基于潜在语义分析与 NIR 的中药材分类研究

陈晓峰 龙长江 牛智有 朱 凯
(华中农业大学工学院, 湖北 武汉 430070)

摘要 基于近红外光谱(NIR)和潜在语义分析(LSA)方法,对 5 种典型壮阳中药材进行分类鉴别研究。利用潜在语义分析对光谱预处理后的 5 种壮阳中药材光谱数据进行特征提取和鉴别分类后,将经光谱预处理和主成分分析(PCA)提取特征后的光谱特征数据分别带入 K 近邻(KNN)、BP 神经网络(BP-ANN)和偏最小二乘支持向量机(LS-SVM)三种典型的分类模型进行分类,并将结果与潜在语义分析模型结果进行对比。在 4119.20~9881.46 cm^{-1} 波数范围内,NIR 光谱数据经多元散射校正(MSC)预处理后,代入潜在语言空间维数为 3 时所建立的 LSA 分类模型,训练集和测试集准确率均达到了 100%。结果表明,在壮阳类中药材的近红外光谱分析鉴别中,潜在语义分析可以作为一种全新的提取光谱信息并分类的方法,具有较好的运用前景和实际意义。

关键词 光谱学;潜在语义分析;近红外光谱;壮阳中药材
中图分类号 O657.3 **文献标识码** A **doi**: 10.3788/AOS201434.0930001

Classification Research of Chinese Medicine Based on Latent Semantic Analysis and NIR

Chen Xiaofeng Long Changjiang Niu Zhiyou Zhu Kai
(*College of Engineering, Huazhong Agricultural University, Wuhan, Hubei* 430070, *China*)

Abstract Five kinds of typical Yang-boosting Chinese herbal medicine are identified and classified based on near infrared spectroscopy (NIR) and latent semantic analysis (LSA) methods. Latent semantic analysis is used for characteristic extraction and classification of preprocessed spectral data of 5 kinds of Yang-boosting Chinese herbal medicine. The spectral characteristic data, after spectral pretreating and characteristic extraction by principal component analysis (PCA), are respectively subjected into the K-nearest neighbour (KNN), BP-artifical neural networks (BP-ANN) and least squares support vector machine (LS-SVM) classification models whose results then are compared with the result of latent semantic analysis model. In the characteristic wavenumber range of 4119.20~9881.46 cm^{-1}, spectral data pretreated by multiplicative scatter correction (MSC) are substituted to LSA classification model when spacing dimension of underlying language is 3, and accuracy rates of both training set and test set are 100%. The results show that latent semantic analysis, which has a good application prospect and practical significance, can be used as a new method for spectral information extraction and classification in the near-infrared spectroscopy identification of Yang-boosting Chinese herbal medicine.

Key words spectroscopy; latent semantic analysis; near-infrared spectroscopy; Yang-boosting Chinese medicine
OCIS codes 300.6340; 070.4790; 070.5010

1 引 言

壮阳是指用温补药材强壮人体的心肾阳气,一般分为壮心阳与壮肾阳两种。壮心阳多用人参和附子等中药材,壮肾阳则多用鹿茸、巴戟天和锁阳等中药材[1]。通常所说的壮阳,范围较窄,单指壮肾阳。由于壮阳中药材在人们日常生活中应用较多,它们

收稿日期: 2014-04-01; **收到修改稿日期**: 2014-05-07
基金项目: 国家自然科学基金(61007058)、中央高校基本科研业务费专项基金(2014JC001)
作者简介: 陈晓峰(1988—),男,硕士研究生,主要从事近红外检测方面的研究。E-mail: charmy_cxf@163.com
导师简介: 龙长江(1975—),男,博士后,副教授,主要从事近红外检测与自动控制等方面研究。E-mail: lcjflow@163.com(通信联系人)
本文电子版彩色效果请详见中国光学期刊网 www.opticsjournal.net

