# 基于 NIR 分析和模式识别技术的葛根品种及产地判别

刘秀明[1],李　涛[2],李源栋[1],马慧宇[3],段焰青[1],吴　宇[2],夏建军[1,*]
(1.云南中烟工业有限责任公司技术中心,云南昆明 650231;
2.科迈恩(北京)科技有限公司,北京 100080;
3.云南烟草质量监督检测站,云南昆明 650106)

**摘　要**:采用模式识别技术对不同品种(柴葛及粉葛)及不同产地(云南、安徽、广西、湖北、四川、重庆、湖南)的葛根进行判定。采集 12 个产地共 120 个葛根样品的近红外光谱数据,对光谱进行预处理并建立共有模式后,进行相似度及 PLS 判别分析,多元统计分析结果显示,除安徽柴葛外,其他组样品之间的相似度较高。分别选择不同的样品为测试集和训练集,基于 PLS-DA 对葛根种类粉葛和柴葛进行模式识别,对种类识别率为 100%,对产地识别率为 84.44%。采用 kNN 分析对葛根产地和品种同时进行模式识别,样品识别率达 100%。实验结果表明,采用 kNN 模式识别可以很好识别不同产地和类别的葛根样品,方法具有可行性和有效性,为利用近红外光谱结合模式识别技术进行葛根品种真伪优劣鉴别、道地性及产地可追溯研究提供了理论依据和实用方法。

**关键词**:近红外光谱,葛根,模式识别,品种及产地

# Recognition of Radix Puerariae Varieties and Origin Based on Pattern Recognition and Near Infrared Spectroscopy Technology

LIU Xiu-ming[1],LI Tao[2],LI Yuan-dong[1],MA Hui-yu[3],DUAN Yan-qing[1],WU Yu[2],XIA Jian-jun[1,*]
(1.Technology Center,China Tobacco Yunnan Industrial Co.,Ltd.,Kunming 650231,China;
2.Chemmind Co.,Ltd.,Beijing 100080,China;
3.Yunnan Province Tobacco Quality Supervision & Test Station,Kunming 650106,China)

**Abstract**:The pattern recognition method was used to determine the Radix Puerariae of different varieties(*Pueraria lobata* Ohwi and *Pueraria thomsonii* Benth) and origins(Yunnan,Anhui,Guangxi,Hubei,Sichuan,Chongqing,Hunan).120 samples from 12 producing areas were collected,then established common patterns after preprocessed NIR spectra.The similarity and PLS discriminant analysis of Radix Puerariae samples were preliminary,the result showed that the sample groups were high similar excepting *Pueraria lobata* Ohwi from Anhui.Different samples were selected as test set and training set,PLS-DA was used to identify the types of *Pueraria lobata* Ohwi and *Pueraria thomsonii* Benth,the identification rate for varieties was 100% and identification rate for origin was 84.44%.The kNN was also used for pattern recognition of varieties and origin at the same time.The identification rate of the samples was 100%.The result indicates that the proposed methods were feasible and effective.Moreover,this investigation provided the theoretical support and practical method for recognition of Radix Puerariae varieties and origin utilizing near infrared spectral data.

**Key words**:near infrared spectroscopy;radix puerariae;pattern recognition;varieties and origin
**中图分类号**:TS201.2　　**文献标识码**:A　　**文 章 编 号**:1002-0306(2018)22-0247-05
doi:10.13386/j.issn1002-0306.2018.22.043
**引文格式**:刘秀明,李涛,李源栋,等.基于 NIR 分析和模式识别技术的葛根品种及产地判别[J].食品工业科技,2018,39(22):247-251.

