# 改进和声搜索算法的近红外光谱特征变量选择

张 磊[1]，丁香乾[1]，宫会丽[1]，吴丽君[2*]，白晓莉[2]，罗 林[2]

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

**摘 要** 近红外光谱分析以其简便、快速、高效、低成本、绿色环保等优点，已广泛应用于诸多领域。然而，近红外光谱同时存在变量维度高、多重共线性、包含冗余信息和高频噪声等问题，直接构建预测模型不但增加建模复杂度，同时也会影响模型的预测性能和泛化能力，因此提出一种基于改进和声搜索算法(HS)的光谱特征变量选择方法。HS常用于解决特征变量优化选择问题。在应用和声搜索算法进行最优光谱变量选择时，首先通过偏最小二乘(PLS)载荷系数计算各光谱点的特征贡献度，作为和声搜索算法改进的扰动权重。算法优选光谱特征变量过程中，引入变量特征贡献度作为激励因子，采用随机遍历和激励因子共同作用的方式生成初始解向量。产生新和声向量时，应用变量特征贡献度作为惩罚项，通过加入平衡因子使选择参数随迭代次数而动态调整，从而适应光谱变量的搜索，增强搜索过程的遍历性和种群的多样性。为验证本算法的有效性，以烟叶样品烟碱、总糖、总氮三个指标的近红外光谱PLS建模应用为例，对采集的原始光谱进行预处理后，应用该方法对光谱变量进行优选，根据变量被选择的累积频次分别计算不同变量个数的模型预测性能，通过校正均方根误差(RMSEC)随变量增加的变化趋势确定最终选择的光谱特征变量。在训练集上分别建立各指标的PLS模型，应用测试集测试模型性能，并与全光谱、无信息变量消除法(UVE)和粒子群算法(PSO)进行比较。实验结果显示，应用该算法所选变量建立的烟碱、总糖和总氮三个模型的决定系数($R^2$)分别为0.921 1，0.925 7和0.941 2，预测均方根误差(RMSEP)分别为0.102 3，1.034 6和0.053 1，与其他方法相比，光谱特征变量更少，同时$R^2$和RMSEP值更优。由此表明，改进的和声搜索算法能有效筛选特征光谱，降低建模复杂度，提升模型预测性能和泛化能力。

**关键词** 近红外光谱；特征变量；和声搜索算法；载荷系数；偏最小二乘法

**中图分类号**：O657.3 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2020)06-1869-07

## 引 言

近红外光谱分析作为光谱测量和化学计量学结合的分析技术，是近年来分析化学领域发展最迅速的高新分析技术，以其简便、快速、低成本、绿色环保、涵盖信息量大和多组分同时测定等优点，广泛应用于烟草、石油、纺织、食品等领域的定量检测和定性分析，是现阶段开展广域范围、大规模样品检测最佳的技术手段[1]。然而，近红外光谱由于变量维度高、包含众多与测定量无关的冗余信息和高频噪声等，对其直接建模不但会增加模型的复杂度，同时也会影响模型的预测性能和泛化能力[2]。因此，如何从高维的光谱变量中筛选与预测指标密切相关的特征变量，从而构建准确、稳定、高效的预测模型就显得尤为重要。

目前，许多学者参与到近红外光谱的特征变量筛选研究当中，取得了一定成果。李倩倩等[3]应用无信息变量消除法(uninformative variables elimination，UVE)剔除了光谱中不含有效信息的光谱点，然后构建总氮和总糖的PLS定量模型。但由于过度依赖光谱和指标关联关系，并且需要单个变量逐一提取，容易忽略光谱变量全局贡献度。徐宝鼎等[4]应用粒子群算法(particle swarm optimization，PSO)进行近红外光谱特征选择，采用特征分层的方式划分初始粒子群。但优化过程中只针对固定的初始粒子群内部进行优选，忽略了不同粒子群光谱之间的比较。

