特征分层结合改进粒子群算法的近红外光谱特征选择方法研究

徐宝鼎[1]，秦玉华[2]，杨 宁[1]，高 锐[3*]，苑程程[1]

1. 中国海洋大学信息科学与工程学院，山东 青岛 266100
2. 青岛科技大学信息科学技术学院，山东 青岛 266061
3. 云南中烟工业有限责任公司技术中心，云南 昆明 650024

摘 要 在近红外光谱数据定量建模中，数据的高冗余和高噪严重影响了建模的稳健性和精确性，因此提出了一种特征分层结合改进粒子群算法(PSO)的特征光谱选择方法。首先通过互信息度量特征的重要性得分，并按特征的重要性降序排序，有效避免了因采用降维方法得到主成分而引起的丢失重要信息的问题。其次，引入了跳跃度概念，并构造了一种特征分层的方法，重要性程度相似的特征并入同一个特征子集，将降序排列的特征集分割为不同的特征子集，避免了筛选特征过程中因人为设定特征重要性得分阈值而导致的不确定性。最后，采用收敛速度快、控制参数少的粒子群算法作为最优特征子集的优化方法，同时对粒子群算法做了两方面改进：引入混沌模型增加种群的多样性，提高了PSO的全局搜索能力，避免陷入局部最优；将特征数目引入到适应度函数中，在迭代前期通过惩罚因子调节特征数目对适应度函数的影响，提高了算法的适应能力。将分层后的数据以特征子集为单位，依次累加并作为改进粒子群算法的输入，从而选择出高辨别力的特征子集。以烟碱指标为例进行了特征选择过程的描述，实验采用尼高力公司的Antaris Ⅱ近红外光谱仪进行近红外光谱数据的采集，光谱扫描范围为4 000～10 000 cm^{-1}。首先，利用互信息理论计算全光谱1 557个特征对待测指标定量建模的重要性得分，得分取30次实验的均值。其次，将所有特征按照重要性得分降序排序，计算所有特征的跳跃度，依据跳跃度寻找特征分层的临界点，将特征划分到不同的特征层中，构建了包含8个特征子集的特征集合$S=\{S'_1, S'_2, S'_3, S'_4, S'_5, S'_6, S'_7, S'_8\}$。然后，依次将特征子集$S'_1$，$\{S'_1, S'_2\}$，$\{S'_1, S'_2, S'_3\}$，…，$\{S'_1, S'_2, S'_3, S'_4, S'_5, S'_6, S'_7, S'_8\}$作为初始粒子群的候选集，以$R/(1+\mathrm{RMSEP})$作为特征子集优劣的评价标准，各自重复实验50次，比值最大的特征子集即为最优特征子集。为验证该算法的有效性，选取了具有代表性烟叶近红外光谱数据作为训练集和测试集，建立了烟碱、总糖两个指标的PLS定量模型，并分别与全光谱、分层后的特征光谱、粒子群算法选出的特征光谱进行了比较。仿真结果表明，本算法所选特征烟碱、总糖的建模相关系数r分别为0.988 5和0.982 2，交互验证均方差RMSECV分别为0.098 4和0.889 3，预测均方根误差RMSEP分别为0.100 7和0.901 6，模型准确率均明显高于其他三种方法。从所选特征数来看，该算法所选特征数最少，有效剔除了原特征集中的弱相关和噪声、冗余信息，所建模型的主因子数最少，降低了模型的复杂性，模型更加稳健，适应性更广。

关键词 特征选择；特征分层；跳跃度；改进粒子群算法；近红外光谱

中图分类号：O657.3 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2019)03-0717-06

引 言

近红外(NIR)光谱分析是化学计量学与光谱测量结合的绿色分析技术，以其无损、高效、简便、高通量等特点越来越引人注目，并在烟草、石油、食品等领域得到广泛应用[1]。在烟草行业，近红外技术主要用于定性识别不同类型、风格的烟叶和定量分析烟草中的常规化学成分[2]。但在实际应用中，近红外光谱具有高维、小样本特点，样本数目一般为几十或上百，而特征波长往往高达几千维，其中含有大量与预测无关的冗余信息和噪声特征[3]。冗余信息和噪声等因素不但影响建模的效率，而且还会降低模型的稳健性和预测精

