第39卷，第12期
2019年12月

光 谱 学 与 光 谱 分 析
Spectroscopy and Spectral Analysis

Vol. 39，No. 12，pp3809-3815
December，2019

近红外光谱 LASSO 特征选择方法及其聚类分析应用研究

李鱼强[1]，潘天红[1,2*]，李浩然[1]，邹小波[3]

1. 江苏大学电气信息工程学院，江苏 镇江 212013
2. 安徽大学电气工程与自动化学院，安徽 合肥 230061
3. 江苏大学食品与生物工程学院，江苏 镇江 212013

摘 要 近红外光谱技术是一种通过分析样本的特征光谱数据，实现定性或定量分析的无损检测方法，特征数据的完整性和代表性决定了所建模型的性能，而现有分析方法只能实现光谱子区间特征筛选，导致分析模型稳定性差、且难以再优化。为实现近红外光谱区间高维数特征提取，有效提高近红外光谱定性分析模型的精度和稳定性，提出一种基于最小绝对收缩和选择算法(LASSO)的光谱特征筛选方法，并以我国特色高值外贸产品云南松茸为分析对象进行聚类应用研究，讨论了该方法对于高维光谱特征筛选的有效性、分析对比了 LASSO 筛选特征变量及主元分析(PCA)降维算法所建松茸真伪甄别及食用菌分类模型的预测精度及稳定性。通过调研发现，云南产鲜松茸因其独特外形易于分辨，而片状的干松茸失去其独有的外形特征，导致国内干松茸掺假事件屡禁不止。选取云南产松茸、杏鲍菇、老人头、姬松茸四种干样共 166 样本数据进行分析，采用光谱范围为 900～1 700 nm 的 NIRQuest512 型近红外光谱仪获得 166×512 维原始光谱数据，剔除异常数据后采用标准正态变换对光谱数据进行预处理。在此基础上，利用 LASSO 筛选出全光谱区间的特征变量，再使用 Kennard-Stone 法并结合典型线性(KNN)和非线性建模(BP)算法，构建松茸真伪甄别模型和食用菌分类模型，对两种模型进行盲样测试，并分析了 LASSO 与 PCA 算法的不同点，最后使用蒙特卡罗方法检测两种模型的稳定性。实验结果表明基于 LASSO 光谱特征选择的松茸真伪甄别模型和食用菌分类模型预测精度和稳定性均高于 PCA 方法，其中基于原始光谱数据所建真伪甄别模型的预测准确率为 69.57％ (BP)和 60.87％ (KNN)，食用菌分类模型准确率为 67.39％ (BP)和 65.22％ (KNN)，基于 LASSO 特征筛选的真伪甄别模型预测准确率分别达到 100％ (BP)和 78.26％ (KNN)，食用菌分类模型预测准确率分别达到 89.13％ (BP)和 80.43％ (KNN)，对两种模型进行 10 次蒙特卡罗实验，其结果平均值分别为 99.93％和 97.22％，由此可知，与 PCA 等数据降维算法相比，LASSO 可实现全光谱区间的光谱特征选择和数据降维，有效地提高了近红外定性分析模型的预测性能，为近红外分析提供了一种新的特征筛选方法。

关键词 近红外光谱；特征选择；LASSO；松茸鉴别；蒙特卡罗方法

中图分类号：Q657.3 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2019)12-3809-07

引 言

近红外光谱分析技术能够快速测定样品中化学成分含量和特性，具有高效、快速、无损等特点，已广泛应用于农业、食品、医药、石油及化工等生产领域，是我国目前产品定性、定量检测分析的重要方法[1]。

根据检测对象及分析指标的不同，近红外光谱分析样本内部所含氢团的种类和数目存在差异，该差异在光谱图上体现为特征峰之间的不同，近红外光谱分析技术正是通过分析特征峰所对应的特征数据，完成对不同对象或同类对象的定性和定量研究[2]。但是，如何有效地从包含大量无关数据的全光谱区间筛选特征数据，一直是光谱分析领域的难点之一[3]。目前近红外光谱特征提取方法主要有两种策略，一种是通过主元分析(principal component analysis，PCA)、UVE 等方法进行光谱数据压缩，如吴习宇等将 PCA 和 DPLS，SVM 相结合，建立了花椒粉掺假定性分析模型，有效降低了模型复杂度[4]；Xu 等利用 PCA 和 BP 算法建立了肌肉腐变

收稿日期：2018-10-31，**修订日期**：2019-02-10

基金项目：国家重点研发计划(2017YFF0211301)，江苏省重点研发项目(BE2018370)资助

作者简介：李鱼强，1995 年生，江苏大学电气信息工程学院硕士研究生 e-mail：1763371201@qq.com

* 通讯联系人 e-mail：thpan@live.com

预测模型，实现了低维数据模型下的定性分析[5]。另外一种方法是通过选择光谱区间进行特征筛选，如李路等通过将竞争性自适应重加权技术和多元线性回归相结合，通过两次波长筛选，完成对稻谷近红外特征数据波长区间的筛选，实现对稻谷脂肪的高精度预测[5]，吴瑞梅等提出间隔偏最小二乘法(siPLS)结合遗传算法(GA)的特征筛选方法，该方法通过对比不同光谱子区间的预测结果，选择性能较优的区间组成联合子区间，实现了对农药乳油中毒死蜱含量的精准测定[7]。然而通过数据压缩或联合区间法筛选光谱特征数据，只能筛选特征变量相对集中的波长区间以减少建模数据，未实现全光谱区间数据降维和特征筛选，但是对于近红外建模而言，由部分特征变量所构建的近红外模型稳定性差、预测精度低，而且模型性能无法再优化。