根据 2015 年《中国药典》[1]所收载的中药材,将葛根分为柴葛(豆科葛属植物野葛 *Pueraria lobata*(Willd.)Ohwi)和粉葛(豆科葛属植物葛的变种甘葛藤 *Pueraria thomsonii* Benth)两个品种,二者在纤维性、葛根素、大豆苷、大豆苷元等含量差异都较大,前者味苦只能入药,后者为药食两用[2]。葛根在我国分

**收稿日期**:2018-01-31
**作者简介**:刘秀明(1985-),女,硕士,工程师,主要从事烟用香精香料方面的研究工作,E-mail:lxmhh2013@163.com。
*通讯作者:夏建军(1979-),男,硕士,研究员,主要从事卷烟产品开发方面的研究工作,E-mail:13508717332@139.com。
**基金项目**:云南中烟工业有限责任公司科技项目(2018CP04);2015 年云南省技术创新人才项目(2016HB009)。

布广泛,资源丰富,但不同产地葛根受环境、气候等因素影响,品质之间差异较大。为此,鉴别葛根的地道性及质量评价一直是热门课题。目前,对于葛根药材及相关中药制剂的质量控制主要是采用色谱及光谱技术测定其中一个或多个有效成分的含量,以含量的多少来评价其质量的优劣。近年来,色谱指纹图谱分析应用较多[3-6]。

近红外光谱(NIR)波长范围在780~2498 nm,NIR光谱属于分子光谱,主要是由分子振动的非谐振性使分子从基态向高能级跃迁时产生,分子在跃迁过程中吸收能量,从而产生了吸收光谱。相对传统的化学分析技术,大多数类型的样品均可采用NIR光谱技术直接进行测定,而不需要进行物理、化学等任何处理,尤其对于固体样品,不需要进行溶剂提取等工艺,直接进行NIR光谱分析,具有快速、简便、高效、准确且成本较低,不破坏样品,不消耗化学试剂,不污染环境优点。因此,NIR光谱分析技术受到越来越多人的青睐,在农业[7]、食品[8]、石油化工[9]、生物医学[10-12]等领域被广泛研究和应用,相对于近红外定量分析方面的发展,近红外模式识别方面的研究进展较慢。近红外光谱模式识别是基于采集到的样品的光谱数据,采用计算机数学建模的方法,对样品进行识别和分类的方法。在化学计量学分析中,用于模式识别的原始数据特征越多,所包含信息越丰富,对于分析实验结果越有利。而近红外光谱往往包含了样品的大量特征信息,因此,将近红外光谱结合模式识别方法,能更加有效地对样品进行等级分类和属性判别。目前,基于NIR光谱信息的模式识别技术已经成为研究热点[13-16],在各个行业的产品真假识别、在线分类判别、原产地鉴定、产品质量监控与分析等方面发挥了重要的作用。近红外光谱模式识别主要分为两部分,首先是特征信息提取,常见的有效方法有主成分分析(PCA)[17]、偏最小二乘(PLS)[11,18]等,其次是分类器算法,常见的有效方法有线性判别分析(LDA)、人工神经网络(ANN)[19]、支持向量机(SVM)[20-21]等。其中特征信息提取是重要的基础性环节,它是对变量(如,波长)和样品对应的数据矩阵进行特征分析和数据降维。k近邻分类(k-nearest neighbor classification,kNN)[20,22]算法根据待识样本在特征空间中k个最近邻样本中多数样本的类别来进行分类,因此具有直观、无需先验统计知识、无师学习等特点,从而成为非参数分类的一种重要方法[23]。

本文通过ChemPattern软件,采用基于多元统计分析PLS-DA及kNN建模,开展基于NIR光谱的不同种类和产地来源的葛根化学模式识别,以期为中药葛根的质量评价与质量控制提供依据。

## 1 材料与方法

### 1.1 材料与仪器

柴葛及粉葛样品　从药店以及香精香料公司购买的安徽、广西、湖北、湖南、四川、云南及重庆等12个不同产地、不同批次共120个样品作为实验样品;详细信息如表1所示。

表1　葛根药材样品

Table 1　Summary of Radix puerariae samples

| 分组 | 产地 | 批次 | 说明 |
|---|---|---|---|
| 1组 | 安徽丰原大药房 | 10 | 柴葛 |
| 2组 | 安徽丰原大药房 | 10 | 粉葛 |
| 3组 | 云南恩典科技公司 | 10 | 粉葛 |
| 4组 | 广西一心医药集团 | 10 | 粉葛 |
| 5组 | 广西一心医药集团滕县 | 10 | 粉葛 |
| 6组 | 湖北同济堂药房 | 10 | 粉葛 |
| 7组 | 湖南香精香料公司 | 10 | 柴葛 |
| 8组 | 广西柳州香精香料公司 | 10 | 粉葛 |
| 9组 | 四川德仁堂药业 | 10 | 粉葛 |
| 10组 | 云南昆明一心堂药店 | 10 | 柴葛 |
| 11组 | 云南曲靖一心堂药店 | 10 | 柴葛 |
| 12组 | 重庆万州和平药房 | 10 | 粉葛(代表性样品) |
| 合计 | | | 120批 |