的分类在现代中药产业中非常重要。传统中药鉴别方法主要依靠一定的技术手段和经验，成本高且效率低下，对于亲缘关系较近、外观较类似的品种很难获得准确的鉴别结果。

近红外光谱（NIR）分析技术以其简捷、无损和环保等优点[2]，提高了鉴别准确性，避免了主观因素的影响，近年来已在中药的真伪[3]、品种[4]、产地[5]和成分[6]等多方面鉴别中得到了良好的应用。但是壮阳中药材由于成分复杂同时有着共同的“壮阳”功效，导致其近红外光谱谱带复杂、重叠严重，传统的定性分析方法不易将其所含定性信息有效提取并进行分类。潜在语义分析（LSA）是一种很好的信息提取和分类方法，它通过统计计算大量文本，来寻找文本中的词与词之间存在的某种映射规则即某种潜在语义结构，进而提取特征并分类[7]。LSA 在文本分类中有着很广泛的应用，但是在光谱的分类中还不曾见到相关研究与应用。

本文基于近红外光谱，将潜在语义分析方法应用于 5 种壮阳中药材分类鉴别，并将 LSA 模型分类效果与经典的 K-近邻法（KNN）、BP 神经网络（BP-ANN）和偏最小二乘支持向量机（LS-SVM）模式识别方法模型分类效果进行对比。

2　仪器与材料

采用美国 Thermo Scientific 公司 Antaris II 型傅里叶变换近红外光谱仪，选择 Result3.0 光谱采集软件采集光谱，Matlab2008 软件进行编程分析光谱。

实验材料为中药店采购的被《中药大辞典》认定的 5 种壮肾阳药材，分别为鹿茸、淫羊藿、巴戟天、锁阳和菟丝子。

3　方　　法

3.1　样品制备与光谱采集

实验前，所有中药材经过干燥、粉碎和筛选后，依次随机称取 4 g 左右的粉末放入玻璃皿作为一个样本。将玻璃皿中的样本充分压实，采用积分球漫反射方式进行样品光谱采集。实验过程中，尽量让室内的温度和湿度保持一致。光谱采集范围为 10000～4000 cm^{-1}，扫描次数为 32 次，分辨率为 8 cm^{-1}。对每个样本在不同时间、不同位置分别采集 3 次信号，取 3 次采集的信号平均值作为该样本的原始光谱。5 种中药材每种药材取 60 个样本，每种样本随机选择 30 个作为校正集，剩下的 30 个作为预测集。校正集样本用于校正模型建立，预测集样本用于对模型的预测性能检验。每种药材取一个样本的原始光谱图，如图 1 所示（R 为默认的漫反射方法吸光度）。

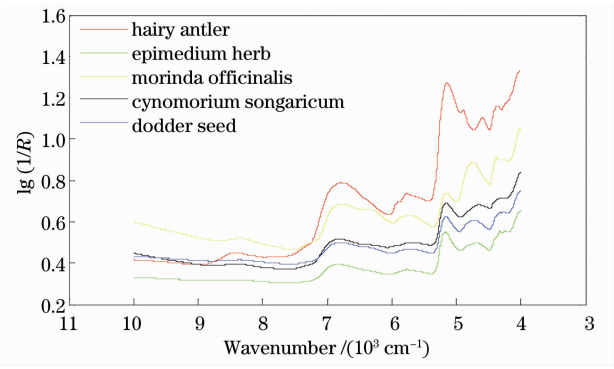

图 1　5 种壮阳中药材原始近红外光谱

Fig. 1　Original near infrared spectra of five kinds of Yang-boosting Chinese medicine

3.2　光谱数据预处理

近红外光谱的模式识别一般要经过光谱预处理、光谱特征提取和模型的建立与预测 3 个过程。光谱预处理中，选择恰当的谱区能够保留适量的样品特征波长建模，选择适当预处理方法可以减弱以至消除各种非目标因素对目标光谱的影响。在 4119.20～9881.46 cm^{-1} 范围内，对比各种预处理方法，最后 LSA 模型选择了多元散射校正（MSC）作为预处理，其他三种常规模型选择了 MSC、一阶导数和 Norris 微分平滑组合作为预处理。经过 MSC、一阶导数和 Norris 微分平滑预处理后的光谱如图 2 所示。