一般而言，只有高维全光谱区间筛选变量同时具有代表性和完整性，才能有效提高近红外光谱分析模型的精度和稳定性，其中代表性即为所筛选变量是特征变量，完整性则为所筛选变量全部特征变量。传统特征数据提取方法无法同时确保筛选变量的代表性和完整性，因此无法实现全光谱区间的特征提取。为此，提出一种基于最小绝对收缩和选择算法(least absolute shrinkage and selection operator，LASSO)的光谱特征选择方法，通过在线性回归的基础上增加范式函数约束，对全光谱区间数据进行特征筛选，只保留与目标变量相关性高的解释变量实现特征变量筛选，基于该特征变量进行建模分析，并将该模型应用于云南产松茸真伪甄别研究。

1 实验部分

1.1 仪器与参数

实验用近红外光谱仪为美国 Ocean Optics 公司生产的 NIRQuest512 型近红外光谱仪，配置波长范围为 360～2 000 nm 的 HL-2000 系列卤钨灯光源，光谱仪分辨率为 3 nm，积分时间为 45 ms，扫描波长范围为 900～1 700 nm，内置具有 512 个像素点、高精度的铟镓砷化物(InGaAs)阵列探测器，扫描次数为 32 次。本实验所有数据处理软件为 Matlab2016b。

1.2 样品制备

云南作为我国最大的食用菌产地，其松茸产量占全国松茸总产量的七成以上。通过调研发现，鲜松茸外形独特，容易分辨，但干松茸则丢失原有生物形态，市场存在的多种相似形态食用菌干样本导致松茸干样本造假事件时有发生。本实验以云南产松茸干样本及其主要假冒替代品作为检测对象，进行真假鉴别和食用菌类别分析研究。2018 年 9 月于云南昆明市木水花食用菌交易市场购买野生松茸(优质、劣质)、姬松茸、老人头、杏鲍菇共 166 个切片干样，为增加样本代表性，所选样本包含检测食用菌各个生长周期内干样本，图 1 为实验所采集样品照片，表 1 所示为所采集部分样本重量。

1.3 数据采集

采用 Kennard-Stone 法选取 38 个样本的 105 组数据作为训练集，18 个样本的 42 组数据作为校正集，其余样本数据做测试集。清理样品表面杂质并编号，放入保温箱内保存，选取样品头部和中部位置各 3 个点采集光谱反射率，取平均值作为该样品相应位置的光谱数据。在近红外建模中，原始光谱数据除了包含检测样本的特征数据之外，还包含较多的冗余变量和由外界因素引起的噪声信号，为提高模型的预测性能，需要在建模之前进行光谱预处理[8]。选取标准正态变量变换(standard normal variate transformation，SNV)作为光谱预处理算法，SNV 算法主要用于消除漫反射数据采集过程中因样本粒径大小分布不均匀、表面不平滑及光程所导致的光谱差异，其计算表达式为

$$x_{i,j}=\frac{x_{i,j}^{\mathrm{org}}-\mu_j}{\sigma_j} \tag{1}$$

式中，$X^{\mathrm{org}}=\{x_{i,1}^{\mathrm{org}},x_{i,2}^{\mathrm{org}},\cdots,x_{i,m}^{\mathrm{org}}\}_{i=1}^{n}$ 为样本原始光谱数据，m，n 分别为波长数与样本数，$\mu_j=\frac{1}{n}\sum_{i=1}^{n}x_{i,j}$ 为样本光谱数据均值，$\sigma_j=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{i,j}-\mu_j)^2}$ 为样本的标准差，$X=\{x_{i,1},x_{i,2},\cdots,x_{i,m}\}_{i=1}^{n}$ 为经过 SNV 预处理后的光谱数据。

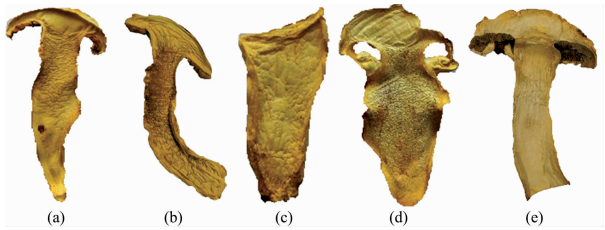

图 1 实验样品照片

(a)：优质松茸；(b)：劣质松茸；(c)：杏鲍菇；(d)：老人头；(e)：姬松茸

Fig. 1 Experimental samples

(a)：High quality matsutake；(b)：Inferior matsutake；(c)：Pleurotuseryngii；(d)：Jujube hilt nipple mushroom；(e)：Agaricusblazei

