样品进行无损采集后,提取光谱有效信息用于样品分类的精准数据建模。由于直接采集的光谱数据包含若干不同级别的光谱特征倍频信息、不同组合的合频信息以及大量冗余信息,特征峰高度混叠,信噪比较低,难以用单一特征对光谱进行解析。因此在解决这类多类样本分类问题时,光谱特征谱段挖掘至关重要,特征选择的正确性直接决定了物质分类的准确性。本文针对光谱数据存在的特征峰高度混叠和大量冗余信息,提出了数据挖掘的特征提取和降维方法。通过自底向下的序列前向选择算法(Sequential forward selection,SFS),以分类贡献度作为量化指标,建立最优分类面的精确预测模型,自适应构建最优特征谱段变量组合来精化样本数据,并训练支持向量机(Support vector machine,SVM)进行多类分类,在光谱数据降维的同时,获得高精度、高鲁棒的分类结果。本文提出的将高光谱问题转化为多光谱问题的普适方法,为研制低复杂度、小型化光谱仪器[11]开辟了新思路。

1 光谱数据特征选择方法

光谱数据特征选择方法需要同时具有分类准确率高、可筛选出特征谱段和方法鲁棒稳定,才能实现对多类样品的分类辨识。特定问题的光谱特征选择通常通过主观经验或构建目标函数实现[12,13],主观经验在进行纯净物样品分类辨别时非常有效,但是大量无损采集的样品未经过化学提纯,尤其近红外光谱谱段的光谱数据特征峰高度混叠、信号噪声高且无法人工辨识。因此需要进行有效的谱段筛选[14,15],常用方法包括:均匀采样、最优求解、主成分分析[16,17]和线性判别式分析(Linear discriminant analysis,LDA)等。均匀采样对所有谱段都设定相同的优先级,不能保证必然选取到对分类贡献最大的维度,分类准确率存在偶然性,其鲁棒性和稳定性不高。最优求解使用分支定界的搜索树方法,其所依赖的判据对特征的单调性在包含无关信息和噪声的高维光谱数据中无法保证,特征增多,判据值不会减少,在进行多类分类时不能做到数据特征的降维,也就无法筛选出特征谱段。主成分分析按照 K-L 展开式对高维数据做降维变换,从而提炼新的特征,但利用全谱数据进行变化的计算时空复杂度高,且降维的特征矩阵和原矩阵不存在线性对应关系,同样无法指导特征谱段筛选。LDA 通过找到一个投影面,使得类间距离最大化,类内距离最小化,达到最好的分类效果和特征压缩以及分类信息抽取的作用。对于矩阵,在 LDA 中包括类内矩阵和类间矩阵,其中类内离散度矩阵为

$$\boldsymbol{S}_w=\sum_{k=1}^{C}S_k=\sum_{k=1}^{C}\sum_{i=1}^{N_k}(\boldsymbol{x}_i-\boldsymbol{m}_k)(\boldsymbol{x}_i-\boldsymbol{m}_k)^{\mathrm{T}}=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{X} \tag{1}$$

式中:C 为类别数,$\boldsymbol{S}_k$ 为 k 类的类内离散度矩阵,N_k 为 k 类的样本数,$\boldsymbol{m}_k$ 为 k 类的均值向量,$\boldsymbol{X}$ 为 $\boldsymbol{S}_w$ 的

基向量矩阵,W 定义为

$$W_{ij}=\begin{cases}\dfrac{1}{N_k} & x_i,x_j\in k\text{-class}\\ 0 & \text{其他}\end{cases} \tag{2}$$

类内矩阵可以表示为

$$\boldsymbol{S}_B=\sum_{i=1}^{C}\sum_{j=1,j\neq i}^{C}(\boldsymbol{m}_i-\boldsymbol{m}_j)(\boldsymbol{m}_i-\boldsymbol{m}_j)^{\mathrm{T}}=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{X} \tag{3}$$

式中:$\boldsymbol{m}_i$ 第 i 类的均值向量,$\boldsymbol{m}_j$ 为第 j 类的均值向量,$\boldsymbol{B}$ 可以表示为

$$\boldsymbol{B}=\boldsymbol{I}_n-\frac{1}{N}\boldsymbol{e}^{\mathrm{T}}\boldsymbol{e},\ \boldsymbol{e}=[1,1,\cdots,1]\in\mathbf{R}^n \tag{4}$$

