基于聚类的烟叶标准库和特征选择

李　航,申金媛,刘润杰,穆晓敏
(郑州大学 信息工程学院,河南 郑州 450001)

摘要:为实现烟叶自动化分级,采用聚类算法来剔除烟叶样本中异样样本,通过计算类间的方差,选取方差值大的特征作为有用特征。利用支持向量机模拟人工操作,进行13个等级分部位、分颜色的识别。结果表明:特征选择后的最佳准确率分别为95.52%、98.22%。在提升准确率的同时,减少了输入特征的个数和采集光谱数据所需的时间。

关键词:烟叶;光谱;聚类分析;特征提取
中图分类号:S572　**文献标识码**:A　**文章编号**:1002-2767(2016)04-0140-04　DOI:10.11942/j.issn1002-2767.2016.04.0140

烟叶是我国主要经济作物之一,烟叶等级的划分影响成品烟的质量。目前人工分级主要依据烟叶的生长部位和颜色确定其等级,存在费时费力、主观性强、分级吻合率不稳定的缺点[1]。很多学者提取烟叶图像[2-3]和光谱[4-5]特征进行分级,图像特征不能反映烟叶内部化学成分含量。由于获取光谱维数高,降维的方法有小波分解[6]、主成分分析[7]、粒子群法[8]、聚类法[9]。本文依据物以类聚思想剔除13类中异样样本,并选取类间方差大的特征为有用特征,采用对小样本、高维数据分类效果好的支持向量机模型进行分部位、分颜色。

1　材料与方法

1.1　材料

郑州市烟草公司提供的13类样本分别为B2F、B3F、B4F、C2F、C3F、X2F、X3F、X4F、C2L、C3L、X2L、X3L、X4L,共计642片。其中B、C、X分别表示烟叶的生长部位为上、中、下,F、L分别表示烟叶的颜色为橘黄、柠檬黄。由岛津UV3600型号仪器获取烟叶1 500～2 400 nm波长(间隔2 nm)反射光谱数据,仿真平台为MATLAB R2009a。

1.2　方法

为消除由于基线漂移带来的误差,对获得的光谱数据进行如下预处理:

$$y_i = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)}$$

其中 x_i 为预处理前的数据,y_i 为预处理后的数据。

由于提供的样本当中可能存在异样样本,数据预处理后对各类进行系统聚类分析。每个样本先自成一类,规定类与类之间的距离为相关系数,选择距离最小的两类合并为一类。计算新类与其它类的距离,再将距离最小的合为一类,直至所有样本都归为一类。聚类后剔除同一类当中那些与大部分样本差异性大的样本,余下的样本选取近一半为标准库(训练集),其余为测试集。相关系数计算公式为:

$$d_{rs} = 1 - \frac{(x_r - \bar{x}_r)(x_s - \bar{x}_s)'}{\left[(x_r - \bar{x}_r)(x_r - \bar{x}_r)'\right]^{\frac{1}{2}} \left[(x_s - \bar{x}_s)(x_s - \bar{x}_s)'\right]^{\frac{1}{2}}}$$

其中 $\bar{x}_r = \sum_j x_{rj}$,$\bar{x}_s = \frac{1}{n}\sum_j x_{sj}$

计算每类样本在各个特征的均值,用均值代表其类别属性。求取13类均值在各个特征上的方差,选取大于不同方差值所对应的特征进行烟叶分部位、分颜色。

采用投票式支持向量机进行部位和颜色的判别。每两个类构建一个支持向量机分类器,M个类别需构建M(M-1)/2。模型构建好后,对于未知类别样本用所有支持向量机进行判别,统计各个支持向量机对该样本的判别,将其归为得票最多的那一类。本文采用线性核函数实现向量向高维的映射,判决函数为:

收稿日期:2016-01-28
基金项目:河南省烟草公司科技计划资助项目(No.M201335)
第一作者简介:李航(1989-),男,河南省开封市人,在读硕士,从事模式识别、光谱分析研究。E-mail:lh121144@163.com。
通讯作者:申金媛(1966-),女,山西省晋中市人,博士,教授,从事模式识别、光谱分析、数据挖掘研究。