表 1 检测样本重量(g)

Table 1 Weight of test samples (g)

指标	最大值	最小值	平均值
杏鲍菇	1.11	0.43	0.74
老人头	2.41	0.66	1.25
姬松茸	1.90	0.69	1.34
松茸优质	2.80	0.96	1.64
松茸劣质	2.16	0.86	1.60

1.4 LASSO 光谱特征数据选择

原始数据维度高、建模复杂度高，基于光谱区间优化和数据压缩等方法无法实现全光谱区间特征变量选择及数据降维，LASSO 算法可对全光谱区间数据进行特征筛选，有效降低建模数据维度。LASSO 方法是在线性回归的基础上，通过增加范式函数，将模型回归系数的绝对值约束在某一个设定的阈值，并最小化模型残差平方和[8]。该算法通过优化目标函数将相关性小于阈值的变量压缩为 0 并进行剔除，剩余变量即为所需特征变量。

设定线性回归模型为

$$\boldsymbol{Y}=\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}+\boldsymbol{\varepsilon} \tag{2}$$

式中，$\boldsymbol{X}=[x_1, x_2, \cdots, x_i, \cdots, x_n]^{\mathrm{T}}$，$\boldsymbol{x}_i=[x_{i,1}, x_{i,2}, \cdots, x_{i,m}]\in R^{1\times m}$ 为经过 SNV 预处理的光谱数据，$\boldsymbol{Y}=[y_1, y_2, \cdots, y_n]^{\mathrm{T}}\in R^{n\times 1}$ 为响应变量，$\boldsymbol{\beta}=[\beta_1, \beta_2, \cdots, \beta_m]^{\mathrm{T}}\in R^{m\times 1}$ 为模型系数，$\boldsymbol{\varepsilon}=[\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n]^{\mathrm{T}}\in R^{n\times 1}$ 为误差向量。线性回归模型的普通最小二乘法估计为 $\min\left[\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{m}x_{i,j}\beta_j\right)^2\right]$，可 $\hat{\boldsymbol{\beta}}_{\mathrm{ols}}=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$，当增加约束函数时，即 LASSO，具体表示为

$$\underset{\beta}{\operatorname{argmin}}\left[\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{m}x_{i,j}\beta_j\right)^2+\lambda\sum_{j=1}^{m}|\beta_j|\right] \tag{3}$$

式中参数 λ 为参数估计的惩罚系数，其大小一般通过交叉验证或者基于信息准则模型确定，参数 α 通过交叉验证确定。

LASSO 回归算法的主要解法有坐标轴下降法(coordinate descent)和最小角回归法(least angle regression)。其中坐标轴下降法是 LASSO 回归的最快解法，但是其变量计算过程只能沿坐标轴进行，而最小角回归法是一种基于前向选择算法和前向梯度算法的变量筛选算法，能够得到更为精确的特征向量，本文采用最小角回归法[10]，具体描述如下：

(1)前向选择算法的计算过程为：在 $\boldsymbol{X}=[x_1, x_2, \cdots, x_i, \cdots, x_n]^{\mathrm{T}}$ 中选择和目标变量 y_k 最为接近的自变量 $\boldsymbol{x}_k=[x_{k,1}, x_{k,2}, \cdots, x_{k,m}]$，有

$$\bar{y}_k=\boldsymbol{x}_k\beta_k \tag{4}$$

其中，系数 β_k 由式(5)确定

$$\beta_k=\frac{\langle \boldsymbol{x}_k, y_k\rangle}{\|\boldsymbol{x}_k\|_2} \tag{5}$$

变量残差为

$$y_{\mathrm{res},k}=y_k-\bar{y}_k \tag{6}$$

将变量残差定义为新的目标变量，同时将不含 $\boldsymbol{x}_k$ 的集合 $\boldsymbol{X}$ 作为新的自变量集合，重复上述过程，直至残差小于设定范围或自变量集合个数为零，算法终止。

(2)前向梯度算法每次选取一个相关性最大的特征变量 $\boldsymbol{x}_k$ 逼近目标变量 y_k，跟前向选择算法不同的是，其残差定义为

$$y_{\mathrm{res},k}=y_k-\boldsymbol{x}_k\boldsymbol{\beta}_k \tag{7}$$

将残差作为新的目标函数，原变量集 $\boldsymbol{X}=[x_1, x_2, \cdots, x_i, \cdots, x_n]^{\mathrm{T}}$ 作为变量集，根据式(7)重新计算，直至残差 $y_{\mathrm{res},k}$ 小于设定阈值范围，得到最优解。

LASSO 算法流程如图 2 所示，具体步骤如下：

Step1：[目标变量] 根据式(4)和式(5)求解与目标函数相关度最高的变量 $\boldsymbol{x}_k$，并将其从变量集合中剔除，根据式(7)确定新目标变量；