LDA 的模型可以写成

$$\min\mathrm{Tr}(\boldsymbol{S}_w-u\boldsymbol{S}_B)=\min\mathrm{Tr}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{X}-u\boldsymbol{X}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{X})=\min\mathrm{Tr}(2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{X}-u\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})=\min\mathrm{Tr}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{X}-\lambda\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}) \tag{5}$$

LDA 的投影矩阵可以变为

$$\min_{\boldsymbol{A}}\mathrm{Tr}(\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{X}\boldsymbol{A}^{\mathrm{T}}-\lambda\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{A}^{\mathrm{T}}) \tag{6}$$

然后加入 $\boldsymbol{A}$ 的 $L_{2,1}$-norm 惩罚项,式(1~5)可以写为

$$\min_{\boldsymbol{A}}\mathrm{Tr}(\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{L}\boldsymbol{X}\boldsymbol{A}^{\mathrm{T}}-\lambda\boldsymbol{A}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{A}^{\mathrm{T}})+\gamma\|\boldsymbol{A}\|_{2,1} \tag{7}$$

式中:λ,γ 都为系数,一般可以根据实际情况取经验值,最后从式(7)解出最优解 $\boldsymbol{A}$,得到对应最大非零特征值的特征向量。然而通过特征变换得到的数据降维结果,利用全谱数据信息进行特征变化,不仅增加了运算量,依旧无法得到特征谱段。为了实现基于特征谱段的高准确率分类,解决光谱特征峰高度混叠难以进行特征选择的问题,本文从模式识别的分类模型入手,通过自适应的谱段筛选算法,最终达到鲁棒、稳定的分类识别。

2 光谱数据降维处理方法

2.1 光谱数据预处理

光谱数据预处理目的是滤除异常数据和噪声。使用马氏距离表示样品间的差异,对异常的光谱数据进行筛选[18]。如图 1 所示,黄色光谱曲线是需要过滤的异常数据,异常数据对于建模精度有很大的影响。光谱信息噪声主要包括背景信息、杂散光和电噪声。求解光谱曲线的一阶导数,突出变化趋势[19],去除噪声影响,减弱光谱特征谱段的混叠,并获得更高信噪比的数据。导数光谱获取方式包括直接插分和最小二乘拟合等方法,本文的一阶导曲线使用直接插分方法,利用光谱数据逐点相减得到,即有

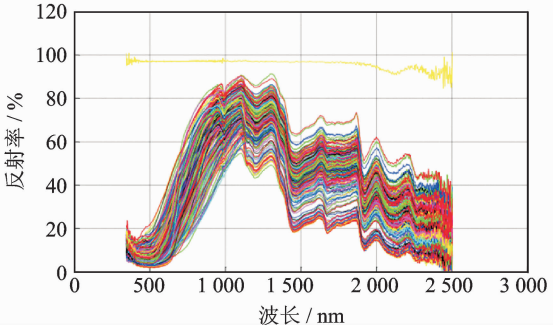

图 1 光谱中的异常数据

Fig. 1 Abnormal data in spectrum

$$\frac{\mathrm{d}x}{\mathrm{d}\lambda}=\frac{x_{i+1}-x_i}{2\Delta\lambda} \tag{8}$$