收稿日期：2018-01-18，**修订日期**：2018-05-11

基金项目：国家重点研发计划项目(2016YFB1001103)，云南中烟工业有限责任公司项目(2017XX02，2018JC01)资助

作者简介：徐宝鼎，1990年生，中国海洋大学信息科学与工程学院硕士研究生 e-mail：xbd991@163.com

*通讯联系人 e-mail：gaorui177@163.com

度。因此，如何从上千维特征中筛选与建模密切相关的特征就显得尤为关键。束茹欣等[4]先用 PCA 对数据降维，再用支持向量机(SVM)进行分类识别，构建烟叶样品产地识别模型。但由于 PCA 是一种以最大方差理论为依据的降维方法，方差小的主成分也可能含有对建模有用的信息，因降维丢弃可能影响预测精度。陈孝敬等[5]利用模拟退火(SA)算法进行特征筛选，用最小二乘支持向量机(LS-SVM)作为识别器，对三种品牌润滑油进行了识别。但模拟退火算法本身全局搜索能力不足，易陷入局部最优。邹小波等[6]将整个光谱等分为 N 个特征子区间，对子区间利用遗传算法进行波长选择，并构建了 PLS 定量模型，取得了一定效果，但特征子区间的选择导致部分相关性强的特征变量未被选择，从而影响模型的适用性和准确性。

针对上述问题，提出了特征分层结合改进粒子群算法[7]的特征光谱选择方法。首先通过互信息[8]度量特征的重要性得分，有效避免了因采用降维方法得到主成分而引起的丢失重要信息的问题。其次，为了避免在筛选特征过程中因人为设定特征重要性得分的阈值而导致的不确定性，引入了跳跃度的概念并构造了一种特征分层的方法，将重要性程度相似的特征并入同一个特征子集。最后，采用收敛速度快、控制参数少的粒子群算法作为最优特征子集的优化方法，同时为避免陷入局部最优，提高算法的适应能力，通过引入混沌模型和将特征数目引入到适应度函数中对粒子群算法进行了改进，从而得出最优特征子集。

1 特征分层结合改进粒子群算法的近红外光谱特征选择算法

1.1 互信息

互信息(mutual information)用来衡量两个变量之间关系的强弱程度，它表示一个随机变量中包含的关于另一个随机变量的信息量。给定两个随机变量 X 和 Y，它们之间的互信息 $I(X;Y)$定义为

$$I(X;Y)=\sum_{x\in X}\sum_{y\in Y}p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{1}$$

其中 X 和 Y 为设定的两个随机变量，$p(x,y)$为随机变量 X 和 Y 的联合概率分布。由定义知，当变量 X 和 Y 相互独立或者完全无关时，X 与 Y 的互信息最小为 0，这说明 X 和 Y 变量之间不存在相同的信息；反之，它们相互依赖程度越高，互信息 $I(X;Y)$的值就越大，所包含相同的信息也越多。

由于实际应用中数据的真实概率分布通常未知，所以计算互信息之前须先近似估计随机变量的概率密度分布。利用高斯核函数的性质近似估计变量的概率密度分布

$$\hat{p}(x)=\frac{1}{n}\sum_{i=1}^{n}G(x-x_i,\Sigma) \tag{2}$$

其中 $G(z,\Sigma)$为高斯核函数，即

$$G(z,\Sigma)=\frac{1}{(2\pi h)^{m/2}\mid\Sigma\mid^{1/2}}\exp\left(-\frac{z^{\mathrm{T}}\Sigma^{-1}z}{2h^2}\right) \tag{3}$$

变量 X 和 Y 的联合概率分布为

$$\hat{p}(x,y)=\left(\frac{1}{n}\right)^2\sum_{i=1}^{n}\sum_{j=1}^{n}G(xi-yj,\Sigma_1+\Sigma_2) \tag{4}$$