鉴于此，本文提出了基于改进和声搜索(harmony search，HS)算法的光谱特征变量选择方法。首先通过偏最

**收稿日期**：2019-04-15，**修订日期**：2019-08-04

**基金项目**：国家重点研发计划课题(2018YFB1701703)，云南中烟工业有限责任公司科技项目(2016XX01)资助

**作者简介**：张 磊，1987年生，中国海洋大学信息科学与工程学院博士研究生 e-mail：zhanglei_0036@163.com

*通讯联系人 e-mail：wallis8@126.com

小二乘(partial least square，PLS)载荷系数计算各光谱点的特征贡献度，作为和声搜索算法改进的扰动权重。然后应用和声搜索算法进行最优光谱变量选择，为了避免陷入局部最优或者收敛速度过慢，引入变量特征贡献度(变量扰动权重)对初始和声记忆库生成方法进行优化，并对和声搜索算法的参数进行动态调整，提高了算法的自适应能力，为近红外光谱的特征变量选择提供了更好的解决方案，有效降低了建模复杂度并提高了模型性能。

## 1 理论与算法

### 1.1 光谱变量特征贡献度

为了计算光谱变量的特征贡献度[5]，首先应用 PLS 方法计算各光谱点的载荷(X-Loading)系数 $p_i$。通过定义光谱矩阵提取因子 $t_i$ 和预测指标矩阵提取因子 $u_i$ 之间最大的协方差，求解如式(1)的最优化问题

$$\max\{\mathrm{Cov}(t_1,u_1)\}=\max\langle E_0w_1,F_0c_1\rangle \quad (1)$$
$$\text{s.t.}=\begin{cases}w_1^{\mathrm{T}}w_1=1\\c_1^{\mathrm{T}}c_1=1\end{cases}$$

利用拉格朗日乘数法求出 $w_1$ 和 $c_1$ 满足式(2)

$$\begin{cases}E_0^{\mathrm{T}}F_0F_0^{\mathrm{T}}E_0w_1=\theta_1^2w_1\\F_0^{\mathrm{T}}E_0E_0^{\mathrm{T}}F_0c_1=\theta_1^2c_1\end{cases} \quad (2)$$

其中，$E_0$ 和 $F_0$ 分别为 $X$ 与 $Y$ 的标准化数据，$w_1$ 是 $E_0^{\mathrm{T}}F_0F_0^{\mathrm{T}}E_0$ 的单位特征向量，$\theta_1^2$ 是对应的特征值即目标函数值的平方，$c_1$ 是 $F_0^{\mathrm{T}}E_0E_0^{\mathrm{T}}F_0$ 最大特征值 $\theta_1^2$ 的单位特征向量。求出 $w_1$ 和 $c_1$ 即可得成分 $t_1=E_0w_1$，然后分别求 $E_0$ 和 $F_0$ 对 $t_1$ 的回归方程，如式(3)所示

$$\begin{cases}E_0=t_1p_1^{\mathrm{T}}+E_1\\F_0=t_1r_1^{\mathrm{T}}+F_1\end{cases} \quad (3)$$

由此可得 $X$ 第一维的载荷矢量为：$p_1=E_0^{\mathrm{T}}t_1/\|t_1\|^2$，$E_1$ 和 $F_1$ 为回归方程的残差矩阵。用残差矩阵 $E_1$ 和 $F_1$ 取代 $E_0$ 和 $F_0$，相应求出 $w_2$ 和 $c_2$ 以及第二个主成分 $t_2$，同理求得第二维的载荷矢量 $p_2=E_1^{\mathrm{T}}t_2/\|t_2\|^2$。通过循环迭代，计算得出每一维的载荷矢量 $p_i$。

最后应用式(4)计算每个光谱变量的贡献度：其中，$n$ 为光谱维数。

$$g_i=\frac{|p_i|}{\sum_{i=1}^{n}|p_i|} \quad (4)$$

### 1.2 和声搜索算法的改进和光谱特征变量选择过程

和声搜索算法是一种相对较为新颖的启发式优化算法，在 2001 年由 Geem 等提出，主要用于解决组合优化和特征变量选择问题，在许多研究领域展现了良好的性能。与演奏家们的目标是合奏出最优美的和声一样，HS 的标准就是优化问题的目标函数[6]。