式中 x 为光谱 DN(Digital number)值。

2.2 特征谱段的数据挖掘

为了剔除冗余信息,本文提出一种自底向下序列前向选择的特征谱段自适应数据挖掘方法,通过量化各谱段变量的分类贡献度,采用"适者生存"的方式对变量进行筛选,对分类的贡献度越大,其权值越高,最终过滤出权值最高的谱段,视为剔除了冗余信息的最优变量组合,放入 SVM 模型的输入 $\boldsymbol{V}$ 中。具体而言,选择单独维度最优的谱段作为第 1 个特征变量,进而从备选集合中依次迭代选择,与前 $i-1$ ($i\geqslant 2$)个已选特征组合在一起后得到最优的第 i 个特征,即每次选取对分类贡献度最大的谱段加入特征集合中。迭代结束后,获得 M 个特征谱段变量子集。算法实现流程如图 2 所示。算法步骤为:(1) 设原始变量为 $\boldsymbol{S}$,当迭代次数 $i=1$ 时,计算 $\boldsymbol{S}$ 中各个变量的分类贡献度。(2) 挑选贡献度最大的变量,计作 $\boldsymbol{S}_i$ 并从原始变量中剔除,i 值自增 1。(3) 若 $1<i\leqslant M$,执行步骤(4),若 $i>M$,执行步骤(5)。(4) 计算 S 中各波长的分类贡献度,执行步骤(2)。(5) M 次迭代后,将最终结果$[S_1, S_2, S_3, \cdots, S_M]$定义为特征向量 $\boldsymbol{V}$,$\boldsymbol{V}$ 作为输入建立 SVM 模型。

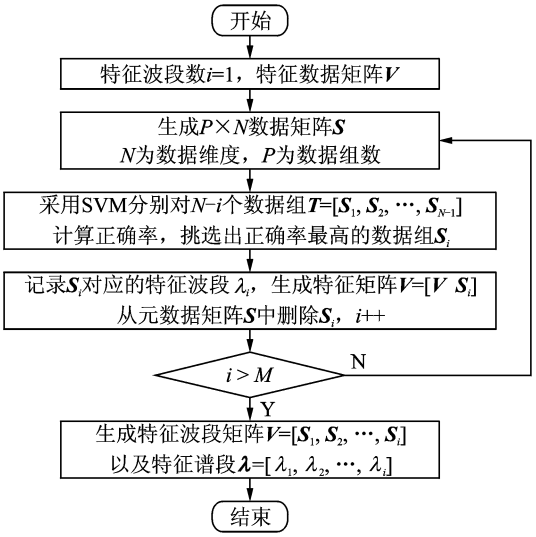

图 2 特征谱段的数据挖掘算法流程图

Fig. 2 Flow chart of data mining algorithm for feature spectrum

2.3 交叉验证

采用可以充分利用数据集的 k 折交叉验证(k-fold cross validation)对算法效果进行测试。将样本分为 k 份,从 k 份样本中取出一个样本作为测试集,剩下的 $k-1$ 份作为训练集设计分类器[22]。交叉验证 k 次,将这 k 次结果的平均值作为对算法精度的估计。k 取 10 时,即为 10 折交叉验证,它是一种常用的算法精度测试方法,能够获得准确的误差估计,因此本文采用 10 折交叉验证来评价和比较分类结果的准确率和鲁棒性。

2.4 SVM 分类

SVM[20,21]是一种优异的机器学习方法,能够实现全局最优的鲁棒分类。其将待解决的问题转化成一个二次规划的凸优化问题,在解决小样本和高维向量分类问题上表现出很多优势。对于两类分类的

问题,SVM 的思想就是寻找一个最优分类面将这两类分离,而距离任何样本缝隙最大的分类面就是要找的最优分类面。最优分类面方程为

$$\boldsymbol{w}\boldsymbol{x}+\boldsymbol{b}=0 \tag{9}$$

其中 $\boldsymbol{w}$ 和 $\boldsymbol{b}$ 未知,转换后的待求解的最优化问题为

$$\min \frac{1}{2}\|\boldsymbol{w}\|^2$$

$$\text{s.t. } y_i[(\boldsymbol{w}x_i)+\boldsymbol{b}]-1\geqslant 0 \quad i=1,2,\cdots,l \tag{10}$$

式中:l 为样本个数,求解得 $\boldsymbol{y}_i$ 和 $\boldsymbol{b}$,即得到最优分类面。对于一些在低维空间中线性不可分的情况,可以通过核函数将其映射到高维空间中使其变为线性可分问题。常用的核函数有多项式内积核函数、径向基核函数和 Sigmoid 内积核函数。多项式内积核函数为

$$K(\boldsymbol{x},x_i)=\left[\frac{1}{256}(\boldsymbol{x}\cdot x_i)\right]^q \tag{11}$$

径向基核函数为

$$K(\boldsymbol{x},x_i)=\exp\left\{-\frac{|\boldsymbol{x}-x_i|^2}{256\sigma^2}\right\} \tag{12}$$