其中 x_i 为第 i 个样本数据，n 为样本数，m 为样本维度，Σ 为协方差矩阵。Σ_i 为第 i 个变量的协方差矩阵，它们经常取相同值。

1.2 特征分层方法

因后续步骤要采用粒子群算法对重要性评分较高的特征子集进一步优化，而重要性评分的高低是一个模糊的概念，现有的方法需要相关工作人员根据经验人工设定特征重要性评分阈值，以筛选掉大量的无关特征(重要性得分低的特征)。这种方法不仅要求工作人员具有相当丰富的经验，而且效果很难达到最优。因此引入跳跃度的概念构造一种特征分层方法，对特征集进行分层，重要程度相似的特征并入同一层。

文献[9]给出跳跃度的概念：设 $X_{(1)}$，$X_{(2)}$，…，$X_{(n-1)}$，$X_{(n)}$为来自总体分布 $F(x;\theta)$的样本容量为 n 的次序统计量，$\hat{u}k$ 为仅依赖与 $X_{(1)}$，…，$X_{(k)}$的期望 u 的点估计，则称$\frac{\hat{u}k+1}{\hat{u}k}$为 $\hat{u}$ 在 k 点的跳跃度。

特征集 S'按照重要性得分降序排列($v_1>v_2>v_3>\cdots>v_n$)，其中 $v_n(n\in N)$为对应特征的重要性得分值，利用公式 $\hat{u}k=\frac{\sum_{i-1}^{k}v_i+(n-k)v_k}{k}$ 计算 $t_k=\frac{\hat{u}_{k+1}}{u_k}$。若 $t_k((k\geqslant 2)$小于前面$((k-1)$个特征点的跳跃度，则 v_k 为不同层之间的临界点，将之前$(k-1)$个特征作为一个特征层 U_1；将去掉 U_1 后的剩余特征重复上述步骤，计算出所有特征点的跳跃度，直到所有特征分层为止。

1.3 改进粒子群算法

粒子群算法(particle swarm optimization，PSO)是 Kennedy 和 Eberhart 受人工生命研究结果的启发、通过模拟鸟群觅食过程中的迁徙和群聚行为而提出的一种基于群体智能的全局随机搜索算法[10]。主要用来解决数据的组合优化问题。但是由于粒子群算法离散数据的优化处理不佳，1997 年 Eberhart 与 Kennedy 又提出了离散二进制粒子群算法(BPSO)用来解决离散数据的全局优化问题[11]，BPSO 算法简单、高效，但是存在易陷入局部最优等缺点。针对这些缺点作出了改进并进行了特征选择。

D 维空间中，有 m 个粒子，其中第 i 个粒子表示为一个 D 维向量 $x_i=(x_{i1},x_{i2},\cdots,x_{iD})^{\mathrm{T}}$，$i=1,2,\cdots,m$，即 x_i 为第 i 个粒子在 D 维空间中的位置(坐标)，每一个粒子的位置都可能是全局最优解。将 x_i 代入适应度函数计算出其适应度，根据适应度的大小衡量 x_i 的优劣。第 i 个粒子的速度也是一个 D 维的向量，记为 $v_i=(v_{i1},v_{i2},\cdots,v_{iD})^{\mathrm{T}}$。记第 i 个粒子迄今为止搜索到的最优位置为 $P_g=(p_{g1},p_{g2},\cdots,p_{gD})^{\mathrm{T}}$。那么根据式(5)和式(6)更新粒子的速度与位置

$$v_{id}^{n+1}=\omega v_{id}^{n}+c_1r_1(p_{id}^{n}-x_{id}^{n})+c_2r_2(p_{gd}^{n}-x_{id}^{n}) \tag{5}$$

$$x_{id}^{n+1}=x_{id}^{n}+\eta v_{id}^{n+1} \tag{6}$$

其中 ω 为惯性权重；c_1 和 c_2 是学习因子或加速常数，一般为正常数，通常均设置为 2；r_1 和 r_2 是[0，1]之间的随机数；η 为约束因子，通常设置为 1；$n=1,2,\cdots$为迭代次数。