和声搜索算法的改进和特征变量选择步骤如下。

#### 1.2.1 初始化和声记忆库

和声记忆库(harmony memory，HM)中解向量的每一个分量都是用该决策变量的下界和上界之间均匀分布的随机数(范围$[Lx_i,Ux_i]$，$1\leqslant i\leqslant N$)来初始化[7]。第 $j$ 个解向量的第 $i$ 个分量通过式(5)给出

$$x_i^j=Lx_i+(Ux_i-Lx_i)\mathrm{rand}[0,1] \quad (5)$$

其中，$j=1,2,3,\cdots,HMS$，rand[0，1]是 0－1 之间均匀分布的随机数。

第 $j$ 个解向量的目标函数值用 $f(x^j)$来表示，置为解向量的最后一个分量。对应 HM 的矩阵结构由式(6)生成

$$HM(j,1:N)=x^j$$
$$HM(j,N+1)=f(x^j) \quad (6)$$

由于上述和声记忆库的初始解向量完全是随机生成的，其在整个解空间 $X_i$ 中的分布可能存在极端情况，经过不断迭代很容易陷入局部最优的状态，在一定程度上影响算法的搜索性能[8]。为了增强全局搜索能力，在各分量生成时引入该分量的特征贡献度 $g_i$ 作为激励因子，采用随机遍历和激励因子共同作用的方式生成初始解向量，如式(7)所示

$$x'_i=\mathrm{Sigmoid}(x'_i+g_i)=\frac{1}{1+\exp[-(x'_i+g_i)]} \quad (7)$$

本工作中，输入变量个数为 1 555 个光谱点，各光谱点依次横向在和声向量中排列。如果该光谱点被选定，则其值设为 1，否则设为 0。每一行的最后一位表示该解向量目标函数。$HM$ 矩阵结构如图 1 所示。

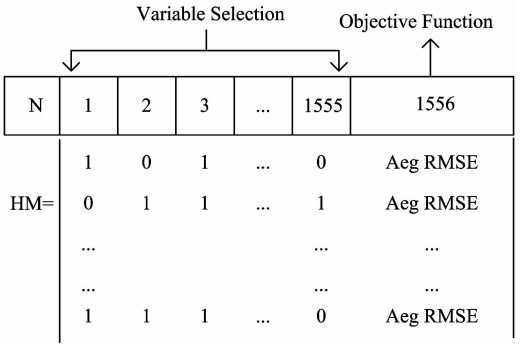


**图 1 和声记忆库(HM)矩阵结构**

**Fig. 1 Harmony memory matrix structure**

#### 1.2.2 从和声记忆库生成新的光谱特征选择向量

HS 算法综合考虑随机扰动性和搜索效率，产生新的和声向量 $x'=x'_1,x'_2,\cdots,x'_n$，具体步骤如下：

(1)在[0－1]之间生成一个随机数 $r_1$，与和声库取值概率(harmony memory considering rate，HMCR)进行比较，若 $r_1<HMCR$，从 HM 中随机取出一个和声变量；若 $r_1\geqslant$HMCR，则从解空间随机生成一个和声变量，如式(8)所示

$$x'_i=\begin{cases}x'_i\in HM,\ x'_1,x'_2,\cdots,x'_n, & r_1<\mathrm{HMCR}\\x'_i\in X_i, & r_1\geqslant\mathrm{HMCR}\end{cases} \quad (8)$$

(2)若这个和声变量是从 HM 中得到的，则对这个和声变量进行微调。在[0－1]之间生成一个随机数 $r_2$，与音调微调概率(pitch adjusting rate，PAR)进行比较，若 $r_2<$PAR，根据微调带宽(band width，BW)对和声变量进行调整，得到一个新的和声变量；若 $r_2\geqslant$PAR，不做任何调整，直接得到新的和声变量[9]。和声变量调整如式(9)所示

$$x'_i=\begin{cases}x'_i\pm r_3\times BW, & r_2<\mathrm{PAR}\\x'_i, & r_2\geqslant\mathrm{PAR}\end{cases} \quad (9)$$