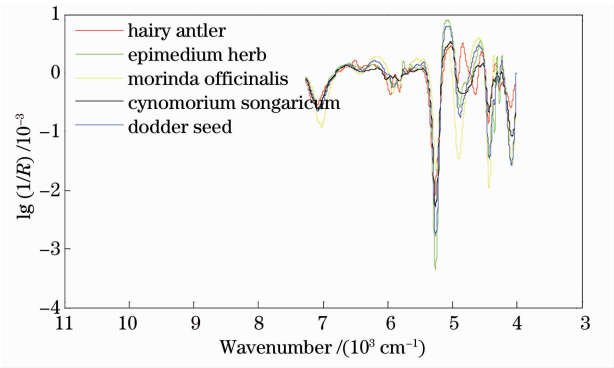

图 2　5 种壮阳中药材预处理后的近红外光谱

Fig. 2　Pretreated near infrared spectra of five kinds of Yang-boosting Chinese medicine

3.3　潜在语义分析模型原理与近红外模型转换

LSA 是一种信息检索方法，常用于文本的检索与分类。LSA 不仅消除了基于关键词检索方法中

存在的同义词、多义词问题，还通过奇异值分解(SVD)来选择合适的空间维度，有效地提高了检索速度和检索结果的准确率[8]。常规的模式识别方法是通过寻找输入和输出间存在的某种映射规则，LSA 则是通过寻找文本中的词汇与词汇之间存在的潜在语义结构，而这些语义结构可以通过构造词汇－文本矩阵来具体量化[9]。它的出发点是假设文本中存在决定词语语义相关性的某种潜在的语义结构，因此在不需要了解词语含义的情况下，能够根据与某语义词汇伴生词语出现的频次，将相近语义的词汇归类。中药由许多具有功效的成分构成，这些成分总是一起出现，协同发挥作用，文本由众多词汇构成，词汇具有语义，不同文本中很多词汇总是协同出现[10]，这为在近红外光谱基础上，将 LSA 运用到中药材分类中提供了可能。

在 LSA 模型中，词汇是含有重要信息的、组成文本的基本单元，文本是阐述某个主题的词汇集合。而在近红外光谱分析模型中，光谱波峰波段是含有物质组成信息的、组成特征光谱的单元，某类药材的光谱则蕴含着该类物质类别的信息。基于上述类似性，LSA 映射到近红外光谱模型中定义为将近红外某波峰段光谱对应成词汇、将某类中药材的近红外光谱对应成一个文本、将某个待预测样本的近红外光谱对应成提问式。

3.4 光谱数据的潜在语义分析计算

参照陈洁华[11]LSA 的方法，结合近红外光谱分析原理进行计算。

3.4.1 寻找特征波段

这里主要是寻找每个样本的光谱数据中，具有一定高度、宽度和个数的特征波段集。将近红外光谱曲线用函数 $y=f(x)$ 表示，其中 x 为波长，y 为是波长处的吸光度。再利用波峰查找函数 G，寻找波峰处波长 p，然后构造矩形窗口函数 ψ 来根据 p 截取波峰段集，这里称为特征波段集 T，即

$$T=\psi(p,\lambda)=\psi\{G[f(x)],\lambda\}, \tag{1}$$

式中 λ 是控制截取窗口宽度的参数。

为了提取每类样本的特征，即提取每类样本的共性，运用波段选取函数 ζ，并调整每类阈值 λ_j，在每类所有样本的 T 中选择特征性高的、对最后类别识别精度影响大的 k 个特征波段，此时特征个数继续增加识别精度将无明显增加，对应的特征组合即为最佳特征组合。利用此函数选取每类样本 k 个最优特征波段组合为

$$T_{\text{best}}=\zeta(T)=\{T_{\text{best1}},T_{\text{best2}},T_{\text{best3}},\cdots,T_{\text{best}k}\}, \tag{2}$$