上述算法只适用于连续问题的求解，而对于离散问题则

需要用离散二进制粒子群算法[12-13]来求解，即 BPSO 算法：粒子速度的更新方式不变，将粒子位置的每一维分量仅用 0 或 1 表示(特征选择时，0 代表特征不选择，1 代表特征被选择)。根据速度的 sigmoid 函数控制粒子的位置更新

$$\text{Sigmoid}(t)=\frac{1}{1+\mathrm{e}^{-t}}$$

$$\begin{cases} x_{id}=1, & \text{rand}<\text{Sigmoid}(v_{id}) \\ x_{id}=0, & \text{otherwise} \end{cases} \tag{7}$$

对 BPSO 算法改进如下：

(1)引入混沌模型增加种群的多样性[14]。传统粒子群算法随机初始化种群，即 $x_{id}=\text{rand}>0.5$，缺点是可能种群分布单一而导致粒子陷入局部最优。为了增加种群的多样性，避免粒子群陷入局部最优，本文在初始化种群的过程中引入了混沌模型，初始化过程为

$$c_r=\text{rand}$$

$$x_{id}=\begin{cases} 4c_r(1-c_r)>0.5 & \text{otherwise} \\ (c_r+0.1\text{rand})>0.5 & c_r=\{0,\ 0.25,\ 0.5,\ 0.75\} \\ (c_r-0.1\text{rand})>0.5 & c_r=1 \end{cases} \tag{8}$$

(2)改进适应度函数。传统适应度函数设为

$$F=R/(1+\text{RMSEP}) \tag{9}$$

其中 R 为采用 PLS 算法建模交互验证的相关系数，RMSEP 为预测均方根误差。

但是，在解决实际问题中，如果两个特征子集 $R/(1+\text{RMSEP})$ 相同，则特征数目少的特征子集应该被选中。所以适应度函数要综合考虑 $R/(1+\text{RMSEP})$ 和特征维数(S)，在构造适应度函数时，以特征维数的倒数为惩罚项，引入 α 作为 $R/(1+\text{RMSEP})$ 和特征维数之间的平衡因子。改进后的适应度函数如下

$$F=(1-\alpha)\frac{R}{1+\text{RMSEP}}+\frac{\alpha}{S}$$

其中

$$\alpha=\begin{cases} t/T, & t<T/2 \\ 0.5, & t\geqslant T/2 \end{cases} \tag{10}$$

其中，t 为当前迭代次数；T 为迭代次数。

1.4　算法步骤

Step1：基于互信息理论计算特征 $x(x\in S_N)$ 与预测结果 y 的重要性得分(特征与预测结果的相关性程度)$v_n(n\in N)$ 并按得分降序排序，得到排序后的特征集 S'。其中 N 为特征总数，S_N 为全体特征集合。

Step2：将得到的排序后的特征集 S' 利用特征分层方法对特征进行分层。引入跳跃度作为临界点的评定指标，将重要性程度相近的特征归为同一个特征子集，避免在筛选特征过程中因人为设定特征重要性得分的阈值而导致的不确定性。得到分组后的特征集合可以表示为：$S'=\{S'_1, S'_2, \cdots, S'_k\}$。

Step3：将 Step2 中得到的分组特征子集 $\{S'_1, S'_2, \cdots, S'_k\}$ 依次累加作为改进后的二进制粒子群算法(BPSO)的输入，得到最优特征子集。具体先将特征重要性得分最高的特征子集 $S'1$ 作为改进后二进制粒子群算法输入，之后再依次累加一个重要性得分最高的特征子集，直到特征集为 S'，得到 K 个特征子集，以此作为最优特征子集的候选集。

Step4：用 Step3 中得到的 K 个最优特征子集的候选集分别建立待测指标的 PLS 定量模型，以 $R/(1+\text{RMSEP})$ 作为特征子集优劣的评价标准，选出比值最大的特征子集，即为最优特征子集。