其中，$r_3$ 是[0－1]之间的随机数，若 $r_3>0.5$，取“＋”号；若 $r_3\leqslant 0.5$，取“－”号。

(3)对和声变量更新方式的改进。根据和声变量微调后的 Sigmoid 函数[10]控制变量的更新，在[0－1]之间生成一个随机数 $r_4$，与 Sigmoid($x'_i$)进行比较，若 $r_4<$Sigmoid($x'_i$)，将 $x'_i$ 的值置为 1；若 $r_4\geqslant$Sigmoid($x'_i$)，将 $x'_i$ 的值置为 0。特征变量选择时，1 表示该变量被选择，0 表示该变量不选择，计算过程如式(10)所示

$$\text{Sigmoid}(x'_i)=\frac{1}{1+\mathrm{e}^{-x'_i}}$$
$$\begin{cases}x'_i=1, & r_4<\text{Sigmoid}(x'_i)\\ x'_i=0, & r_4\geqslant\text{Sigmoid}(x'_i)\end{cases}\tag{10}$$

传统 HS 算法中，$BW$ 在整个迭代过程中是不变的，在某种程度上会影响搜索过程的遍历性和种群的多样性，容易陷入局部最优[11]。因此，本文以光谱变量特征贡献度 $g_i$ 作为惩罚项，引入 $\alpha$ 作为 $BW$ 和 $g_i$ 的平衡因子，并根据式(11)使 $BW$ 随迭代次数而动态调整，从而适应光谱选择问题特定阶段的搜索。

$$BW=(1-\alpha)BW-\alpha(BW-g_i)\tag{11}$$
$$\alpha=\begin{cases}t/T, & t<T/2\\ 0.5, & t\geqslant T/2\end{cases}$$