式中 T_{best1}，T_{best2}，…，$T_{\text{best}k}$ 为每类样本中的 k 个最优特征波段。

3.4.2 文本特征波段集、提问式特征波段集和关键词特征波段集

设中药材总的种类数为 d，第 j 类中药材(即第 j 个文本)的阈值 λ_j，求解及本特征波段集。现求得第 j 类的最优特征组合集 T^j_{best}，则 d 类中药可以求出 d 个最优特征组合集 $\{T^2_{\text{best}},T^2_{\text{best}},T^3_{\text{best}},\cdots,T^d_{\text{best}}\}$，对上述 d 个最优特征波段集求并集，计为 T_s，且有 s 个波段，则

$$T_{\text{s}}=\bigcup_{j=1}^{d}T^j_{\text{best}}=\{T_{\text{s1}},T_{\text{s2}},T_{\text{s3}},\cdots,T_{\text{ss}}\}. \tag{3}$$

求提问式特征波段集 T_{q} 中，根据特征波段集求法求得某个提问式 q 的特征波段集 T_{q}。求关键特征波段集 T_{sq} 中，对于某个提问式 q，取 T_{q} 与 T_{s} 的交集，得到该提问式与文本的关键特征波段集 T_{sq}，记其一共有 m 个波段，则有

$$T_{\text{sq}}=T_{\text{s}}\cap T_{\text{q}}=\{T_{\text{sq1}},T_{\text{sq2}},T_{\text{sq3}},\cdots,T_{\text{sq}m}\}, \tag{4}$$

式中 $T_{\text{sq1}},T_{\text{sq2}},\cdots,T_{\text{sq}m}$ 为 m 个波段提问式与文本的关键特征波段集。

3.4.3 计算词汇-文本频率矩阵和提问式频率矩阵

LSA 中，最为重要的是频率矩阵的求取，这里先定义频率的计算。对于第 j 类中药的第 n 个样本在第 i 个波段的吸光度可以表示为 $f_n(T_i)$，某个测试集样本(提问式)在给定的某个波段 T_i 对应的吸光度与训练集中第 k 个样本，在波段 T_i 的吸光度的面积差函数为

$$\Delta a_i=\frac{\int_0^{L(T_i)}|f_n(T_i)-f_k(T_i)|\,\mathrm{d}t}{L(T_i)}, \tag{5}$$

式中 $L(T_i)$ 表示波段 T_i 的长度，t 为波长变量。

如果 Δa_i 超过每类给定的阈值 β_j 时，当前样本吸光度与给定波段的吸光度相差较大，计频数为 0；而当如果 Δa_i 小于给定的阈值 β_j 时，则计频数为 1。评价函数

$$\kappa(x,\beta_j)=\begin{cases}1, & x<\beta_j\\ 0, & x\geqslant\beta_j\end{cases}. \tag{6}$$

根据上述的频率计算定义，在关键词集 T_{sq} 的波段区内，词汇-文本频率矩阵中频数 $x_{ij}=\sum_{k=1}^{n}\kappa(\Delta a_i,\beta_j)$，则词汇-文本矩阵 $\boldsymbol{X}$ 为

$$\boldsymbol{X}=\{x_{ij}\}=\left\{\sum_{k=1}^{n}\kappa\left[\frac{\int_0^{L(T_i)}|f(T_i)-f_k(T_i)|\,\mathrm{d}t}{L(T_i)},\beta_j\right]\right\}. \tag{7}$$