Antaris FT-NIR光谱仪,配备InGaAs检测器的漫反射积分球,及直径为5 cm石英底采样杯和旋转台　美国Thermo Fisher Scientific。

### 1.2 实验方法

1.2.1　葛根样品处理　先将葛根样品敲成小碎块,然后用旋风磨粉碎,过60目筛,装入密封袋中备用。

1.2.2　光谱数据的采集　在室内温度24~28 ℃下,相对湿度≤70%,开机预热光谱仪2 h;采集背景光谱后,把混匀的固体粉末样品放入样品杯中,使用压样器轻压平整,样品厚度≥10 mm;将装好样品的样品杯置于旋转台上,采集样品近红外漫反射光谱并保存,每个样品重新装样并连续进行3次平行采集。仪器参数为光谱扫描范围4000~10000 $cm^{-1}$;分辨率:8 $cm^{-1}$;扫描次数不低于64次。采集完成后,用95%乙醇2~3次洗净样品杯上的残留物,待乙醇挥发完毕后,进行下一个样品光谱的采集。

1.2.3　葛根品种及产地的多元统计分析　采集12个产地共120个葛根样品的近红外光谱数据,对光谱进行预处理并建立共有模式,对全部样品进行相似度分析、PLS判别分析及部分样品(除差别较大的安徽柴葛)的PLS分析,初步对样品种类及产地情况进行判定。

1.2.4　葛根品种及产地的模式识别　分别选择不同的样品为测试集和训练集,基于PLS-DA对葛根的种类(粉葛和柴葛)进行模式识别,另外对比PLS-DA和kNN两种方法,对葛根产地以及产地和种类同时进行识别,以样品识别率为依据,选定较为合适的模式识别方法。

### 1.3 数据处理

数据处理软件:ChemPattern化学计量学与化学指纹图谱系统解决方案软件2017版[科迈恩(北京)科技有限公司(Chemmind Technologies Co.,Ltd.)]。

## 2 结果与讨论

### 2.1 多元统计分析

将光谱数据导入ChemPattern软件,所有葛根样

品的红外透过率叠加图谱如图 1 所示，可以看出，安徽柴葛的红外光谱和其他组有明显的差异，剩余各组样品间的光谱曲线差异很小，很难对葛根的分类进行区分。

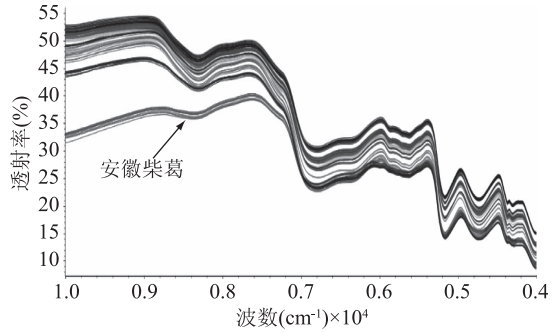


图 1　葛根样品的 NIR 透过率叠加图

Fig.1　Near infrared transmission spectra of Radix Puerariae

对光谱进行校正后，设置重庆万州的 10 批样品为代表性样品生成共有模式，如图 2 所示。利用所建立的共有模式，采用欧氏距离计算相似度，结果如图 3(A)所示，结果显示，除安徽柴葛外，其他组样品之间的相似度较高，仅可大致区分出安徽柴葛和其他组两个大类，但不能进行全部区分。对样品进行偏最小二乘判别分析，分析结果如图 3(B)所示。由图可知，偏最小二乘判别分析显示，安徽柴葛明显区别于其他组别的样本。