Sigmoid 内积核函数为

$$K(\boldsymbol{x},x_i)=\tanh\left(\frac{\boldsymbol{b}(\boldsymbol{x}\cdot x_i)}{256}-c\right) \tag{13}$$

对于映射到高维空间中仍然线性不可分的情况,引入松弛变量,在容许少量错分的情况下实现对样本的线性分类。对于多类问题 SVM 有“1 对多”和“1 对 1”两种方法。“1 对多”的方法适用于测试样本出现类别重叠的情况,无法建立单一标签,“1 对 1”方法适用于样品具有独立标签的分类,因此在设计分类时采用“1 对 1”的方法。“1 对 1”对于 n 类样本,两两类别训练一个两类分类器,共训练 $n\times(n-1)/2$ 个分类器,对测试样本逐一进行测试并进行投票,得票最多的类别即为其类别标签。

2.5 分类方法比较

近邻法(K-nearest neighbor algorithm,KNN)[23]的基本思想是对于待测数据,与已知类别的数据逐一计算距离,根据统一的距离判别准则找到和已知样本最接近 k 个的数据,并且以该已知数据的类别作为待测数据的类别。具体表述为:样品集合 $S_N=\{(x_1,\theta_1),(x_2,\theta_2),\cdots,(x_N,\theta_N)\}$,$x_i$ 为第 i 个样本,θ_i 为第 i 个样本的类别,常见的距离判别准则有欧式距离和马氏距离,则

$$\delta(x,x')=\min_{j=1,\cdots,N}\delta(x,x^j) \tag{14}$$

那么 x 的类别就是 δ',这就是最近邻的决策思想。与 KNN 相比,结合了特征谱段挖掘的 SVM 分类方法,在光谱数据的多类分类问题中,性能略优,鲁棒性强。SVM 具有可控的模型训练过程,在光谱数据降维后做数据建模,可以获得稳定的分类模型,再用这个模型直接对新的样本测试集进行分类,因此适用于已经获得最大样本集情况下的分类问题,实现光谱数据降维的同时获得高分类准确率。而对于 KNN,通过将训练数据与训练数据进行距离度量来实现分类,没有数据的训练过程;另外,KNN 在样本不均匀的情况下,尤其是样本个数严重失调的情况下会对结果尤其是投票机制下的结果产生错的判断,因此选用时要避免样本的不均匀,对多类样品的训练集合还需要采取同数量的选择方式。神经网络算法[24,25]用于光谱数据样本的多类分类问题,训练不同样本的分类正确率随机性较强,模型鲁棒性较低,并且选用时需要调节和配置的参数较多,包括层数、神经元个数等,无法稳定获得全局最优,其模型分类准确程度不如 SVM 方法和 KNN 方法。

3 实验结果与分析

3.1 实验设备和实验样本

红木样本的光谱采集设备采用美国 ASD 公司的 FieldSpec 4 光谱仪，波长范围为 35～2 500 nm，光谱分辨率为 3 nm@700 nm，8 nm@1 400/2 100。每个红木样本使用光纤探头无损测量。实验以红木为样本，采集遍及 4 属 6 类的 11 种红木样本，每个红木样本约 100 处采样点，每采样点多次重复采集，取均值作为该样本采样点的光谱，共获取 1 150 条采样数据，样品编号如表 1 所示。

表 1 红木样品编号说明

Tab. 1 List of mahogany sample

红木编号	类型	种类	样品个数	采样点
1	紫檀木类	檀香紫檀	8	80
2	花梨木类	大果紫檀(缅甸)	8	100
3	花梨木类	大果紫檀(老挝)	7	100
4	香枝木类	降香黄檀	4	100
5	黑酸枝木类	伯利兹黄檀	12	110
6	红酸枝木类	巴里黄檀	4	100
7	红酸枝木类	交趾黄檀(泰国)	4	100
8	红酸枝木类	交趾黄檀(老挝)	15	110
9	红酸枝木类	微凹黄檀(尼加拉瓜)	11	100
10	红酸枝木类	微凹黄檀(巴拿马)	4	80
11	鸡翅木类	非洲崖豆木	8	100