计算提问式频率矩阵时，对于某个提问式 T_q 集，及其对应的关键词集 T_{sq}，将 T_q 和 T_{sq} 对比根据频率定义的计算方法，得到提问式的频率矩阵 $\boldsymbol{X}_q$，则提问式在潜在语义空间的表示为

$$\boldsymbol{D}_q = \boldsymbol{X}_q^{\mathrm{T}} \boldsymbol{T}_k \boldsymbol{S}_k^{-1}, \tag{8}$$

式中 $\boldsymbol{T}_k$ 为 k 维词汇矩阵，$\boldsymbol{S}_k$ 为 k 维奇异值矩阵，$\boldsymbol{D}_q$ 即为提问式在 k 维语义空间内的坐标向量。词汇、文本和提问式三者的坐标向量构成了潜在语义空间。

3.4.4　奇异值分解

一般得到的词汇-文本矩阵较为庞大，需要进行奇异值分解来提取有效信息并简化计算。分解后，选取奇异值为

$$\boldsymbol{X} = \boldsymbol{TSD} \approx \boldsymbol{X}_k = \boldsymbol{T}_k \boldsymbol{S}_k \boldsymbol{D}_k, \tag{9}$$

式中 $\boldsymbol{X}_k$ 即为降维后的语义结构，$\boldsymbol{D}_k$ 为 k 维文本矩阵，分别决定词汇和文本在 k 维潜在语义空间内的位置。

3.4.5　计算相似度并分类

对提问式坐标向量 $\boldsymbol{D}_q$ 与文本矩阵 $\boldsymbol{D}$ 的每一行向量分别进行比较，就可分别计算出提问式与每篇文本间的相关程度，即所需要的测试集里某个样本与训练集每个样本的相似度。采用计算夹角余弦值来计算相关程度，余弦值越大，则相关度越大，相似度为

$$\boldsymbol{C}_q = \frac{\sum_{i=1}^{k} \boldsymbol{D}_{q_i} \boldsymbol{D}_i}{\sqrt{\sum_{i=1}^{k} (\boldsymbol{D}_{q_i})^2} \cdot \sqrt{\sum_{i=1}^{k} (\boldsymbol{D}_i)^2}}, \tag{10}$$

式中 $\boldsymbol{C}_q$ 为提问式在 k 维空间内的坐标向量 $(\boldsymbol{D}_{q_1}, \boldsymbol{D}_{q_2}, \cdots, \boldsymbol{D}_{q_k})'$ 与某文本在 k 维空间内的坐标向量 $(\boldsymbol{D}_1, \boldsymbol{D}_2, \cdots, \boldsymbol{D}_k)'$ 之间的夹角余弦。

计算完提问式 q 与每一个文本之间的距离之后，那么提问式 q 就最接近距离最小的文本，模式分类得以实现。

4　结果与讨论

4.1　潜在语言空间维数 k 的选取

建立 LSA 模型中，对经过预处理的训练集和测试集光谱数据，将所有训练集样本数据求得 T_s，再依次取一个测试集样本求得 T_q 和 T_{sq}。求得该测试集样本下的词汇-频率矩阵 $\boldsymbol{X}$ 和提问式频率矩阵 $\boldsymbol{D}_q$。再对 $\boldsymbol{X}$ 进行奇异值分解和 $\boldsymbol{D}_{sq}$ 潜在语义空间转换，再进行相似度计算并分类。

潜在语言空间维数 k 的选择非常重要，k 值过大噪音过多，计算量增加，效率降低；k 值过小，会丢失有用信息，计算的准确率下降。根据翟琳琳提出的奇异值分解中 k 值自动选择算法[11]，选定 $k=3$，在效率和准确率间达到平衡。

4.2　LSA 模型的建立与测试

根据以上分析，现选用 4119.20～9881.46 cm^{-1} 范围内的光谱数据，经过 MSC 预处理后，选用 $k=3$ 建立测试最佳模型。当 k 为 3 时，LSA 模型训练集和测试集准确率均达到了 100%，如表 1 所示。