Step2：[相关变量] 重复 Step1，直至得到新的变量 $\boldsymbol{x}_l$ 与目标变量 $y_{\mathrm{res},k}$ 的相关度和变量 $\boldsymbol{x}_k$ 与 $y_{\mathrm{res},k}$ 的相关度相同；

Step3：[特征变量] 在 $\boldsymbol{x}_k$ 和 $\boldsymbol{x}_l$ 的角平分线上，利用式(7)重新逼近，得到变量 $\boldsymbol{x}_t$，使得 $\boldsymbol{x}_t$ 与 $y_{\mathrm{res},k}$ 的相关度和 $\boldsymbol{x}_k$，$\boldsymbol{x}_l$ 与 $y_{\mathrm{res},k}$ 的相关度一样，将变量 $\boldsymbol{x}_t$ 添加到特征集合中，并以该集合的共同角平分线作为新的逼近方向；

Step4：[循环] 循环上述过程，直至 $y_{\mathrm{res},k}$ 足够小或者变量集合为空，最终特征集合为所求特征变量。

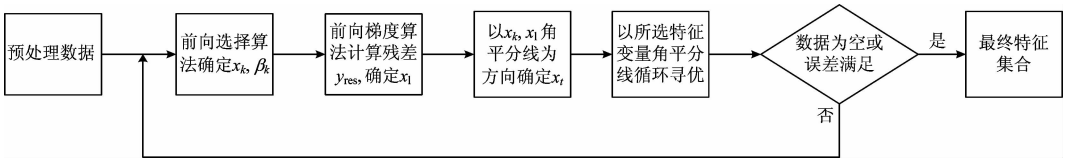

图 2　LASSO 技术流程图

Fig. 2　Flowchart of LASSO

2　结果与讨论

2.1　近红外光谱数据

对所有切片样本的头部和中部进行光谱数据采集，并对姬松茸和野生松茸干样本的根部进行数据采集，共获得 166 组样本数据，所得原始光谱如图 3 所示。图 3 显示原始光谱数据的特征峰不明显，基于该数据所建模型稳定性差，需采用预处理去噪，图 4 为经过 SNV 预处理后的光谱数据。

2.2　LASSO 特征筛选

使用 SNV 光谱预处理后的光谱数据谱线平滑，特征峰明显，但该数据仍是高维数据，基于该数据所建模型复杂度高、计算量大、预测性能低，因此采用 LASSO 算法对经过预处理的光谱数据进行特征筛选，挑选相关性较高的光谱数据进行建模分析。

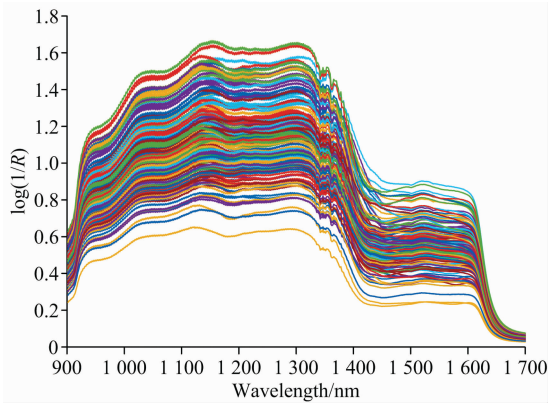

图 3　原始光谱数据

Fig. 3　Original spectral data

LASSO 算法是光谱分析领域一种较新的分析方法，但是分析效果不逊于传统的数据优化算法，区别于其他算法其

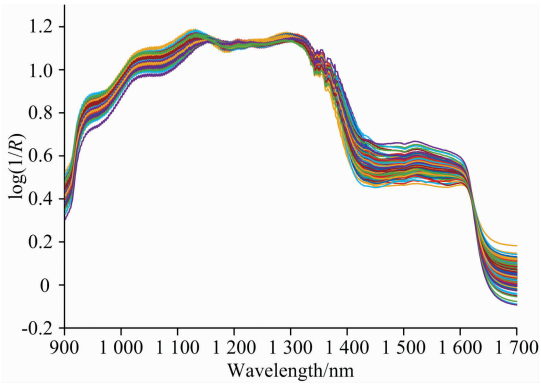

图 4 预处理光谱数据

Fig. 4 Preprocessed spectral data

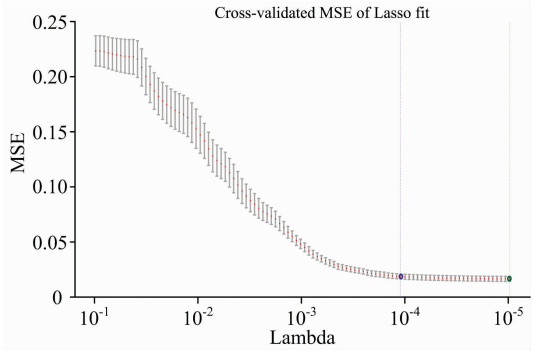

图 5 交叉验证确定 λ 参数

Fig. 5 Cross validation determination of λ value

主要特点在于可通过调节式(3)中 λ 参数决定目标函数惩罚系数，将无关变量压缩为 0，降低数据维度。本实验 λ 参数通过 10 次交叉验证确定，建模参数选取过程如图 5 所示，参数 λ 与系数矩阵 $\boldsymbol{\beta}$ 的关系图如图 6 所示。