3.2 数据预处理和特征谱段数据挖掘

红木采样的原始光谱数据如图 3 所示，红木样品的反射光谱特征具有相似的走势，难以直接区分。对原始数据求解一阶导数，结果如图 4 所示。显而易见，图 4 所示数据消除了采样间基线各异带来的影响，突出了数据的变化部分，更适合开展分类研究。将图 4 所示一阶导数曲线，在 350～2 500 nm 内，以 10 nm 作为一个区间进行划分，每 10 nm 作为一个谱段变量，迭代上限 M 取 20，特征谱段选择结果如图 5 所示。引入多分类贡献度，各个变量子集对分类的影响如图 6 所示。由图 6 可见，累积贡献度已近 90%，可代表光谱数据分类模型的主要特征。

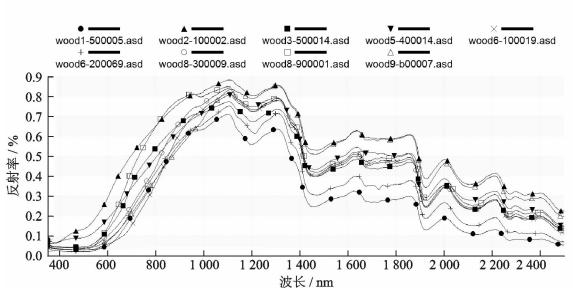

图 3 原始光谱曲线

Fig. 3 Raw spectra curve

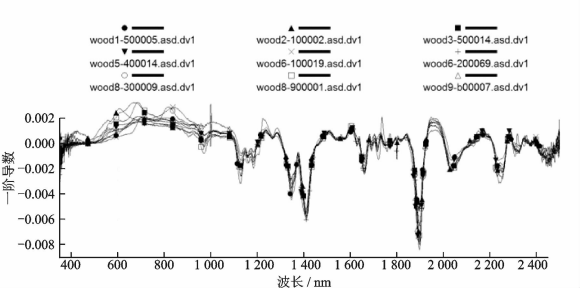

图 4 一阶导光谱曲线

Fig. 4 First-order spectrum curve

3.3 3 种多类分类器实验结果比较

使用 Matlab 中 Libsvm 工具包实现分类器，结果如表 2 所示。使用 SVM 对红木进行精确到产地的光谱数据分类，正确率为 98%。样品 2 和 3，7 和 8，9 和 10 为同一物种，将其归为一类进行精确到种类

的红木鉴别,分类准确率提高至99%。综上所述,红木分类问题使用了2 151维度中20个光谱维度,即0.93%的光谱数据进行SVM数据建模,实现11类样品的有效分类,解决红木精确分类问题效果突出,极具应用推广价值。表3给出了KNN分类结果,精确到产地的光谱数据分类正确率为93%,精确到种类的分类正确率为95%,与SVM分类结果对比,SVM比KNN分类准确率略高。神经网络精确到产地和种类的分类正确率仅分别为66.7%和60%,其模型分类准确程度较差,不具有应用价值。

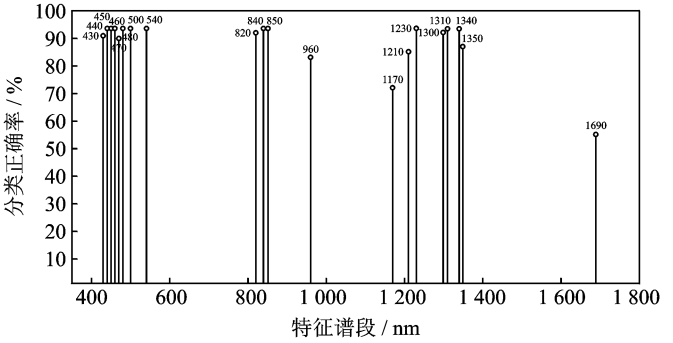

图5 特征谱段选择结果

Fig. 5 The characteristic band selection result

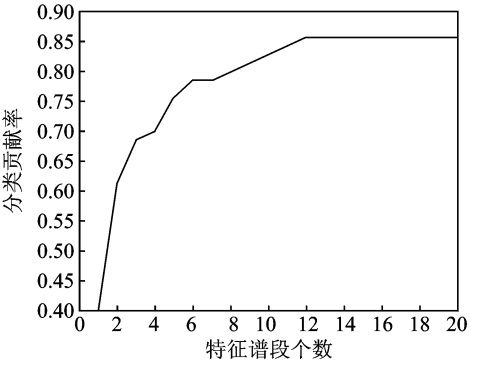

图6 特征谱段对分类贡献度的影响

Fig. 6 Influence of characteristic spectrum on classification contribution

表2 SVM分类结果(惩罚因子 $C=100$)