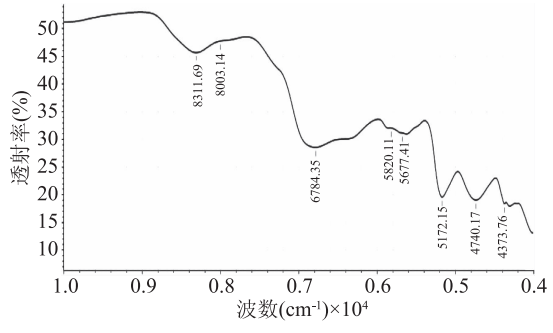


图 2　葛根样品模式

Fig.2　Common pattern of Radix Puerariae Transmission spectra

对除安徽柴葛外的样品做 PLS 分析如图 4(A)，可以看出，绝大部分地区样本都可以很好地区分，但四川粉葛和重庆万州(图 4A 实线圈)出现了重叠，四川和重庆地理位置相对比较接近，可以用于解释造成该两组样本相似度高的原因。此外，粉葛和柴葛的区别，从图中的分布也可以大致体现出来(图 4A、4B 虚线圈为柴葛样品)。以 LV1、LV2 和 LV3 进行作图 4B(实线圈)可以看出，原本重叠的四川粉葛和重庆万州也可以完全区分开。

## 2.2　葛根品种及产地的模式识别

2.2.1　葛根品种识别　从柴葛和粉葛每组中随机挑选 1/5 的样本作为测试集，以剩余的粉葛(58 个)和柴葛(32 个)作为训练集，进行 PLS-DA 模式识别，采用留一交叉验证选择潜变量数目。结果如图 5 所示，根据训练集留一交叉验证结果，选出潜变量个数为 3，训练集交叉验证葛根种类识别率为 100%。同时，利用测试集对所建 PLS-DA 模型进行评价，5 次随机分组建模测试集葛根种类识别率平均结果为 100%。以上结果表明所选的 PLS-DA 葛根种类识别模型准确可靠，可用于粉葛和柴葛的准确判别。

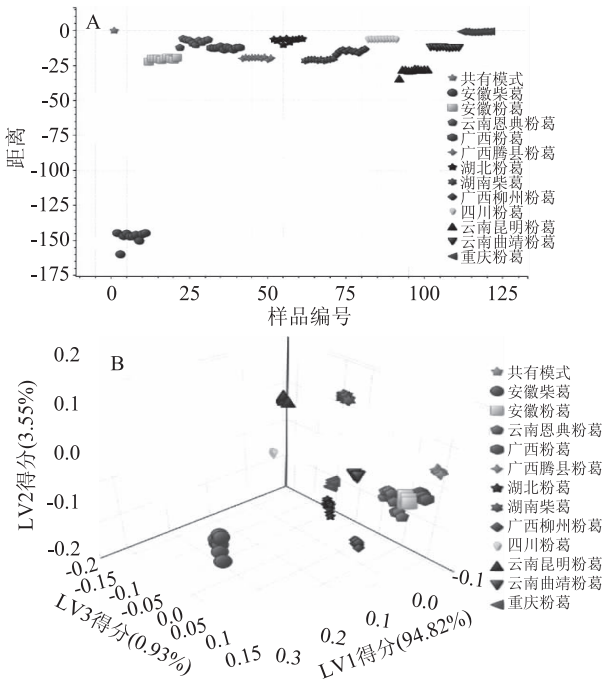


图 3　葛根样品相似度分析(A)及 PLS 潜变量分析(B)

Fig.3　Similarity analyses(A)and PLS-DA scores plot(B)of all Radix Puerariae sample