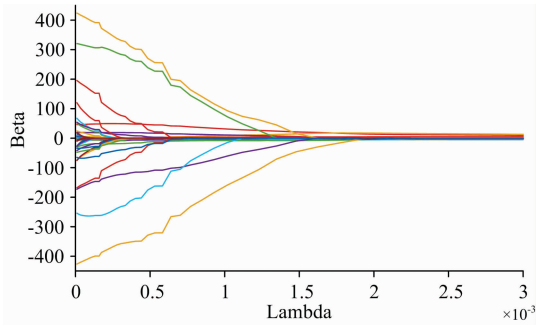

图 6 参数 λ 与系数 $\boldsymbol{\beta}$

Fig. 6 The relationship between λ and $\boldsymbol{\beta}$

2.3 松茸真伪甄别

将野生松茸样本标记为 1(真)，其余样本标记为 0(伪)，以 LASSO 算法所筛选的特征变量作为输入变量，相应属性作为输出变量，分别选取了线性建模(k-nearest neighbors algorithm，KNN)算法和非线性建模(back propagation neural network algorithm，BP 神经网络)进行松茸真伪鉴别分析。KNN 算法通过计算待分类样本与已训练样本之间的距离完成分类，是目前应用最广泛的线性分析方法[11]；BP(back propagation)神经网络是一种基于误差反向传播的多层前馈神经网络，可根据建模对象的不同，调节其各层神经元个

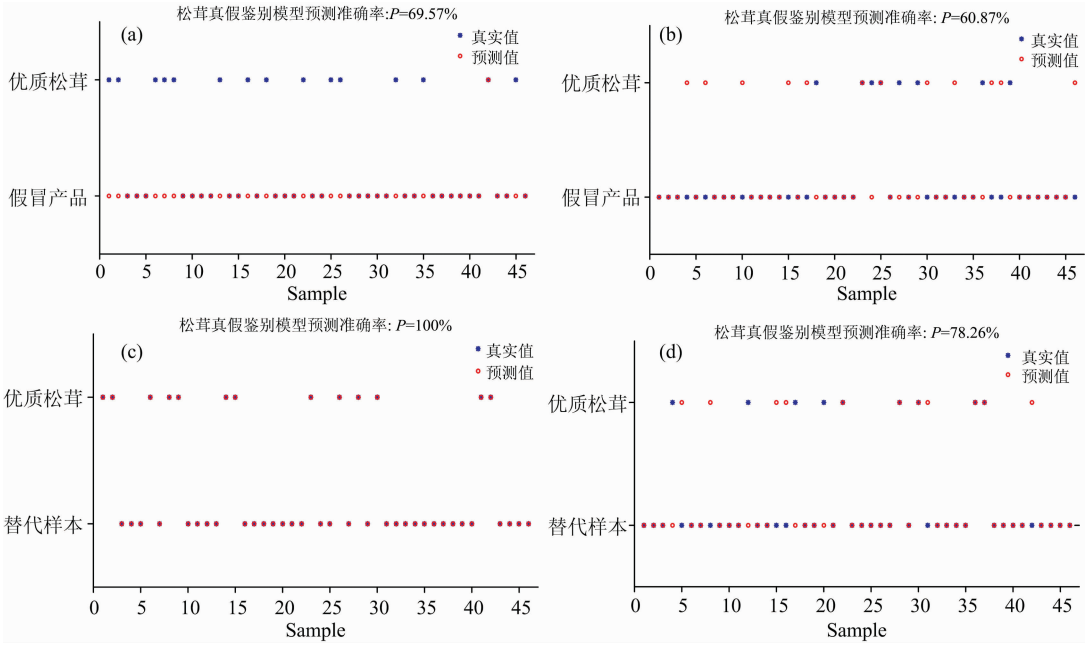

图 7 筛选变量前后 BP 和 KNN 模型预测结果对比

(a)：基于原始数据的 BP 模型预测结果(69.57%)；(b)：基于原始数据的 KNN 模型预测结果(60.87%)；
(c)：基于 LASSO 筛选特征的 BP 模型预测结果(100%)；(d)：基于 LASSO 筛选特征预 KNN 测结果(78.26%)

Fig. 7 Comparison between BP and KNN models before and after extraction feature

(a): Original data prediction results of BP model (69.57%); (b): Original data prediction results of KNN model (60.87%);
(c): Prediction results of BP model using LASSO algorithm (100%); (d): Prediction results of KNN model using LASSO algorithm (78.26%)

数，其柔性的网络结构使其具有很强的非线性映射能力，能够实现任何非线性逼近[12]。以两种模型的预测准确率为检测指标，分别进行建模分析，所得结果如图 7 所示。

通过对比图 7 的预测结果可知，基于 LASSO 特征筛选的建模变量均能提高 KNN 模型和 BP 模型的预测精度，但是基于原始数据和 LASSO 特征筛选的 KNN 模型预测性能不如 BP 模型，基于 LASSO 算法的 BP 模型预测准确率可达到 100%。

