文章编号:1000-2367(2021)01-0045-05 DOI:10.16366/j.cnki.1000-2367.2021.01.007

基于提取不同中红外光谱特征信息的烟叶部位判别研究

赵娟娟[1a],叶顺[2],徐可[1b],陈栋骅[2],岳宝华[1a],李敏杰[1a],刘太昂[1a],陆文聪[1a]

(1.上海大学 a.化学系;b.电子信息材料系,上海 200444;2.上海烟草集团有限责任公司 技术中心,上海 200082)

摘 要:中红外光谱(MIR)分析技术在烟草中有广泛的应用,利用中红外分析可以获取烟草中大量化学信息.为了提高谱图的信噪比,需要对谱图数据进行预处理.研究发现对烟叶中红外光谱数据进行一阶导数结合 Savizky-Golay 的预处理,不仅提高了信噪比,而且增加了烟叶部位分类判别的准确率.另外,对谱图数据进行降维处理,有利于提取中红外谱图信息,减少冗余数据,减少计算时间.本文对比了基于原始中红外谱图数据、连续投影算法(SPA)特征提取后数据、偏最小二乘法(PLS)降维特征提取后数据的烟叶部位分类判别准确率,结果表明 PLS 降维特征提取可以有效提取烟叶中红外光谱数据的特征信息,有利于烟叶部位分类判别准确率的提高.利用 PLS 提取烟叶中红外特征信息数据建立的烟叶部位支持向量机(SVM)分类判别模型,其建模、留一法和独立测试集的准确率分别为:96.00%、89.60%和 80.65%.

关键词:中红外光谱;连续投影算法;支持向量机;烟叶部位

中图分类号:O69 **文献标志码**:A

红外光谱包括近红外光谱(Near-infrared spectroscopy,NIR)和中红外光谱(Mid-infrared spectroscopy,MIR),可以作为一种检测物质中含有 C-H、N-H 等化学键的分析工具[1].光谱技术在物质组分分析方面有着快速、简单、灵敏等优势,在食品[2]、烟草[3]和土壤[4]等方面得到广泛应用.中红外光谱能呈现有机物的分子震动信息,具有吸收峰窄,谱峰重叠不严重、信息量较大、信息提取更容易、样品信息表达更丰富、分子选择性更好等优点[5],因此,中红外光谱分析技术在烟草行业越来越受到重视.李沅等[6]开发设计了基于烟草主流烟气红外光谱的氨浓度检测系统,首先由红外探测器得到主流烟气的红外光谱,再结合光谱分析算法与光谱数据库,通过比尔朗伯定律计算主流烟气中的氨浓度.吴舜等[7]利用中红外光谱技术分别测定了烟梗粉和烟叶粉的木质素含量.TERPUGOV E L 等[8]利用傅里叶变换红外光谱对烟叶进行了无损检测研究.

在进行中红外光谱分析时分辨率越高,所包含的样本化学成分信息越多,但也提高了数据采集时间和成本,同时也不可避免地会包含很多与目标变量无关的光谱信息,因此,在利用中红外光谱分析技术时,进行光谱数据预处理和光谱特征信息提取是十分必要的[9-10].本文基于烟叶萃取液的中红外光谱谱图,分析对比了不同预处理方法和不同特征提取方法,希望挖掘出有效提取中红外光谱数据特征信息方法,为中红外光谱分析技术在烟叶中的应用提供支持,为烟叶质量管理和部位识别提供参考.

1 实验和算法

1.1 中红外光谱

实验选取 156 个具有代表性的烟叶样本,其中包含 53 个上部烟叶样品,71 个中部烟叶样品,32 个下部烟叶样品.在试管中放置 1 g 烘干后的烟叶粉末,加入正己烷 10 mL,然后用超声萃取 30 min,静置 30 min 后,用 0.22 μm 滤膜过滤至小试管中,取 5 mL 至静置挥发 3 d,获得烟草萃取液样品.利用 ThermoFisher 公

收稿日期:2020-05-17;**修回日期**:2020-12-01.