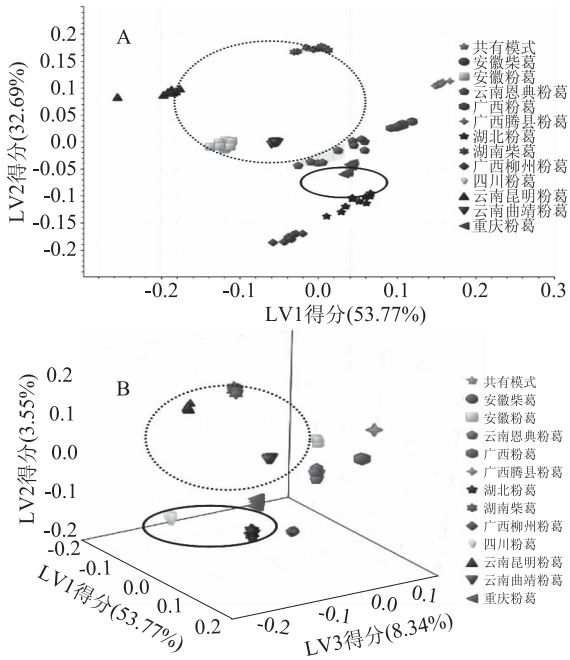


图 4　葛根(除安徽柴葛)PLS 潜变量分析图

Fig.4　PLS-DA scores plot

(Radix Puerariae sample from An hui excluded)

注：A：LV1/LV2，B：LV1/LV2/LV3。

2.2.2　葛根地点的识别　根据葛根的地点分布，首先将所有样品随机分为训练集(4/5)和测试集(1/5)，然后采用 PLS-DA 建立葛根产地识别模型。PLS-DA 的潜变量数用留一交叉验证确定，结果如图 6 所示，可以看出潜变量数为 16 时，模型的识别率最

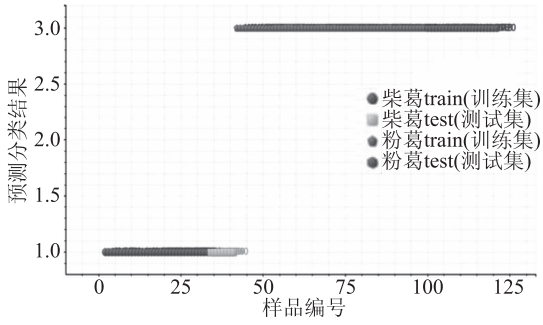


图 5　粉葛和柴葛的 PLS-DA 预测效果图

Fig.5　Performance of PLS-DA Radix Puerariae Varieties classification model

大仅为 84.43%，可能的原因是 PLS-DA 为线性模型，而红外光谱和地点信息之间可能是非线性的关系，所以导致模型的识别效果不够理想。

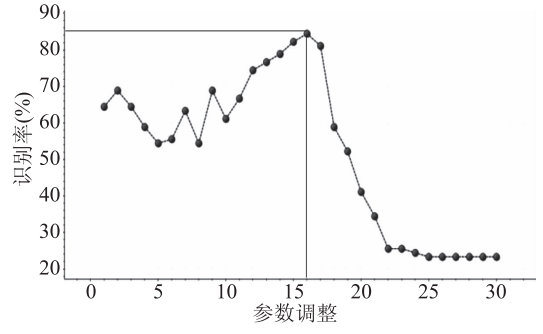


图 6　PLS-DA 建模潜变量数与模型识别率关系图

Fig.6　Accuracy of classification versus number of latent variable

采用 kNN 进行建模，留一交叉验证进行邻近样本数目 K 值的选择，结果如图 7 所示，可以看出 K 值为 1 或 2 时，结果最好；K 值增加到 3 时，模型效果有较大下降，而 K 越大模型越不容易过拟合，因此 K 值确定为 2。采用非线性的方法 kNN 进行建模结果如图 8 所示，可以看出 kNN 模型对各地点葛根可以进行很好的识别，模型训练集和测试集识别率均为 100%，表明模型对葛根地点的识别准确可靠，另外也表明葛根地点信息和红外光谱之间可能存在非线性关系。

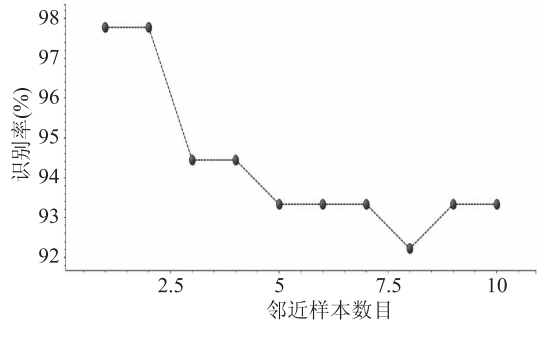


图 7　邻近样本数目与 kNN 葛根地点识别模型准确率关系图

Fig.7　Accuracy of kNN Radix Puerariae origin recognition model versus number of nearest neighbors