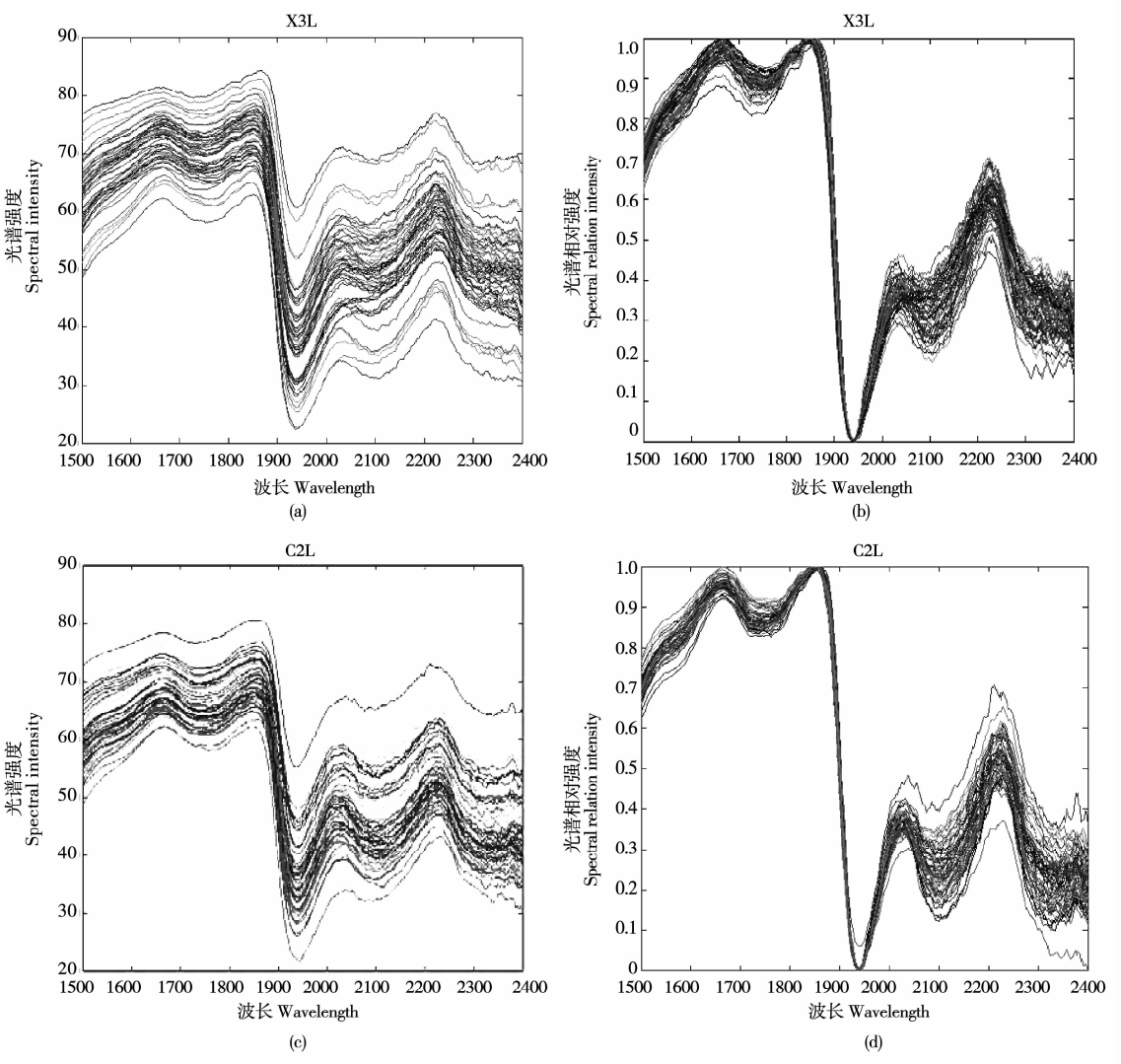
$$g(x) = \mathrm{sgn}[\sum_{i=1}^{n}\alpha_i d_i K(x_i, x) + b]$$