基金项目:国家自然科学基金青年基金(21706156);卷烟烟气重点实验室开放性课题(K2018-1-056P).

作者简介:赵娟娟(1996-),女,安徽安庆人,上海大学硕士研究生,主要从事数据挖掘工作,E-mail:3055182093@qq.com.

通信作者:刘太昂,博士,主要从事近红外光谱技术研究,E-mail:athincat@163.com.

司的 Nicolet iS50 傅立叶变换红外光谱仪扫描每个正己烷萃取后的烟叶萃取液样品,得到不同烟叶样本的中红外光谱谱图,每个样品扫描 4 次,求其平均光谱作为样品的中红外光谱.图 1 是不同部位烟叶萃取液样品的中红外光谱.

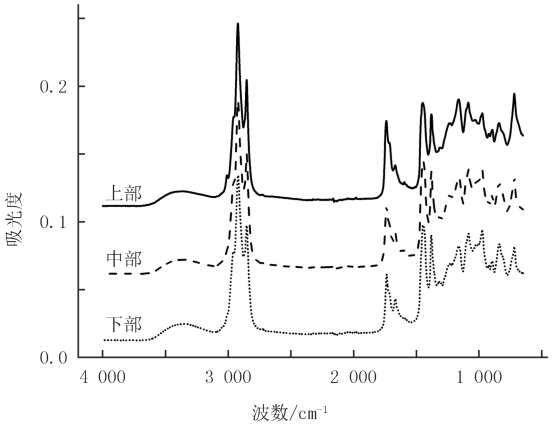

图1 不同部位烟叶中红外光谱谱图

Fig. 1 MIRS of different tobacco leaves parts

1.2 光谱的预处理

由于光谱采集的质量会受到多种因素的影响,如仪器、环境、样品浓度不均匀等,因此需要对中红外谱图数据进行预处理来尽可能地消除各种影响因素带来的测量、预测误差,以提高光谱信噪比.本文采用的预处理方法包括一阶导数、二阶导数和 Savitsky-Golay(S-G)平滑.图 2(a)是一阶导数结合 Savitsky-Golay 平滑预处理图,图 2(b)是二阶导数结合 S-G 平滑预处理图.

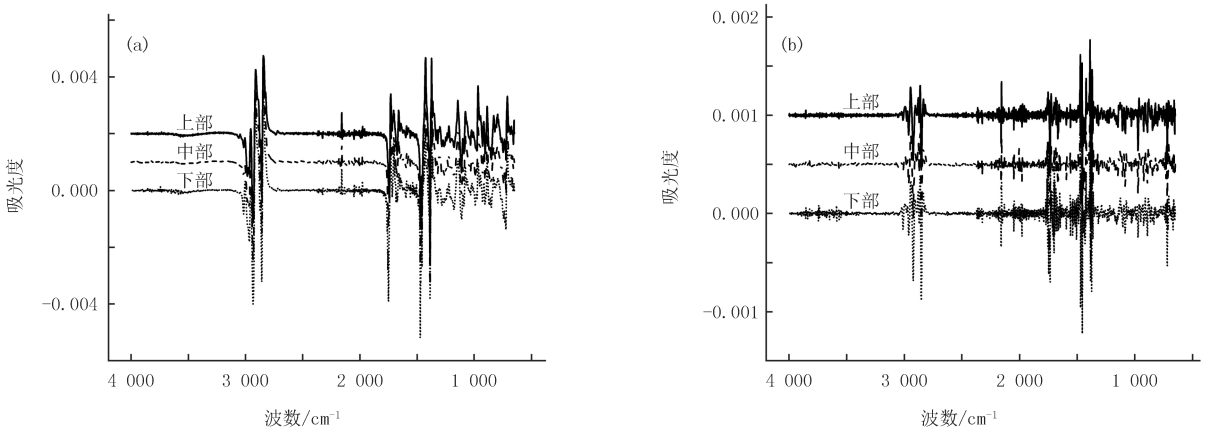

图2 一阶导数+S-G平滑(a)和二阶导数+S-G平滑(b)