表 1　潜在语义分析模型分类结果

Table 1　Classification results of LSA model

Dimension k of latent semantic space	Recognition rate of calibration set /%	Recognition rate of predication set /%
2	94.67	95.33
3	100	100
4	100	100

4.3　与其他模式识别方法分类结果对比

主成分分析(PCA)是一种很好的把原来多个变量用少数个综合变量来表示的降维处理与数据挖掘技术[12]。在主成分分析后，这里分别使用 KNN[13]、BP-ANN[14] 和 LS-SVM[15] 三种典型的模式识别方法来建模并与潜在语义分析模型进行对比。在 4119.20～9881.46 cm^{-1} 光谱范围内，经过 MSC、一阶导数和 Norris 平滑预处理，再分别选择主成分个数后来分别建立分类模型。

在 KNN 模型中，主成分因子数(PCs)和近邻个数 K 都对模型的识别精度和稳定性有着重要影响，这里通过交互验证的方法来同时优化这两个参数。前 9 个主成分时累积贡献率为 98.57%，已突出了大部分光谱特征差异。表 2 为不同 PCs 和 K 时 KNN 模型的判别结果，由表 2 可以看出，当 K 为 5，PCs 为 9 时，所建立的 KNN 模型识别率最高，即取得的模型最佳。此时，模型对训练集与预测集中的样本，识别率分别达到了 94.67 %和 94%。

表 2　不同 PCs 和 K 对 KNN 模型判别结果影响

Table 2　Influences of different PCs and K on discrimination result using KNN model

Number of PCs	Number of neighbors K	Recognition rate of calibration set /%	Recognition rate of predication set /%
8	3	90.67	92
8	4	93.33	92.67
9	4	94.67	91.44
9	5	94.67	94
10	5	93.33	92.67

在 BP-ANN 模型中，建立一个三层的 BP 人工神经网络结构，各层传递函数都用 S 型函数。神经网络的输出层节点数为类别数，输入层节点数为主成分个数。隐层的节点个数 n_1[16]，和输入层神经节点数 n_2、输出层神经节点数 n_3 及 1～10 之间的常数 a 有如下关系：$n_1=\sqrt{n_2+n_3}+a$。不同拓扑结构对 BP-ANN 模型判别结果的影响如表 3 所示，最后确定网络输入层节点数为 15，隐含层节点数为 12，输出层节点数为 5，即拓扑结构为 15-12-5。同时网络参数通过实验设置如下：训练速率为 0.01，训练目标误差为 0.001，训练迭代次数为 1000。此时，网络性能最好，校正集和预测集分类正确率分别为 95.33%和 96%。

表 3 不同拓扑结构对 BP-ANN 模型判别结果影响

Table 3 Influences of different topologies on discrimination result using BP-ANN model

Topology structure	Recognition rate of calibration set /%	Recognition rate of predication set /%
12-11-5	92.67	94
13-10-5	91.44	94.67
14-14-5	94.67	90.67
15-12-5	95.33	96
16-8-5	93.33	93.33

在 LS-SVM 模型中选定径向基核函数(RBF)后，通过交叉验证和网格搜索，对惩罚系数 γ 和核函数参数 σ^2 这两个模型参数进行全局寻优选择[17]，确定最优 γ 和 σ^2 分别为 53.7334 和 1.5416。不同主成分数和模型参数对 LS-SVM 模型判别结果影响如表 4 所示，当输入为前 12 个主成分时校正集和预测集判别正确率分别达到 96%和 96.67%，此时的成分作为输入的 LS-SVM 模型结果良好。再增加主成分数时，模型预测性能变差，所以确定主成分数为 12。

表 4 不同主成分数和模型参数对 LS-SVM 模型判别结果影响

Table 4 Influences of different principal components and parameters on the results of LS-SVM model