2.2.3　葛根种类地点同时识别　由前 2.2.2 可知，葛根地点和红外光谱之间存在非线性关系，因此将样

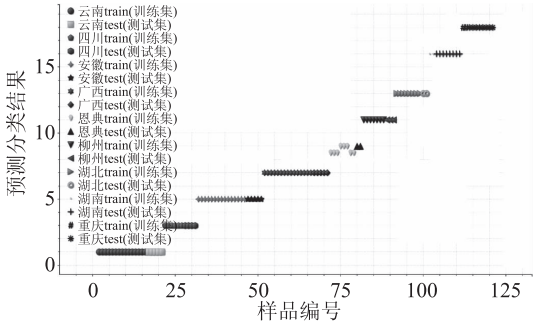


图 8　葛根地点的 kNN 模式识别效果

Fig.8　Performance of kNN origin identification model for Radix Puerariae

本分为训练集(4/5)和测试集(1/5)，采用 kNN 进行建模。如图 9 所示，采用留一交叉验证选出 K 值为 2，训练集交叉验证识别率为 99.30%。利用测试集评价所建 kNN 模型的识别效果，结果如图 10 所示，可以看出此时模型对训练集和测试集的识别率均为 100%，表明 kNN 模型可对葛根的产地和种类同时进行准确的识别。

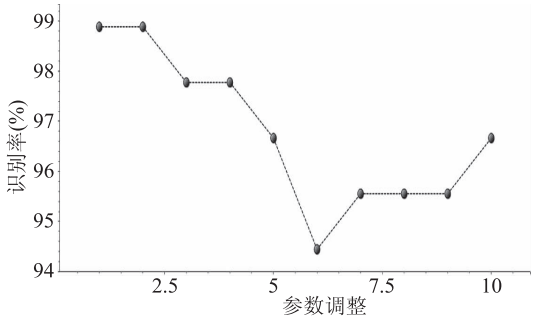


图 9　邻近样本数目与 kNN 葛根地点种类模型准确率关系图

Fig.9　Accuracy of kNN Radix Puerariae varieties and origin recognition model versus number of nearest neighbors

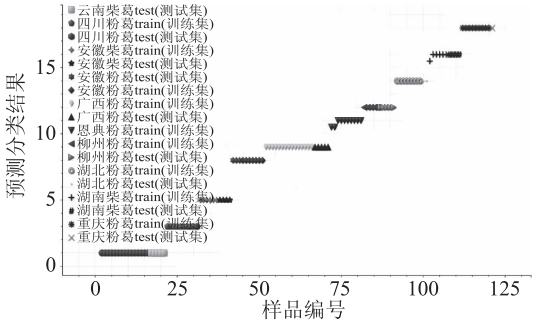


图 10　kNN 建模对葛根地点种类模式识别结果

Fig.10　Performance of kNN Radix Puerariae Varieties classification model

## 3　结论

基于多元统计分析，对 12 个产地共 120 个葛根样品进行相似度及 PLS 判别分析，结果显示，除安徽柴葛外其他组样品之间的相似度较高。结合药材外观，可以观察到安徽柴葛的纤维性强、颜色较深，其性状与其他野葛样品亦有不同，推断该样品可能为生长年限较长的野生品种。分别选择不同的样品为

测试集和训练集，基于 PLS-DA 对葛根的两个种类粉葛和柴葛进行模式识别，识别率达 100%，另外采用该方法对葛根产地的识别率为 84.44%，采用非线性的 kNN 后识别率提升为 100%，表明葛根地点信息和红外光谱间可能存在非线性关系。当采用 kNN 对葛根产地和品种同时进行模式识别，样品识别率达 100%。本文建立了基于近红外光谱的化学计量学模式识别方法，为葛根的质量评价及控制提供了可靠的评价新方法。