2.4　食用菌类别分析

将所采集的样本分为老人头、杏鲍菇、姬松茸、松茸优质、松茸劣质 5 类，分别讨论基于 LASSO 算法 KNN 和 BP 模型的预测准确率，图 8 为 BP 模型使用 LASSO 算法前后模型训练及预测散点图，图 9 为 LASSO 特征提取前后两种模型性能对比结果。如图8所示，在食用菌多分类问题中，模

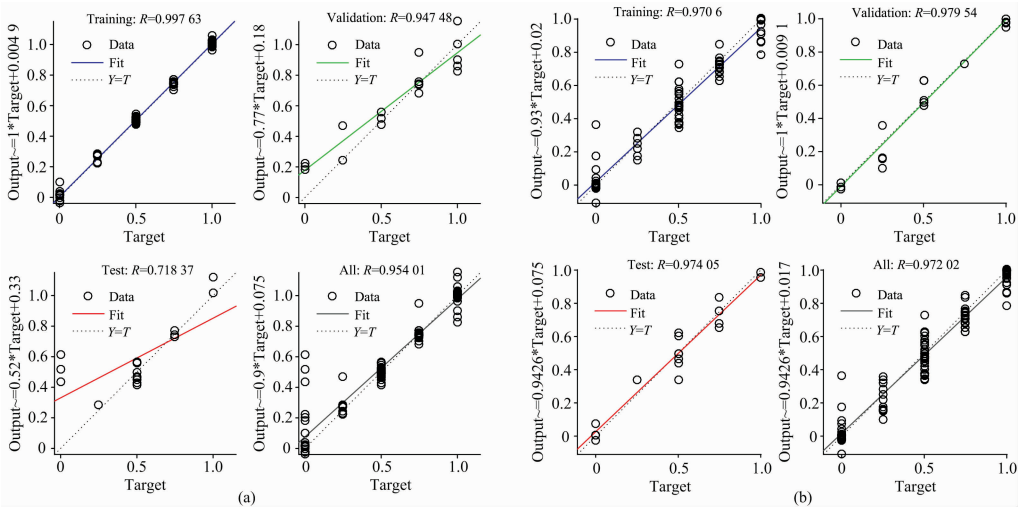

图 8　基于 SNV-LASSO 特征提取食用菌类别分析模型性能对比

(a)：原始数据的 BP 模型；(b)：基于 LASSO 筛选特征的 BP 模型

Fig. 8　Performance of edible fungus classification model using SNV-LASSO feature extraction

(a)：Traditional BP model；(b)：BP model based on LASSO algorithm

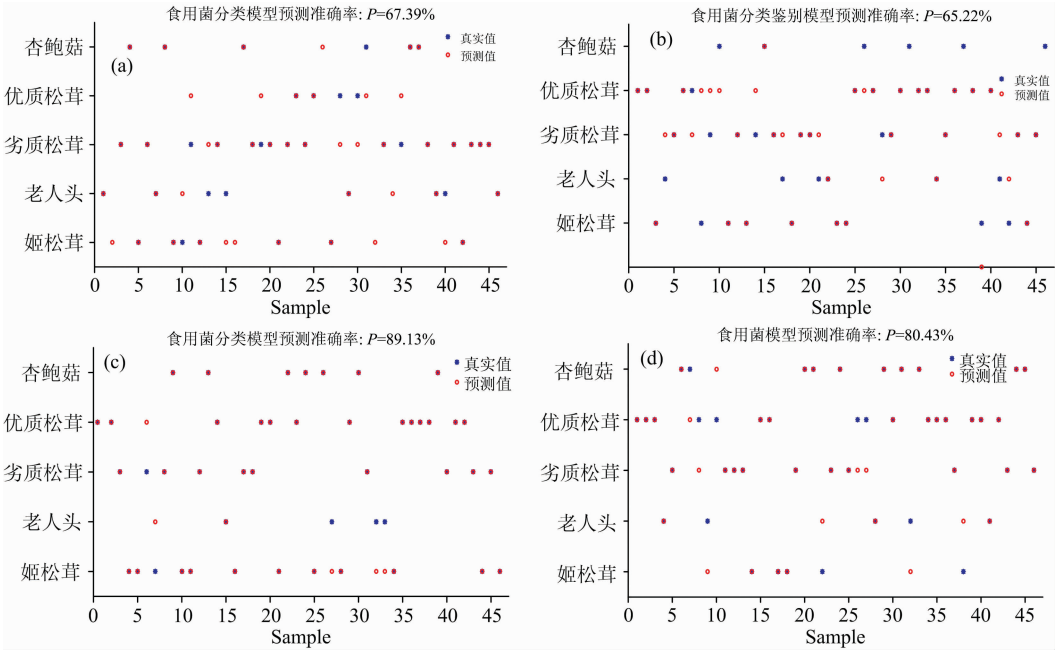

图 9　筛选变量前后 BP 和 KNN 模型预测结果对比

(a)：基于原始数据 BP 模型预测结果(67.39%)；(b)：基于原始数据 KNN 模型预测结果(65.22%)；

(c)：基于 LASSO 筛选特征 BP 模型预测结果(89.13%)；(d)：基于 LASSO 筛选特征 KNN 模型预测结果(80.43%)