Number of PCs	Accumulative contribution rate /%	Parameter		Recognition rate of calibration set /%	Recognition rate of predication set /%
		γ	σ^2		
10	98.24	45.1322	24.0211	96	93.33
11	99.01	86.1104	17.8230	94.67	94.67
12	99.13	53.7334	1.5416	96	96.67
13	99.18	33.4564	3.4742	95.33	94

对比可知，KNN 是最直接的基于距离的分类判别方法，结构简单，运算速度快，同时分类准确率也较高。而 BP-ANN 和 LS-SVM 需要分别建立神经网络和支持向量机，结构较 KNN 复杂，但是二者分类效果更好，同时稳定性也更高。LSA 需要提取特征光谱和建立潜在语义空间模型，运行时间比前面三种方法都长，但是它表现出了优异的校正和预测性能。

5 结 论

利用近红外光谱结合 LSA 识别 5 种成分复杂、功能近似的壮阳中药材，取得了比较理想的结果，证明了利用 LSA 在壮阳中药材鉴别分类上的可行性。同时，将 LSA 分类模型与 KNN、BP-ANN 和 LS-SVM 三种典型常规模式分类模型对比，结果表明 LSA 模型运算时间尽管比一般模式识别方法稍长，但是它在校正和预测性能上都有较明显的优势，作为一种全新的基于近红外光谱的分类方法，为后面的分类鉴别提供了一种新的思路与方法。最为重要的是它能够给出每种药材分类的光谱依据，这是传统方法所不具有的，LSA 具有很好的应用前景。

参 考 文 献

1 Li Jingwei. Dictionary of Chinese Medicine [M]. Beijing: People's Medical Publishing House, 1995. 712－712.
李经纬. 中医大辞典[M]. 北京：人民卫生出版社，1995. 712－712.

2 Hu Yongchuan, Tian Xiaoxin, Liu Lei, *et al*.. Advances in identification of Chinese medicines by NIRS [J]. China Journal of Chinese Materia Medica, 2012, 37(8): 1066－1071.
胡咏川，田晓鑫，刘 蕾，等. 近红外光谱技术鉴定中药的进展[J]. 中国中药杂志，2012，37(8)：1066－1071.

3 Gao Hongbin, Liu Hao, Xiang Bingren. Rapid and nondestructive identification of pinellia rhizome and its pseudo-product arisaema rhizome by near-infrared diffuse reflectance spectrometry [J]. Chinese Journal of Spectroscopy Laboratary, 2012, 29(2): 899－902.
高鸿彬，刘 浩，相秉仁. 半夏及其伪品天南星的近红外漫反射快速无损鉴别[J]. 光谱实验室，2012，29(2)：899－902.

4 Gong Haiyan, Bai Yan, Song Ruili, *et al*.. The discrimination of

tigun yam and baiyu yam using near infrared spectroscopy [J]. Chinese Journal of Hospital Pharmacy, 2010, 30(9): 735-737.
龚海燕，白　雁，宋瑞丽，等. 近红外光谱结合聚类分析鉴别铁棍山药和白玉山药[J]. 中国医院药学杂志，2010，30(9)：735-737.

5 Du Min, Gong Yin, Lin Zhaozhou, *et al*.. Rapid identification of wolf berry fruit of different geographic regions with sample surface near infrared spectra combined with multi-class SVM [J]. Spectroscopy and Spectral Analysis, 2013, 33(5): 1211-1214.
杜　敏，巩　颖，林兆洲，等. 样品表面近红外光谱结合多类支持向量机快速鉴别枸杞子产地[J]. 光谱学与光谱分析，2013，33(5)：1211-1214.

6 Zhang Jing, Geng Zhipeng, Fan Gang, *et al*.. A new method for analysis of six akaloids in coptis chinensis franch by near infrared diffuse reflectance spectroscopy [J]. Lishizhen Medicine and Materia Medica Research, 2011, 22(10): 2393-2394.
张　静，耿志鹏，范　刚，等. 近红外光谱技术测定黄连中6种生物碱含量的新方法[J]. 时珍国医国药，2011，22(10)：2393-2394.