Fig. 2 First derivative +S-G(a) VS Second derivative +S-G(b)

1.3 算 法

1.3.1 连续投影算法

连续投影算法(Successive Projection Algorithm,SPA)[11-12]可以从中红外光谱数据中提取有效的波长点信息,提取出变量之间共冗余信息最小的变量组合,这样最大程度上减少建模所需的变量数量,从而提高建模的效率.SPA 算法计算有两个阶段.第一个阶段对中红外光谱矩阵进行投影产生 K 个集合,且每个集合有 M 个变量.在这 K 个集合中,被挑选出来的每一个特征与之前的特征之间的线性关系要最小.第二个阶段,SPA 可以通过不同的预测性能指标来评估第一阶段 K 个集合中的变量的候选子集,本文以预报准确率的最大值所对应的子集即为最优特征波长集.

1.3.2 偏最小二乘法

偏最小二乘法(Partial Least Squares,PLS)为统计学中的进行数据分析的常用方法之一,其方法主要是用来找到多因变量数据与多自变量数据之间的统计关系,从而建立回归模型.PLS 是空间变换的主要数学方法之一.在低维的 PLS 空间,进行模式识别可以用来对特征众多的中红外光谱数据进行降维.偏最小二乘降维是一种非常有效的数据降维方法,广泛地应用于光谱数据分析领域[13-14].

1.3.3 支持向量机

VAPNIK[15]提出的支持向量机(Support Vector Machine，SVM)的主要思想是在特征空间上，构造出最优超平面，并要求该超平面与不同类样本集之间的距离间隔最大，以便达到其有最大的泛化能力.支持向量机算法将数据映射到高维特征空间中，将问题转化为低维输入空间上的一个简单的函数计算，解决了原始空间中数据线性不可分的问题，并且通过解决一个二次规划问题，获得全局最优.由于支持向量机的计算复杂性与其计算和输入数据的维数并不直接相关，而和选择的支持向量样本数有关，可以很好地解决小样本、非线性数据集的分类问题[15-16].

2 结果与讨论

2.1 预处理方法的选择

随机提取 156 个样本的 20%即 31 个样本作为预报集，剩余的 80%即 125 个样本作为建模集.表 1 对比了利用原始谱图、一阶导数结合 S-G 平滑和二阶导数结合 S-G 平滑预处理的烟叶部位 SVM 分类判别模型的建模、留一法和预报准确率.结果表明：对烟叶中红外光谱数据进行一阶导数结合 S-G 平滑的预处理，可以提高烟叶部位 SVM 分类判别模型的准确率.

表 1 不同预处理方法建模性能的比较

Tab. 1 The accuracies of different preprocessing methods

预处理方法	建模准确率/%	留一法准确率/%	预报准确率/%
原谱图＋PLS 降维＋SVM	96.80	85.60	61.29
一阶导数＋S-G 平滑＋PLS 降维＋SVM	96.00	89.60	80.65
二阶导数＋S-G 平滑＋PLS 降维＋SVM	95.20	88.00	70.97

2.2 特征信息提取

由图 1 和图 2 可以看出无论是烟草原始的中红外光谱或者求导平滑预处理后的数据都存在数据量大，信息冗余度高等特点，其存储所需要的空间大，处理时间长，波段波数多，容易出现维数灾难现象，因此中红外谱图数据的特征信息提取是非常重要的一个环节.本文主要采用 SPA 和 PLS 降维对中红外光谱进行降维、特征提取.