Tab. 2 Result of classification by SVM(penalty factor $C=100$)

编号	名称	产地鉴别准确率/%	种类	种类分类正确率/%
1	檀香紫檀	100	檀香紫檀	100
2	缅甸大果紫檀	100	大果紫檀	100
3	老挝大果紫檀	89		
4	降香黄檀	99	降香黄檀	98
5	伯利兹黄檀	99	伯利兹黄檀	100
6	巴里黄檀	97	巴里黄檀	100
7	交趾黄檀(泰)	94	交趾黄檀	95
8	交趾黄檀(老)	99		
9	微凹黄檀(尼)	99	微凹黄檀	99
10	微凹黄檀(巴)	100		
11	非洲崖豆木	100	非洲崖豆木	100
	综合	98	综合	99

表 3 KNN 分类结果

Tab. 3 The result of classification by KNN

编号	名称	精确到产地分类正确率/%	种类	种类分类正确率/%
1	檀香紫檀	94	檀香紫檀	94
2	缅甸大果紫檀	94	大果紫檀	94
3	老挝大果紫檀	87		
4	降香黄檀	90	降香黄檀	94
5	伯利兹黄檀	98	伯利兹黄檀	98
6	巴里黄檀	93	巴里黄檀	93
7	交趾黄檀(泰)	93	交趾黄檀	99
8	交趾黄檀(老)	92		
9	微凹黄檀(尼)	93	微凹黄檀	96
10	微凹黄檀(巴)	100		
11	非洲崖豆木	100	非洲崖豆木	100
	综合	93	综合	95

4 结束语

本文提出的用于多类分类的光谱数据解耦方法,利用光谱特征数据挖掘进行数据降维,经实践检验,可自适应的实现高维光谱数据"去繁存精",在对光谱数据降维的同时,实现了高精度的数据分类。将光谱特征峰高度混叠条件下的高光谱分类难题化简为易求解的多光谱问题,可推广至针对特定应用的低复杂度、小型化特定光谱仪器研制领域。

参考文献:

[1] Cao X, Du H, Tong X, et al. A prism-mask system for multispectral video acquisition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(12): 2423-2435.

[2] Ma C, Cao X, Tong X, et al. Acquisition of high spatial and spectral resolution video with a hybrid camera system[J]. International Journal of Computer Vision, 2014, 110(2): 141-155.

[3] Lin X, Liu Y, Wu J, et al. Spatial-spectral encoded compressive hyperspectral imaging[J]. ACM Transactions on Graphics (TOG), 2014, 33(6): 233.

[4] Bian L, Suo J, Situ G, et al. Multispectral imaging using a single bucket detector[J]. Scientific Reports, 2016, 6: 24752.

[5] Lin X, Wetzstein G, Liu Y, et al. Dual-coded compressive hyperspectral imaging[J]. Optics Letters, 2014, 39(7): 2044-2047.

[6] Wang Y, Suo J, Fan J, et al. Hyperspectral computational ghost imaging via temporal multiplexing[J]. IEEE Photonics Technology Letters, 2016, 28(3): 288-291.

[7] Suo J, Bian L, Chen F, et al. Bispectral coding: Compressive and high-quality acquisition of fluorescence and reflectance[J]. Optics Express, 2014, 22(2): 1697-1712.

[8] Cao X, Yue T, Lin X, et al. Computational snapshot multispectral cameras: Toward dynamic capture of the spectral world [J]. IEEE Signal Processing Magazine, 2016, 33(5): 95-108.

[9] An D, Suo J, Wang H, et al. Illumination estimation from specular highlight in a multi-spectral image[J]. Optics Express, 2015, 23(13): 17008-17023.

[10] Gao Y, Ji R, Cui P, et al. Hyperspectral image classification through bilayer graph-based learning[J]. IEEE Transactions on Image Processing, 2014, 23(7): 2769-2778.