2　实验部分

2.1　数据来源

为保证实验的有效性，采用某卷烟企业提供的具有代表性的 268 个烟叶样品，随机选取其中的 201 个样品作为训练集，剩余 67 个样品作为测试集。本实验烟叶的化学成分指标含量由权威机构根据常规化学分析方法测定。

2.2　近红外光谱数据采集和光谱预处理

外光谱数据采集选用尼高力公司的 Antaris Ⅱ近红外光谱仪，光谱扫描范围为 4 000～10 000 cm^{-1}。将样品放置在 60 ℃烘箱烘干 4 h，用旋风磨粉碎，过 40 目筛，样品常温下避光储存放在密封袋中。每个样品称重 20 g，将样品放置于干净的样品杯(小杯直径 5 cm)中，用压样器压实样品，放入近红外光谱仪中扫描，实验室温度在 18～25 ℃之间、湿度＜60%，采用漫反射方式，扫描样品杯底部 7 个不同位置光谱取平均值。为降低样品均匀性不一致等因素的影响，每个样品均重复装样扫描三次，计算三次扫描的平均值作为该样品光谱。

采用二阶导数加 Norris(11)点平滑预处理方法，以消除基线漂移、仪器随机误差和背景噪声等干扰对光谱数据的影响。

3　结果与讨论

为了验证本算法的有效性，选取了烟碱、总糖两个指标分别进行实验。由于烟碱、总糖的特征选择过程相似，以烟碱指标为例进行特征选择过程的描述。

3.1　特征重要性得分

选取4 000～10 000 cm^{-1} 作为建模谱段。首先，利用互

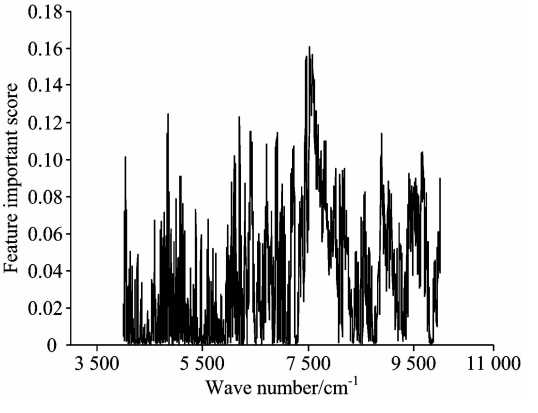

图 1　特征重要性得分图

Fig. 1　The importance score of the features

信息理论根据式(1)计算全光谱 1 557 个特征对待测指标定量建模的重要性得分。重复计算 30 次，重要性得分取 30 次的均值，所有特征重要性得分如图 1 所示。通常重要性得分越高的特征对定量模型的构建越重要。

3.2 特征分层

将所有特征按照重要性得分降序排序，计算所有特征的跳跃度，依据跳跃度寻找特征分层的临界点，将特征划分到不同的特征层中，构建不同的特征子集。分层结果如图 2 所示，图中黑圈代表了层与层之间的临界点。

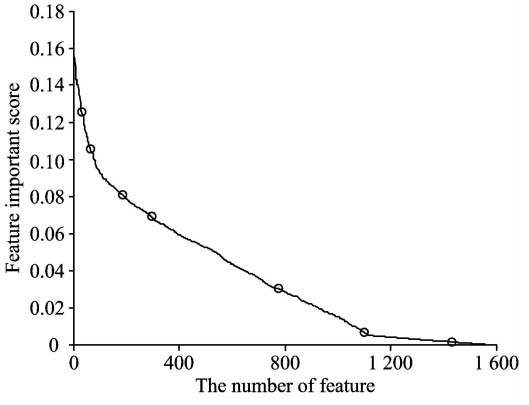

图 2 特征分层图

Fig. 2 Feature stratification diagram

由图 2 得知，利用分层法把所有特征划分为 8 个特征子集，特征重要性得分最高的子集记为 S'_1，次高的记为 S'_2，最低的记为 S'_8。可得所有特征集合为 $S=\{S'_1, S'_2, S'_3, S'_4, S'_5, S'_6, S'_7, S'_8\}$。