2.2.1 SPA 特征提取

SPA 提取了 36 个特征信息，对应的特征波数分别为 668.213 7、687.498 4、706.783 1、726.067 7、745.352 4、803.206 4、822.491 1、841.775 8、861.060 4、899.629 8、918.914 4、957.483 8、976.768 4、996.053 1、1 015.338 0、1 034.622 0、1 053.907 0、1 073.192 0、1 092.476 0、1 111.761 0、1 150.330 0、1 169.615 0、1 188.900 0、1 208.184 0、1 227.469 0、1 246.754 0、1 381.746 0、1 439.600 0、1 458.885 0、2 847.381 0、2 866.666 0、2 885.950 0、2 905.235 0、2 924.520 0、2 943.804 0 和 2 963.089 0.结果如图 3 所示.

2.2.2 PLS 降维后提取特征

PLS 降维后，以烟叶部位 SVM 分类判别模型的预报准确率为标准进行 PLS 因子个数选择，当选择 12 个PLS 因子时，模型的预报准确率可以达到 80.66%.结果如图 4 所示.

2.3 不同模型结果比较

由表 2 对比了利用全部原始谱图、SPA 特征提取、PLS 降维特征提取后对应的烟叶部位 SVM 分类判别模型的准确率.可以看出：原始谱图数据有 3 475 个特征，包含的信息量最大，虽然建模准确率达到了 100.00%，但其留一法和预报准确率都非常低，可能是在 3 475 个特征峰中也包含了噪声信息，模型出现了过拟合问题，因此留一法和预报结果差.SPA 算法在进行特征提取时只选择了 36 个特征，虽然大大减少了特征信息的数目，有效地简化了模型，但其预报准确率也较低，这是因为 SPA 虽然挖掘出最低冗余度信息，但也损失了比较多的有效信息，因此其泛化能力较差.利用 PLS 降维提取特征，其建立的烟叶部位 SVM 分类判别模型的建模、留一法和预报准确率分别为：96.80%、89.60%和 80.66%，明显高于基于全部原始谱图和 SPA 特征提取后的两个模型.一方面原因是虽然只选择了 12 个 PLS 因子，但每个 PLS 因子都是一阶导数结

合 S-G 平滑后全部中红外光谱数据的线性组合,因此保留的信息量足够大;另一方面原因是 PLS 降维后又只选择前 12 个 PLS 因子,这又删除了对模型泛化能力影响小的冗余信息,因此 PLS 降维特征提取方式无论是建模、留一法或是预报准确率都比较高.

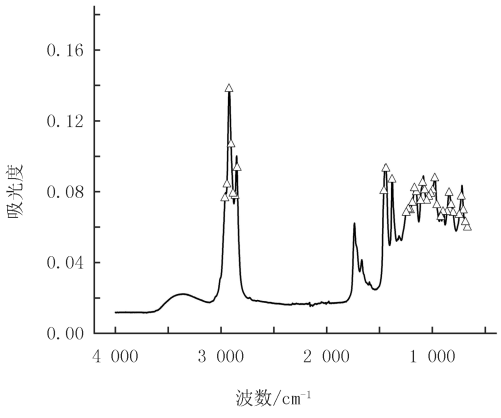

图3 SPA特征选择图

Fig.3 Feature selection of SPA

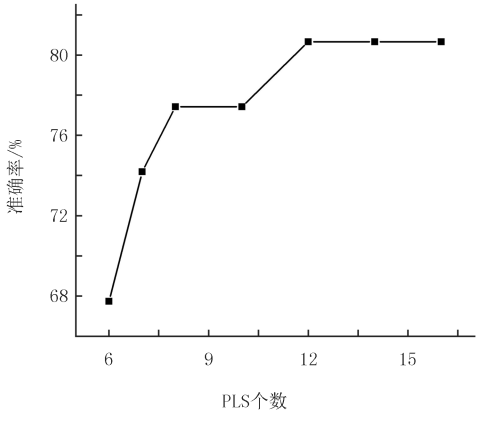

图4 PLS因子个数选择

Fig.4 Selection PLS numbers

表 2 不同建模方法的准确率比较

Tab. 2 The accuracies of different methods

建模方法	特征个数	建模准确率/%	留一法准确率/%	预报准确率/%
全部原始谱图+SVM	3 475	100.00	71.12	64.52
SPA+SVM	36	92.80	75.20	70.97
PLS 降维+SVM	12	96.00	89.60	80.65