## 参考文献

[1]国家药典委员会.中华人民共和国药典[M].一部.2015.
[2]王苏静,赵新杰.葛根素的药理作用研究进展[J].内蒙古中医药,2010,29(2):107-108.
[3]黄芳,祝婧云,梁新丽.葛根药材中化学成分 HPLC 指纹图谱研究[J].世界中医药,2016,11(12):2789-2792.
[4]Lang Y,Wang X,Tan X,et al.Rapid screening of antioxidant active constituents from puerariae lobatae radix based on the investigation of quantitative patteren-activity relationship[J].Current Analytical Chemistry,2015,11(4).
[5]丁宁,何轶,何玉梅,等.HPLC-DAD 法测定冠脉宁片中葛根素和丹酚酸 B 的量及其指纹图谱研究[J].首都医科大学学报,2015,36(6):958-963.
[6]尤春雪,张振秋,李峰,等.HPLC 波长切换技术对葛根中 8 种成分的测定及指纹图谱研究[J].中草药,2013,44(5):616-621.
[7]Pan T,Lim,Chen J.Selection method of quasi-continuous wavelength combination with applications to the near-infrared spectroscopic analysis of soil organic matter[J].Applied Spectroscopy,2014,68(3):263.
[8]宋雪健,钱丽丽,张东杰,等.近红外光谱技术在食品溯源中的应用进展[J].食品研究与开发,2017,38(12):197-200.
[9]褚小立,许育鹏,陆婉珍.用于近红外光谱分析的化学计量学方法研究与应用进展[J].分析化学,2008,36(5):702-709.
[10]谢军,潘涛,陈洁梅,等.血糖近红外光谱分析的 Savitzky-Golay 平滑模式与偏最小二乘法因子数的联合优选[J].分析化学,2010,38(3):342-346.
[11]牟倩倩,贺敬霞,张建琪,等.近红外漫反射光谱法结合 PLS 法快速测定红景天药材中水分和红景天苷的含量[J].中国药房,2017(30):4260-4264.
[12]尹智伟.近红外光谱运用于中成药口服液多糖的快速分析[D].广州:暨南大学,2017.
[13]吴功煌,史新元,乔延江.近红外模式识别技术在中药质量控制中的应用研究进展[J].世界科学技术-中医药现代化,2010,12(2):265-270.
[14]Guo H,Pan T,Chen J,et al.Vis-NIR wavelength selection for non-destructive discriminant analysis of breed screening of transgenic sugarcane[J].Analytical Methods,2014,6(21):8810-8816.
[15]张峰,连芬燕,蓝洪桥,等.基于烟叶信息的近红外光谱和化学成分模式识别比较[J].江苏农业科学,2015,43(5):291-295.
[16]刘桂松,郭昊淞,潘涛,等.Vis-NIR 光谱模式识别结合 SG 平滑用于转基因甘蔗育种筛查[J].光谱学与光谱分析,2014,34(10):2701-2706.
[17]岳显可,杜伟锋,凌珏,等.基于高效液相色谱指纹图谱及聚类分析、主成分分析的市售片姜黄质量研究[J].时珍国医国药,2017(5):1095-1098.
[18]蒋程,丁静,孙云峰,等.基于偏最小二乘-判别分析算法的参麦注射剂不良反应关键影响因素分析[J].中国临床药学杂志,2017(4):221-225.
[19]陶益,任玉超,陈西,等.一种基于均匀设计及人工神经网络的中药组分配伍优化方法:中国,CN106266230A[P].2017.
[20]蔡洪民,王庆香.基于深度学习的入侵检测技术研究[J].网络安全技术与应用,2017(11):62-64.
[21]黄有强.基于 FTIR 离散平稳小波特征提取的 SVM 应用于中药材紫花地丁与同属植物的鉴别研究[J].中国现代应用药学,2017,34(5):692-696.
[22]王承伟,宾俊,范伟,等.基于近红外光谱技术结合随机森林的烟叶成熟度快速判别[J].西南农业学报,2017,30(4):931-936.
[23]李萌,沈炯.基于自适应遗传算法的过热汽温 PID 参数优化控制仿真研究[J].中国电机工程学报,2002,22(8):145-149.