Fig. 9　Comparison between BP and KNN models before and after extraction feature

(a)：Original data prediction results of BP model (67.39%)；(b)：Original data prediction results of KNN model (65.22%)；

(c)：Prediction results of BP model using LASSO algorithm (89.13%)；(d)：Prediction results of KNN model using LASSO algorithm (80.43%)

型使用 LASSO 特征变量选取前后预测准确率相差明显，传统检测方法的预测准确率为 71.83%，LASSO 特征选取后模型的预测准确率为 97.4%，LASSO 算法明显提高了 BP 模型食用菌类别分析性能。

如图 9 所示，通过对比应用 LASSO 算法前后 BP 模型及 KNN 模型的预测结果发现，基于传统分析方法所建模型的预测准确率分别是 67.39%(BP)，65.22%(KNN)，基于 LASSO 特征筛选变量所建模型预测准确率分别是 89.13%(BP)，80.43%(KNN)。由此可知，LASSO 算法能够提取全光谱区间高维光谱数据特征，降低建模数据维度的同时提高了 BP 模型及 KNN 模型的预测性能，实现了线性及非线性建模方法的高精度定性分析。

为验证 LASSO 算法对于高维光谱数据特征提取及数据降维的有效性，将 LASSO 算法与常规数据压缩算法 PCA 对比，采用盲样测试分析了两种算法特征数据筛选前后模型性能的变化[13]，结果如表 2 所示，实验结果表明 PCA 数据压缩对于高维数据处理能力有限，并且在建模数据维度较大时模型性能无明显提高，而 LASSO 算法能够有效筛选特征数据，所得模型特征筛选预测性能得到明显提高，能够实现二分类与多分类的高精度预测。

为分析基于 LASSO 算法聚类分析模型的稳定性，针对真伪甄别和食用菌类别分析模型分别采用蒙特卡罗方法进行多次试验分析[14]。两个模型均进行 10 次建模测试，分别记录各自模型每次计算的预测精度和 RMSE，结果如表 3 所示。结果表明，基于 LASSO 特征选择的松茸真伪甄别模型预测准确率平均值为 98.08%，食用菌类别分析模型的准确率平均值为 97.22%，基于 LASSO 特征筛选所建聚类分析模型稳定性好、预测精度高。

表 2 LASSO 与 PCA 算法建模性能对比

Table 2 Modeling performance comparison of LASSO and PCA algorithms

算法	二分类模型/%	多分类模型/%
PCA	78.09	69.96
LASSO	99.89	88.78

表 3 蒙特卡罗方法测量结果

Table 3 Results of Monte Carlo method

指标	真假鉴别				食用菌分类			
	训练集准确率/%	RMSE	预测集准确率/%	RMSE	训练集准确率/%	RMSE	预测集准确率/%	RMSE
最大值	99.99	0.008 6	99.98	0.170 1	98.81	0.034 0	99.75	0.100 0
最小值	99.86	0.001 4	90.45	0.001 0	96.63	0.010 0	93.94	0.004 4
平均值	99.93	0.004 1	98.08	0.042 2	97.71	0.022 1	97.22	0.034 3

因此，与传统建模方法相比，LASSO 算法能够实现高维光谱数据特征变量选择、降低建模数据维度，提高其聚类分析模型预测精度，对于近红外分析等建模数据维度较高领域，LASSO 算法是一种有效的特征提取方法。

3 结 论

采用 LASSO 进行光谱特征筛选，结合 BP 和 KNN 进行松茸聚类建模分析，研究了 LASSO 特征选择算法对于近红外定性分析的作用。通过对比变量筛选前后模型的预测性能，发现 LASSO 特征筛选算法可以实现全光谱区间内光谱特征选择，有效提高了光谱分析特征数据的代表性和完整性，与传统分析方法相比较，该算法能够实现高维数据特征筛选和高准确率聚类分析预测，并且基于所筛选特征变量所建模型预测性能良好、稳定性高，在近红外特征变量筛选和定性分析领域具有很大应用前景。

References

[1] WANG Meng-dong, WANG Sheng-peng(王梦东，王胜鹏). Journal of Huazhong Agricultural University(华中农业大学学报), 2015, 34(1): 123.

[2] Yahui L, Xiaobo Z, Tingting S, et al. Food Anal. Methods, 2017: 10: 1034.

[3] Balabin R M, Smirnov S V. Analytica Chimica Acta, 2011, 692(1): 63.

[4] WU Xi-yu, ZHU Shi-ping, WANG Qian, et al(吴习宇，祝诗平，王 谦，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2018, 38(8): 2369.

[5] Xu Y, Kutsanedzie F Y H, Sun H, et al. Food Anal. Methods, 2018, 11: 1199.