3 结 论

通过对比了不同预处理方法对中红外光谱信噪比的影响,发现对烟叶中红外光谱数据进行一阶导数结合 S-G 平滑的预处理,可以提高信噪比,有利于烟叶部位分类判别准确率的提高.通过对比不同特征提取方法发现利用 PLS 降维特征提取方法可以有效提取烟叶中红外光谱数据的特征信息,有利于烟叶部位分类判据准确率的提高.这为中红外光谱分析技术在烟叶中的应用提供信息支持,为烟叶质量管理和部位识别提供参考.

参 考 文 献

[1] 何勇,郑启帅,张初,等.基于中红外光谱和化学计量学算法鉴别核桃产地及品种[J].光谱学与光谱分析,2019,39(9):2812-2817.
HE Y,ZHENG Q S,ZHANG C,et al.Identification of walnut origin and varieties based on mid-infrared spectroscopy and chemometrics algorithm[J].Spectroscopy and Spectral Analysis,2019,39(9):2812-2817.

[2] SUEHARA K,KAMEOKA T,HASHIMOTO A.Sugar uptake analysis of suspension Arabidopsis,tobacco,and rice cells in various media using an FT-IR/ATR method[J].Bioprocess and Biosystems Engineering,2012,35:1259-1268.

[3] LIU T A,ZHANG Q,CHANG D P,et al.Characterization of tobacco leaves by near-infrared reflectance spectroscopy and electronic nose with support vector machine[J].Analytical Letters,2018,51(12):1935-1943.

[4] DUNNE K S,HOLDEN N M,O'ROURKE S M,et al.Prediction of phosphorus sorption indices and isotherm parameters in agricultural soils using mid-infrared spectroscopy[J].Geoderma,2020,358:1-9.

[5] 王佰华,李正章,王军华.中-近红外双光路光谱法快速测定汽油辛烷值[J].化学工程与装备,2015(2):170-171.
WANG B H,LI Z Z,WANG J H.Rapid determination of gasoline octane number by mid-near infrared dual-path spectroscopy[J].Chemical Engineering and Equipment,2015(2):170-171.

[6] 李沅,陈智刚,李凯.基于红外特征吸收法的烟草主流烟气中氨浓度检测系统[J].光谱学与光谱分析,2013,33(6):1521-1524.
LI Y,CHEN Z G,LI K.Ammonia concentration detection system for mainstream smoke of tobacco based on characteristic infrared absorption method[J].Spectroscopy and Spectral Analysis,2013,33(6):1521-1524.

[7] 吴舜，楼宏铭，莫贤科，等.中红外光谱法测定烟草中的木质素[J].烟草科技，2014，10：67-70.
WU S，LOU H M，MO X K，et al.Determination of lignin in tobacco by mid-infrared spectroscopy[J].Tobacco Technology，2014，10：67-70.

[8] TERPUGOV E L，DEGTYAREVA O V，SAVRANSKY V V.Possibility of light-induced mid-IR emission in situ analysis of plants[J].Journal of Russian Laser Research，2016，37(5)：507-510.

[9] 黄华，朱洁，刘广昊，等.近红外光谱多核并行谱区选择任务调度策略研究[J].农业机械学报，2018，49(10)：270-283.
HUANG H，ZHU J，LIU G H，et al.Task scheduling strategies of parallel near infrared spectral region selection on multi core and its application[J].Journal of Agricultural Machinery，2018，49(10)：270-283.