其中 $K(x_i, x)$ 为线性核函数以实现输入样本的映射，x_i 为支持向量。d_i 取值为 1 或者 −1，对应为支持向量机的输出，x 为验证集。

2　结果与分析

2.1　数据预处理

由图 1 可知，减最值处理可有效的消除基线漂移，降低噪声的影响。



a、c为 X3L和C2L级别原始光谱，b、d为X3L和C2L预处理后的光谱

a and c are the original spectrum of X3L and C2L,b and d are the pretreatment spectrum of X3L and C2L

图 1　烟叶 X3L、C2L 光谱数据预处理前后对比

Fig. 1　Comparison diagram of pretreatment spectrum and original spectrum of X3L and C2L

2.2　聚类结果与分析

由图 2 可知，X3L 聚类效果较好，C2L 中相关系数大于 5×10^{-3} 的可能为异样样本，进行分部位、分颜色时将其剔除。

2.3　类间方差分析

通过比较 13 类的均值属性，可以看出各类均值在某些特征上差异性比较大(见图 3)。计算均值在各个特征上的方差，选取方差大的特征作为有用特征。

2.4　分部位、分颜色

通过改变不同方差阈值选取有用特征，将其作为支持向量机的输入。分部位、分颜色吻合率随方差阈值变化见图 4。随输入特征数目的变化，分部位、分颜色吻合率均呈现先增加再减少的

趋势。分部位和分颜色的测试集正确率最大值分别为 95.52%和 98.22%，此时对应的方差阈值分别为 3.4×10^{-3}和 2.8×10^{-3}，统计图 3 中纵坐标大于该阈值和特征个数，分部位和分颜色对应的特征数目分别为 275 个和 383 个，比全光谱 451 个特征有明显减少，选取有用特征可减少光谱数据的采集，同时识别率也有一定的提高。