3.3 最优特征子集

最后利用改进粒子群算法进行最优特征子集的选择，具体步骤如下：

依次将特征子集 S'_1，$\{S'_1, S'_2\}$，$\{S'_1, S'_2, S'_3\}$，…，$\{S'_1, S'_2, S'_3, S'_4, S'_5, S'_6, S'_7, S'_8\}$作为初始粒子群的候选集，粒子群个数为 8，学习因子为 $c_1=c_2=2$，ω 取值为 0.1～0.6，线性递减。适应度函数为式(10)，迭代次数 $K=100$。本实验以 $F=R/(1+\text{RMSEP})$作为评价特征子集优劣的标准，F 越大则所筛选出的特征子集越优。8 个粒子群各自重复实验 50 次，F 值为 50 次的均值，即 $F_i=\frac{1}{50}\sum_{j=1}^{50}F_{ij}$。其中 i 为粒子群的标号，j 为实验次数。

所有粒子群的 F 值如表 1 所示。

由表 1 得知，随着重要性得分较高的样品特征子集依次加入，F 不断变大，直到样品的特征子集为$\{S'_1, S'_2, S'_3, S'_4\}$时，$F$ 达到最大，当特征子集继续增加时，F 随着特征子集的增加而减小。由此可见，随着重要性得分相对较低的特征子集加入，光谱数据的噪声特征和冗余信息随之增加，从而导致 F 减小。所以，特征子集$\{S'_1, S'_2, S'_3, S'_4\}$经 BPSO 筛选后得到的 143 个特征波长即为最优特征子集。

3.4 性能对比

为验证本算法的有效性，分别以全光谱、粒子群算法筛选出的特征集合、特征子集$\{S'_1, S'_2, S'_3, S'_4\}$(烟碱)、$\{S'_1, S'_2, S'_3\}$(总糖)和本文算法筛选出的最优特征集合，构建烟碱、总糖赘标的 PLS 定量模型。以相关系数(r)、交互验证均方差(RMSECV)和预测均方根差(RMSEP)作为模型的评价指标，结果如表 2 和表 3 所示。

表 1 特征子集与 F 关系表

Table 1 The relationship of feature subsets and F

特征子集	BPSO 筛选前特征数	BPSO 筛选后特征数	F
$\{S'_1\}$	31	16	0.753 5
$\{S'_1, S'_2\}$	63	24	0.812 7
$\{S'_1, S'_2, S'_3\}$	186	102	0.871 4
$\{S'_1, S'_2, S'_3, S'_4\}$	294	143	0.898 1
$\{S'_1, S'_2, S'_3, S'_4, S'_5\}$	776	316	0.884 2
$\{S'_1, S'_2, S'_3, S'_4, S'_5, S'_6\}$	1 099	584	0.859 7
$\{S'_1, S'_2, S'_3, S'_4, S'_5, S'_6, S'_7\}$	1 431	847	0.837 4
$\{S'_1, S'_2, S'_3, S'_4, S'_5, S'_6, S'_7, S'_8\}$	1 557	882	0.835 3

表 2 不同特征子集烟碱模型性能对比

Table 2 Performance contrast of nicotine model with different feature subsets

特征光谱子集	因子数	特征个数	r	RMSECV	RMSEP
全光谱	7	1 557	0.958 2	0.133 5	0.147 1
BPSO	7	882	0.951 4	0.131 2	0.145 3
$\{S'_1, S'_2, S'_3, S'_4\}$	6	294	0.982 3	0.112 5	0.129 1
本文算法	6	143	0.988 5	0.098 4	0.100 7