[6] LI Lu, HUANG Han-ying, LI Yi, et al(李 路，黄汉英，李 毅，等). Food and Fermentation Industries(食品与发酵工业), 2018, 44(2): 87.

[7] Saptoro A, Tadé M, Vuthaluru H. Chemical Product and Process Modeling, 2012, 7(1): 1.

[8] Wu X, Wu B, Sun J, et al. Journal of Food Process Engineering, 2017, 40(2): 23.

[9] LIU Pi-lian, WANG Xiao, LIU Mu-hua, et al(刘丕莲，王 晓，刘木华，等). Chinese Journal of Pesticide Science(农药学学报), 2014,

16(1): 106.
[10] Frank A F, Hlutkowsky C, Bemis L, et al. NeuroImage, 2019, 184: 68.
[11] Zhang Liguo, Zhang Xin, Ni Lijun, et al. Food Chemistry, 2014, 145: 342.
[12] Yosra A, Estrella Funes L Gabriel Beltran M, et al. Journal of Near Infrared Spectroscopy, 2015, 23(2): 111.
[13] Teye E, Huang X Y, Lei W, et al. Food Research International, 2014, 55: 288.
[14] Verleker A P, Shaffer M, Fang Q, et al. Applied Optics, 2017, 56(4): 1131.

NIR Spectral Feature Selection Using Lasso Method and Its Application in the Classification Analysis

LI Yu-qiang[1], PAN Tian-hong[1, 2*], LI Hao-ran[1], ZOU Xiao-bo[3]
1. School of Electrical and Information Engineering, Jiangsu University, Zhenjiang 212013, China
2. School of Electrical Engineering and Automation, Anhui University, Hefei 230061, China
3. School of Food and Biological Engineering, Jiangsu University, Zhenjiang 212013, China

Abstract Near-infrared spectroscopy (NIRS) is a non-destructive detection method for qualitative or quantitative analysis by using spectral feature data. The integrity and representativeness of feature data determine the performance of the analytical model. However, existing analytical methods can only extract the feature data from the spectral subinterval. Then the developed models using these feature extracting methods have poor stability. In order to extract the feature from the high-dimensional NIR spectral data and improve the accuracy and stability of NIR spectral model, a spectral screening method using the Least Absolute Shrinkage and Selection Operator (LASSO) algorithm is proposed in this paper. Furthermore, the Tricholoma Matsutake, one of the high-value foreign trade products in China is taken as example to validate the developed classified model using LASSO algorithm. The effectiveness of the feature screening algorithm for the high-dimensional spectral data is discussed, and predictive accuracy and stability of the Tricholoma Matsutake distinguished and edible fungus classified model using LASSO and PCA are also analyzed. It is well known that the fresh Tricholoma Matsutake has the unique shape and it is easy to distinguish its counterfeit. However, it is difficult to distinguish the dry Tricholoma Matsutake from other mushrooms because all of dry mushrooms have the similar flake shape. As a result, dry Tricholoma Matsutake adulteration incidents have occurred frequently. 166 dry samples of Yunnan Tricholoma Matsutake, Pleurotuseryngii, Jujube hilt nipple mushroom and Agaricusblazei were selected in this experiment, and 166×512-dimensional raw spectral data were obtained by NIRQuest 512 NIR spectrometer with a spectral range of 900～1 700 nm. The standard normal transformation (SNV) was taken to pre-process the spectral data after the anomalous data eliminating. The LASSO was used to extract feature variables from the high-dimensional NIR spectral data based on the spectral pretreatment. Then the typical linear (k-Nearest Neighbor, KNN) and the nonlinear modeling (Back-Propagation neural network, BP) algorithms combined with the Kennard-Stone method were used to construct the Tricholoma Matsutake distinguished and edible fungus classified model. The effectiveness of models using LASSO and PCA were also analyzed. Furthermore, the predictive accuracy and the stability of the developed KNN model and BP model were analyzed by using the Monte Carlo method. The experimental results demonstrated that the prediction accuracy and stability of model using LASSO were better than those of the model using PCA. The prediction accuracy of the distinguished and edible fungus classified models using the original spectral data were 69.57% (BP), 60.87%(KNN) and 67.39% (BP), 65.22% (KNN) respectively. And the prediction accuracy of the distinguished and edible fungus classified models using LASSO algorithm were up to 100% (BP), 67.39% (KNN) and 89.13% (BP), 80.43% (KNN) respectively. The two models were performed by 10 times Monte Carlo method and the average results were 99.93% and 97.22%, respectively. Compared with the conventional feature selection methods (such as PCA), the LASSO algorithm can extract the feature from the high-dimensional NIR spectral data. And the accuracy and stability of the models using NIR spectral data can be improved. Furthermore, the developed algorithm is alternative to be a new feature extraction method for NIR spectral data analysis.

Keywords NIRS; Feature extraction; LASSO; Tricholoma Matsutake discrimination; Monte Carlo method

* Corresponding author

(Received Oct. 31, 2018; accepted Feb. 10, 2019)