7 Jiang Jianhong, Luo Mei. Latent semantic information extraction and classification of online product [J]. Computer & Digital Engineering, 2014, 42(1): 112-115.
蒋建洪，罗　玫. 在线商品的潜在语义信息提取及分类研究[J]. 计算机与数字工程，2014，42(1)：112-115.

8 Liu Bo. Application of Latent Semantic Indexing Chinese Information Retrieve [D]. Beijing: Beijing University of Posts and Telecommunications, 2009. 10-17.
刘　博. 潜在语义索引在中文信息检索中的应用[D]. 北京：北京邮电大学，2009. 10-17.

9 Jian Yan. Chinese Text Clustering Based on Latent Semantic and Its Applications [D]. Shenyang: Northeastern University, 2008, 5-12.
简　艳. 基于潜在语义的中文文本聚类及其应用[D]. 沈阳：东北大学，2008. 5-12.

10 龙长江，万　鹏. 近红外检测技术在中药研究中的应用[C]. 中国农业工程学会2011年学术年会论文集，2011. 1704-1707.

11 Chen Jiehua. Theory and Application of Latent Semantic Analysis [D]. Shanghai: Shanghai University, 2005. 12-21.
陈洁华. 潜在语义分析理论研究及其应用[D]. 上海：上海大学，2005. 12-21.

12 Guo Peiyuan, Lin Yan, Fu Yan, *et al*.. Research on freshness level of meat based on near-infrared spectroscopic technique [J]. Laser & Optoelectronics Progress, 2013, 50(3): 033002.
郭培源，林　岩，付　妍，等. 基于近红外光谱技术的猪肉新鲜度等级研究[J]. 激光与光电子学进展，2013，50(3)：033002.

13 Zhao Jiewen, Jiang Pei, Chen Quansheng. Discrimination of snow lotus from different geographical origins by near infrared spectroscopy [J]. Transactions of the Chinese Society for Agricultural Machinery, 2010, 41(8): 111-114.
赵杰文，蒋　培，陈全胜. 雪莲花产地鉴别的近红外光谱分析方法[J]. 农业机械学报，2010，41(8)：111-114.

14 Niu Xiaoying, Shao Limin, Zhao Zhilei, *et al*.. Nondestructive discrimination of strawberry varieties by NIR and BP-ANN [J]. Spectroscopy and Spectral Analysis, 2012, 32(8): 2095-2099.
牛晓颖，邵利敏，赵志磊，等. 基于BP-ANN的草莓品种近红外光谱无损鉴别方法研究[J]. 光谱学与光谱分析，2012，32(8)：2095-2099.

15 Jiang Shiquan, Zhou Xingcai, Jiang Shiping. Discriminative model for FTNIS analysis on age of Shaoxing rice wine based on PCA and LS-SVM [J]. Chinese Journal of Spectroscopy Laboratory, 2012, 29(2): 806-811.
蒋诗泉，周兴才，蒋诗平. 基于PCA和LS-SVM的傅里叶变换近红外光谱的黄酒酒龄的鉴别模型[J]. 光谱实验室，2012，29(2)：806-811.

16 Shao Yongni, Cao Fang, He Yong. Discrimination years of rough rice by using visible/near infrared spectroscopy based on independent component analysis and BP neural network [J]. Journal of Infrared and Millimeter Waves, 2007, 26(6): 433-436.
邵咏妮，曹　芳，何　勇. 基于独立组分分析和BP神经网络的可见/近红外光谱稻谷年份的鉴别[J]. 红外与毫米波学报，2007，26(6)：433-436.

17 Q Chen, J Zhao, H Lin. Study on discrimination of Roast green tea (Camellia simensis L.) according to geographical origin by FT-NIR spectroscopy and supervised pattern recognition [J]. Spectrochim Acta Part (A), 2009, 72(4): 845-850.

栏目编辑：史　敏