表 3 不同特征子集总糖模型性能对比

Table 3 Performance contrast of total sugar model with different feature subsets

特征波长子集	因子数	特征个数	r	RMSECV	RMSEP
全光谱	6	1 557	0.953 5	1.123 4	1.136 2
BPSO	6	743	0.956 7	1.125 7	1.137 3
$\{S'_1, S'_2, S'_3\}$	5	182	0.977 5	0.988 1	0.993 0
本文算法	5	159	0.982 2	0.889 3	0.901 6

由表 2 和表 3 可以看出，本算法所选特征建模相关系数 r 最高，RMSECV 和 RMSEP 最低，模型准确性均较其他方法高。从所选特征数来看，本算法所选特征数最少，有效剔除了原特征集中的弱相关和噪声、冗余信息，所建模型的主因子数最少，降低了模型的复杂性，模型更加稳健，适应性更广。

4 结 论

特征分层结合改进二进制粒子群算法的近红外光谱波长选择算法有效的提高了模型的稳健性和预测精度。该算法融合了互信息的高效与改进后粒子群算法全局搜索的优点，有

效选出了相关性高的建模特征。同时，通过对特征集分层，依次累加作为改进粒子群的输入，避免了依靠经验设置重要性得分阈值，为粒子群的初始化提供了质量保证。实验结果表明，本文算法所建模型精确性高，稳健性好，因子数少，模型更加简单。该方法可为烟草质量检测与质量分析领域提供科学依据，同时对近红外定量分析模型的建立具有普遍的参考意义。

References

[1] YUAN Tian-jun, WANG Jia-jun, ZHE Wei, et al(袁天军, 王家俊, 者　为, 等). Chinese Agricultural Science Bulletin(中国农学通报), 2013, 29(20): 190.

[2] QIU Jun, ZHANG Huan-bao, SONG Yan, et al(邱　军, 张怀宝, 宋　岩, 等). Chinese Tobacco Science(中国烟草科学), 2008, 29(1): 55.

[3] QIN Yu-hua, DING Xiang-qian, GONG Hui-li(秦玉华, 丁香乾, 宫会丽). Infrared and Laser Engineering(红外与激光工程), 2013, 42(5): 1355.

[4] SHU Ru-xin, SUN Ping, YANG Kai, et al(束茹欣, 孙　平, 杨　凯, 等). Tobacco Science and Technology(烟草科技), 2011, (11): 50.

[5] CHEN Xiao-jing, WU Di, YU Jia-jia, et al(陈孝敬, 吴　迪, 虞佳佳, 等). Acta Optica Sinica(光学学报), 2008, 28(11): 2154.

[6] ZOU Xiao-bo, ZHAO Jie-wen (邹小波, 赵杰文). Acta Optica Sinica(光学学报), 2007, 27(7): 1316.

[7] LIU Xin, YU Sui-huai, CHU Jian-jie, et al(刘　昕, 余隋怀, 初建杰, 等). Computer Engineering and Applications(计算机工程与应用), 2015, 51(7): 1.

[8] TANG Shi-wei, LIU Xian-mei(唐世伟, 刘贤梅). Information Theory(信息论). Harbin: Harbin Engineering University Press(哈尔滨: 哈尔滨工业大学出版社), 2009.

[9] ZHANG De-ran(张德然). Statisitical Research(统计研究), 2003, (5): 53.

[10] Kennedy J, Eberhart R C. International Conference On Neural Networks, 1995. 1942.

[11] Kennedy J, Eberhart R C. International Conference On Systems, Man, And Cybernetics, 1997. 4104.

[12] YANG Rui-qing, LIU Guang-yuan(杨瑞请, 刘光远). Computer Science(计算机科学), 2008, 35(3): 137.

[13] Mahdiyeh Eslami, Hussain Shareef, Azah Mohamed. Journal of Central South University of Technology, 2011, 18: 1579.

[14] LI Ce, WANG Bao-yun, GAO Hao(李　策, 王保云, 高　浩). Computer Technology and Development(计算机技术与发展), 2017, 27(4): 89.