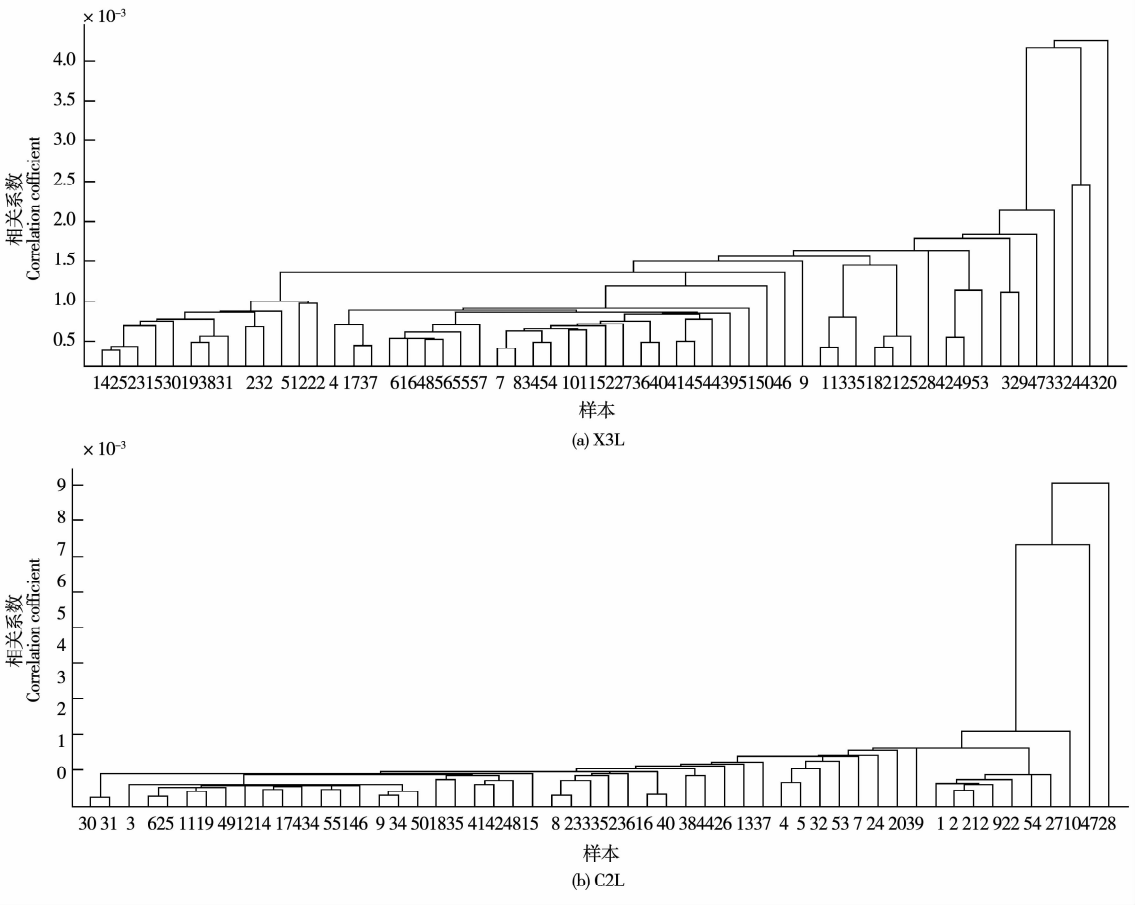


图 2　X3L 和 C2L 级别的聚类结果

Fig. 2　The clustering results of X3L and C2L

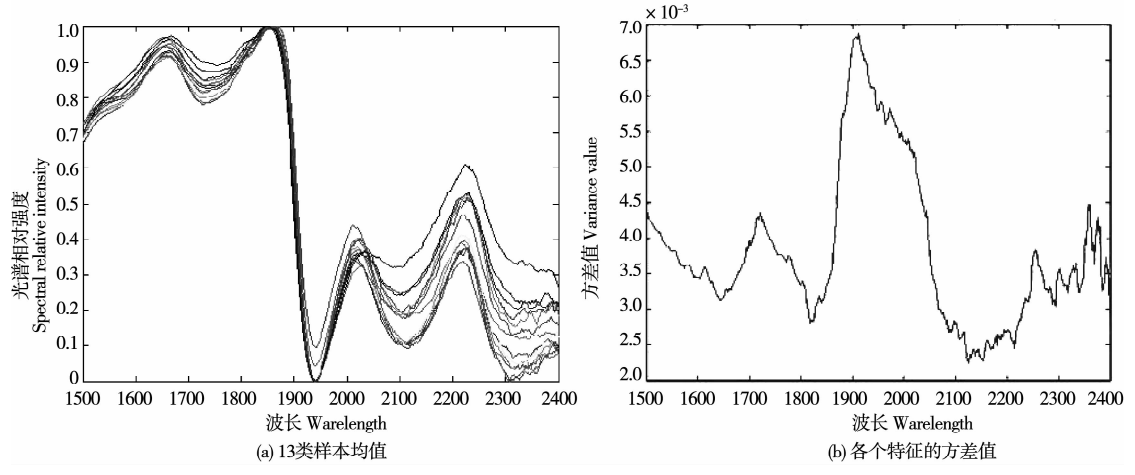


图 3　样本均值和特征方差值

Fig. 3　The sample mean and the characteristics of variance values

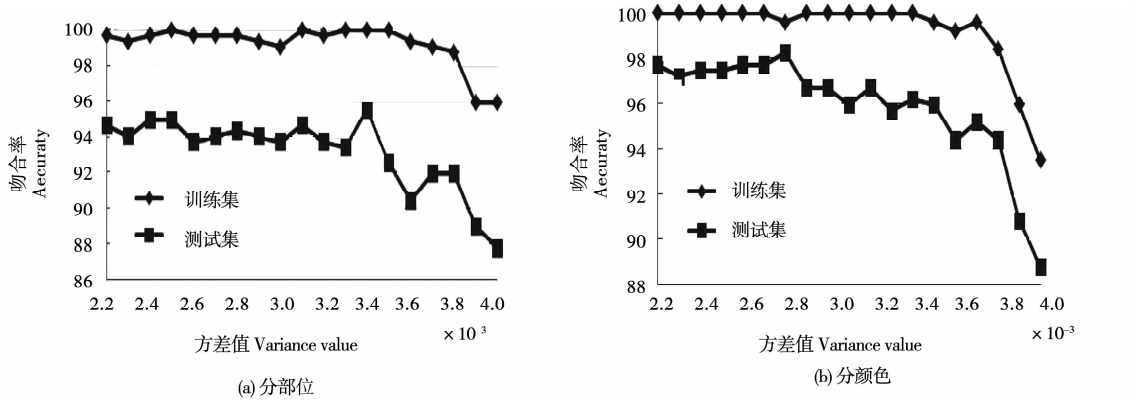


图 4　分部位、分颜色吻合率与方差值的关系曲线

Fig. 4　The curve of relationship between the accuracy and variance value of division and color

3　结论

为降低基线漂移的影响,减最值处理是常用的数据处理方法。由于获得的烟叶样本中可能含有分错类别的样本,对各类进行系统聚类分析,剔除与大部分样本差异性较大的样本是可行的,余下的样本集作为该类的样本库选取类间差别大的特征作为支持向量机的输入。

对 13 等级烟叶数据进行减最值预处理,聚类分析剔除异样样本,由于所采集烟叶的光谱特征中可能含有冗余特征,删减冗余特征不仅可以减少光谱的采集,而且可以提高分级正确率。提取类间差异性大的特征作为有用特征,选用支持向量机模型进行分部位、分颜色。特征数目得以减少且吻合率有一定的提高,为实现自动化烟叶分级奠定了理论基础。

参考文献:

[1]　马文杰. 烟叶图像采集技术规范与烤烟收购质量分级特征研究[D]. 武汉:华中农业大学,2007.

[2]　张惠民,韩力群,段正刚. 基于图像特征的烟叶分级[J]. 武汉大学学报,2003,28(3):359-362.

[3]　Tatters filed G,Forbes K. Classification of tobacco leaves by color and plant position[C]//Proceedings of theNinth Annual Symposium of the Pattern Recognition Association of South Africa ,Cape Town,1998:11.

[4]　王毅,马翔,温亚东,等. 应用近红外光谱分析不同年度工业分级烟叶的特性[J]. 光谱学与光谱分析,2012,32(11):3014-3018.

[5]　Hongmei Li,Jiande Wu,Kuake Huang,et al. Study of fluecured tobacco classification model based on the PSO-SVM[J]. Research Journal of Applied Sciences,Engineering and Technology,2013,5(19):4671-4676.