[11] Das A J, Wahi A, Kothari I, et al. Ultra-portable, wireless smartphone spectrometer for rapid, non-destructive testing of fruit ripeness[J]. Scientific Reports, 2016, 6: 32504.

[12] Guo Z, Zhao J, Chen Q, et al. Application of selecting wavelength regions to determination of free amino acid content in tea by FT-NIR spectroscopy [J]. Optics and Precision Engineering, 2009, 8: 12.

[13] Liu G H, Xia R S, Jiang H, et al. A wavelength selection approach of near infrared spectra based on SCARS strategy and its application[J]. Guang pu xue yu guang pu fen xi= Guang pu, 2014, 34(8): 2094-2097.

[14] 张学工. 模式识别[M]. 北京:清华大学出版社,2010:154-156.
Zhang Xuegong. Pattern recognition[M]. Beijing: Tsinghua University Press, 2010: 198-205.

[15] 张小超.近红外光谱分析技术及其在现代农业中的应用[M].北京:电子工业出版社,2012.122-132.
Zhang Xiaochao. Near infrared spectroscopy technology and its application in modern agriculture[M]. Beijing: Publishing House of Electronics Industry, 2012:122-132.
[16] Jia X, Richards J A. Segmented principal components transformation for efficient hyperspectral remote-sensing image display and classification[J]. Geoscience and Remote Sensing, IEEE Transactions on, 1999, 37(1): 538-542.
[17] Yang Z S, Guo L, Luo X, et al. Research on segmented PCA based on ban selection algorithm of hyperspectral image[J]. Engineering of Surveying and Mapping, 2006, 15(3): 15-18.
[18] 陈斌,邹贤勇,朱文静. PCA结合马氏距离法剔除近红外异常样品[J]. 江苏大学学报:自然科学版,2008,29(4):277-279.
Chen Bin, Zou Xianyong, Zhu Wenjing. Eliminating outlier samples in near-infrared model by method of PCA-Mahalanobis distance[J]. Journal of Jiangsu University: Natural Science Edition, 2008, 29(4): 277-279.
[19] 王学顺,戚大伟,黄安民.木材近红外光谱小波阈值去噪方法[J]. 东北林业大学学报,2009,37(2):32-34.
Wang Xueshun, Qi Dawei, Huang Anmin. Threshold denoising of near infrared spectroscopy of wood based on wavelet transform [J]. Journal of Northeast Forestry University, 2009, 37(2): 32-34.
[20] 苏高利,邓芳萍. 关于支持向量回归机的模型选择[J]. 科技通报,2006,22(2):154-158.
Su Gaoli, Deng Fangping. Introduction to model selection of SVM regression [J]. Bulletin of Science and Technology, 2006, 22(2): 154-158.
[21] Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods[M]. Cambridge: Cambridge University Press, 2000,8-149.
[22] 冯进玫,卢志茂,陈纯锴.一种基于均值更新的分类模型[J]. 计算机系统应用,2012,21(8):123-126.
Feng Jingong, Lu Zhimao, Chen Chunkai. Classification model based on the mean update [J]. Computer Systems and Applications, 2012, 21(8): 123-126.
[23] Zhang M L, Zhou Z H. ML-KNN: A lazy learning approach to multi-label learning[J]. Pattern Recognition, 2007, 40(7): 2038-2048.
[24] Malliani A, Lombardi F, Pagani M. Power spectrum analysis of heart rate variability: A tool to explore neural regulatory mechanisms[J]. British Heart Journal, 1994, 71(1): 1.
[25] Böhm G, Muhr R, Jaenicke R. Quantitative analysis of protein far UV circular dichroism spectra by neural networks[J]. Protein engineering, 1992, 5(3): 191-195.

作者简介:

戴琼海(1964-),男,教授,博士生导师,研究方向:计算摄像学,E-mail:qhdai@tsinghua.edu.cn。

张晶(1986-),女,工程师,研究方向:光谱数据分析、模式识别。

李菲菲(1990-),女,硕士研究生,研究方向:光谱数据采集。

范静涛(1979-),男,讲师,博士,研究方向:计算摄像理论与技术。