Study on Feature Selection of Near Infrared Spectra Based on Feature Hierarchical Combining Improved Particle Swarm Optimization

XU Bao-ding[1], QIN Yu-hua[2], YANG Ning[1], GAO Rui[3*], YUAN Cheng-cheng[1]

1. College of Information Science and Engineering, Ocean University of China, Qingdao　266100, China

2. College of Information Science and Technology, Qingdao University of Science and Technology, Qingdao　266061, China

3. China Tobacco Yunnan Industrial Co., Ltd., Technical Research Center, Kunming　650024, China

Abstract　In the quantitative modeling of near-infrared spectroscopy data, the high redundancy and high noise of the data severely affect the robustness and accuracy of the modeling. Therefore, this paper presents a feature-based spectroscopy combined with improved Particle Swarm Optimization (PSO) Method of choosing. First, we measure the importance score of each feature through mutual information, and then sort the features according to the importance of the features in descending order. This effectively avoids the problem of losing important information caused by using the principal component reduction method. Secondly, the concept of jump degree is introduced and a method of feature stratification is constructed. Similar features of similar importance are merged into the same feature subset, and the descending ordered feature set is segmented into different feature subsets, avoiding the screening uncertainty caused by artificially setting the score of feature importance score during feature process. Finally, the particle swarm optimization algorithm with fast convergence rate and few control parameters is used as the optimal feature subset optimization method. At the same time, particle swarm optimization is improved in two aspects: The chaotic model is introduced to increase the diversity of the population and improve the global searching ability of PSO, so as to avoid getting into local optimum. The number of features is introduced into the fitness function, and the influence of the number of features on the fitness function is adjusted by the penalty factor in the early iteration to improve the adaptability of the algorithm. The strati-

fied data is collected as a feature subset and then added as a modified particle swarm optimization algorithm to select the high-resolution feature subset. In this paper, the nicotine index as an example of the feature selection process is described, using Nicolet company Antaris II near infrared spectrometer near infrared spectrum data acquisition, spectrum scanning range is 4 000～10 000 cm^{-1}. First, we use the mutual information theory to calculate the importance score of 1 557 features of the whole spectrum on the quantitative modeling of the index to be measured, and take the average of 30 experiments. Secondly, all the features are sorted in descending order of importance scores to calculate the jumping degree of all the features. According to the jumping degree, the critical points of the feature stratification are searched, and the features are divided into different feature layers to construct a feature containing 8 feature subsets set $S=\{S'_1, S'_2, S'_3, S'_4, S'_5, S'_6, S'_7, S'_8\}$. Then, the feature subset is in turn $\{S'_1\}$, $\{S'_1, S'_2\}$, $\{S'_1, S'_2, S'_3\}$, …, $\{S'_1, S'_2, S'_3, S'_4, S'_5, S'_6, S'_7, S'_8\}$ as a candidate for initial particle swarm. With $R/(1+\mathrm{RMSEP})$ as the evaluation criteria of the pros and cons of feature subsets, each iterative experiment 50 times, the ratio of the largest feature subset is the optimal feature subset. In order to verify the effectiveness of this algorithm, we select representative tobacco near-infrared spectral data as a training set and a test set, establish a PLS quantitative model of nicotine and total sugar, and compare with the full-spectrum, stratified characteristic spectrum, particle swarm algorithm selected by the characteristic spectra. The simulation results show that the modeling correlation coefficients R of nicotine and total sugar selected by this algorithm are respectively 0.988 5 and 0.982 2, RMSECV of mutual verification are 0.098 4 and 0.889 3 respectively, RMSEP of prediction root mean square error are 0.901 6 and 0.100 7 respectively, Accuracy are significantly higher than the other three methods. From the selected number of features, the proposed algorithm has the least number of selected features, effectively eliminating the weak correlation and noise and redundant information in the original feature set, minimizing the number of main factors of the model and reducing the complexity of the model, and the model is steadier, more adaptable.

Keywords Feature selection; Feature stratification; Jumping degree; Improved particle swarm optimization; Near infrared spectroscopy

(Received Jan. 18, 2018; accepted May 11, 2018)

* Corresponding author