[6]　田高友,袁洪福,刘慧颖,等. 小波变换在近红外光谱分析中的应用进展[J]. 光谱学与光谱分析,2003,23(6):1111-1114.

[7]　彭丹青,申金媛,刘剑君,等. 基于径向基网络的烟叶光谱分级[J]. 农机化研究,2009,53(10):15-18.

[8]　李航,赵海东,申金媛,等. 基于 BPSO 和 SVM 的烟叶近红外有用特征光谱选择[J]. 物理实验,2015,35(6):8-12.

[9]　赵海东,申金媛,刘润杰,等. 基于聚类的烟叶近红外光谱有效特征的筛选方法[J]. 红外技术,2013,35(10):659-664.

Tobacco Standard Library and Feature Selection Based on the Cluster

LI Hang, SHEN Jin-yuan, LIU Run-jie, MU Xiao-min

(School of Information Engineering, Zhengzhou University, Zhengzhou, Henan 450001)

Abstract: There are possibilities that the different samples may exist among the tobacco left samples, so the clustering method is applied to get rid of the different samples. By calculating the variance between the classes, the characteristics of big variance value as a useful feature were selected. Using SVM to simulate the artificial operation, like the recognition of the parts and the colors of thirteen grades of the tobacco left samples. The optima accuracy reached 95.52% and 98.22% respectively after the select of the useful features. Meanwhile, not only the accuracy arising, but the number of input features and the time needed for spectral data were reduced.

Keywords: tobacco leaves; spectrum; clustering analysis; feature extraction