[10] 刘太昂.近红外和电子鼻数据融合及其在烟草中的应用[D].上海：上海大学，2018.
LIU T A.Data fusion of near infrared and electronic nose and its application in tobaccos[D].Shanghai：Shanghai University，2018.

[11] 李阳阳，孙雨安，王国庆，等.基于高光谱的大叶女贞叶片水分定量测定[J].河南师范大学学报(自然科学版)，2017，45(6)：47-51.
LI Y Y，SUN Y A，WANG G Q，et al.Determination of ligustrum leaf water content based on hyperspectral[J].Journal of Henan Normal University(Nature Science Edition)，2017，45(6)：47-51.

[12] ARAUJO M C U，SALDANHA T C B，GALVAO R K H，et al.The successive projections algorithm for variable selection in spectroscopic multicomponent analysis[J].Chemometrics and Intelligent Laboratory Systems，2001，57(2)：65-73.

[13] 鄢悦，张红光，卢建刚，等.基于光谱信息散度的近红外光谱局部偏最小二乘建模方法[J].计算机与应用化学，2017，34(5)：351-355.
YAN Y，ZHANG H G，LU J G，et al.Spectral-information-divergence based on local pls modeling algorithm of near infrared spectroscopy[J].Computer and Applied Chemistry，2017，34(5)：351-355.

[14] 张浩博，刘太昂，束茹欣，等.基于烟叶电子鼻-近红外数据融合的支持向量机分类判别烟叶年份[J].光谱学与光谱分析，2018，38(5)：1620-1625.
ZHANG H B，LIU T A，SHU R X，et al.Using EN-NIR with support vector machine for classification of producing year of tobacco[J].Spectroscopy and Spectral Analysis，2018，38(5)：1620-1625.

[15] VAPNIK.Statistical learning theory[M].New York：Wiley-Interscience，1998.

[16] 沙云菲，王亮，刘太昂，等.基于筛选后主要化学成分对同类植物的品种分类研究[J].计算机与应用化学，2019，36(5)：491-496.
SHA Y F，WANG L，LIU T A，et al.Classification for the varieties of single plant based on important chemical compositions by feature selections[J].Computer and Applied Chemistry，2019，36(5)：491-496.

Research on discrimination of tobacco leaf parts based on extracting different information of MIR

Zhao Juanjuan[1a]，Ye Shun[2]，Xu Ke[1b]，Chen Donghua[2]，Yue Baohua[1a]，Li Minjie[1a]，Liu Taiang[1a]，Lu Wencong[1a]

(1. a. Department of Chemistry；b. Department of Electronic Information Materials，
Shanghai University，Shanghai 200444，China；2. Technology Center of Shanghai Tobacco Group Co.，Ltd.，Shanghai 200082，China)

Abstract：Mid-infrared spectroscopy(MIR)technique has been widely applied in tobacco analysis，which could be used to obtain numerous chemical information. To investigate the IR spectrum data and improve the S/N value that contains large amount of tobacco information，data pre-treatment must be conducted. In this work，first-order derivative coupled with Savizky-Golay data treatment can not only improve the S/N value but also result in a better prediction accuracy of support vector classification(SVC)for tobacco leave part. It is found that dimension reduction can be used to eliminate redundancy information which helps to extract the characteristic data and saves the computation time. Here，Partial least squares(PLS)and successive projection algorithm(SPA)were used to extract the characteristic data from the infrared spectrum. The prediction accuracy of SVC for tobacco leave part showed that the result of dimension reduction via the PLS algorithm was better than that via the SPA algorithm. It can be concluded that the prediction accuracy of SVC for tobacco leave part was improved by using dimension reduction via the PLS to extract the most valuable data from MIR. In a word，the SVC accuracy for tobacco leave part from PLS data of training set，leave one out cross validation and testing set were 96.00%，89.60% and 80.65%，respectively.

Keywords：MIR；SPA；SVM；tobacco leave parts

[责任编校　赵晓华　陈留院]