其中，$t$ 为当前迭代次数，$T$ 为最大迭代数。产生新和声向量的过程如图 2 所示。

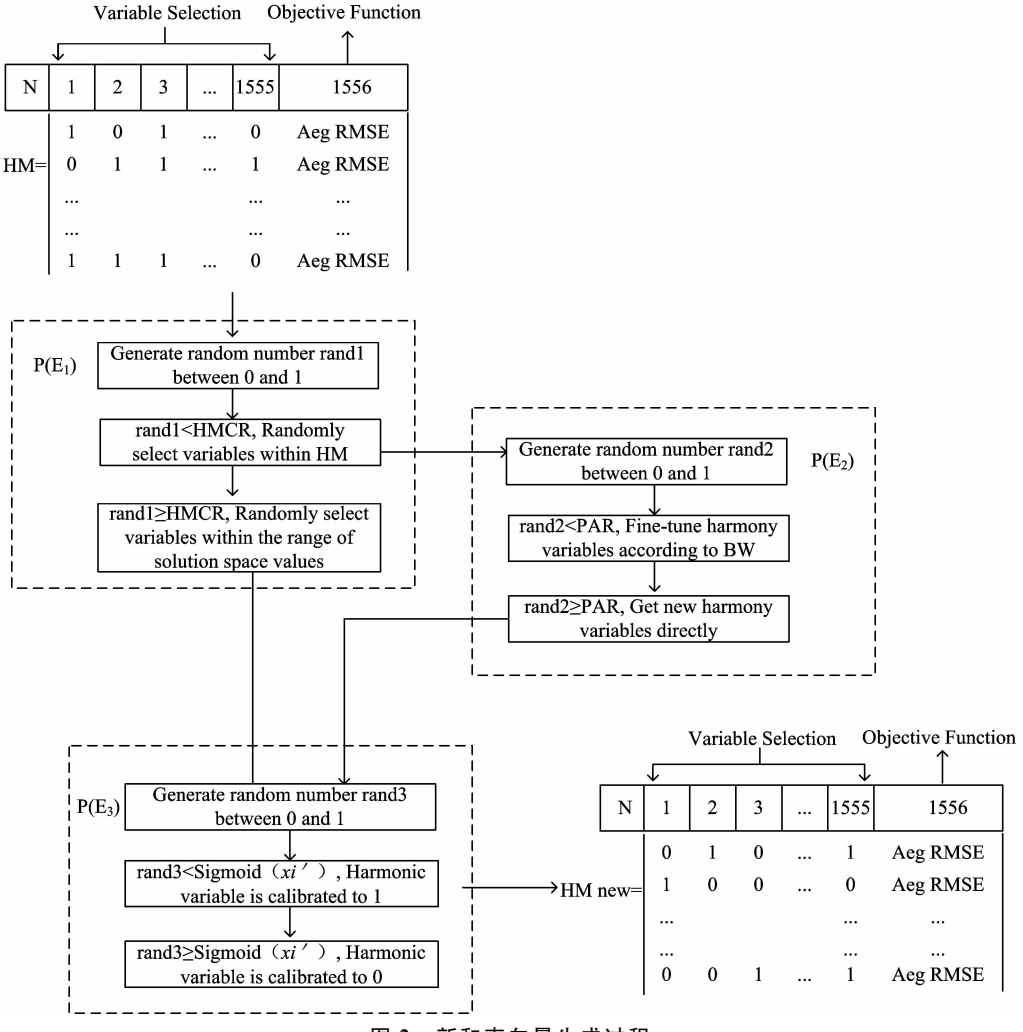


**图 2　新和声向量生成过程**

$P(E_1)$，$P(E_2)$，$P(E_3)$：通过多概率扰动生成新和声；和声库取值概率($E_1$)，音调微调概率($E_2$)，S 型函数概率($E_3$)

**Fig. 2　New harmony vectors generation process**

$P(E_1)$, $P(E_2)$, $P(E_3)$: Probabilities to improvise a new harmony vector by memory considerations HMCR ($E_1$), pitch adjustment rate PAR ($E_2$), and activation function Sigmoid ($E_3$)

### 1.2.3　应用光谱特征选择向量构建 PLS 预测模型

光谱特征向量中各分量表示光谱点选择状态，非零列表示该光谱点被选中。最后一个分量应用各训练子集均方根误差(root mean square error，RMSE)的均值[12]作为光谱特征向量的目标函数，如式(12)所示

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{i=1}^{n}e_i^2}{n}}=\sqrt{\frac{\sum_{i=1}^{n}(p_i-a_i)^2}{n}} \quad (12)$$

其中，$n$ 为样本个数，$e_i$ 为第 $i$ 个样本的预测误差，等于第 $i$ 个样本的预测值 $p_i$ 和真实值 $a_i$ 之间的差值。

针对各训练子集数据，基于 HM 中每个光谱特征向量对应选择的特征光谱点，分别构建 PLS 预测模型，最终计算各训练子集模型 RMSE 的均值作为目标函数值。如果该目标函数值比 HM 中最差的光谱特征向量好，即 $f(X^{\mathrm{new}})<f(X^{\mathrm{worst}})$，则用 $X^{\mathrm{new}}$ 替换函数值最差的光谱特征向量 $X^{\mathrm{worst}}$；否则，和声记忆库不做修改。将 HM 按照目标函数值进行降序排列，重复上述步骤，直到迭代次数达到预先设定的最大周期数为止[13]。

## 2 实验部分

### 2.1 样本数据

为验证算法的有效性，共选取 800 个具有代表性的烟叶样品，随机选取其中的 600 个样品作为训练集，剩余的 200 个作为测试集。进一步，又将训练集随机划分成 3 个数量相等的训练子集，用于增强变量选择算法的泛化能力。样本数据分布情况如表 1 所示。

**表 1 样本数据分布情况**

**Table 1 Sample data distribution**

| 样本数据集 | | 检测项指标 | 样本数量 |
|---|---|---|---|
| Training set | Training subset 1 | nicotine, total sugar, total nitrogen | 200 |
| | Training subset 2 | nicotine, total sugar, total nitrogen | 200 |
| | Training subset 3 | nicotine, total sugar, total nitrogen | 200 |
| Test set | | nicotine, total sugar, total nitrogen | 200 |

### 2.2 仪器设备与采集方法

选用布鲁克公司的 MATRIX-I 型傅里叶变换近红外光谱仪采集近红外光谱，扫描范围为 4 000～10 000 $\mathrm{cm^{-1}}$，分辨率为 8 $\mathrm{cm^{-1}}$，扫描次数为 64。将烟叶样品放置在 60 ℃烘箱中烘干 2 h，用旋风磨磨成粉末，过 40 目筛，放入干燥皿中。每个样品称重 15 g，放置于干净的 5 cm 样品杯中，用压样器自然压实后进行近红外光谱扫描。为保证光谱一致性，每个样品均重复装样扫描三次，采用三次扫描的平均光谱作为该样品的最终光谱，样品三次扫描的平均误差为：$1.5\times10^{-3}$ A，标准差为：$0.4\times10^{-3}$。烟碱、总糖、总氮三个化学指标含量均按照烟草行业规定的标准方法测定。

### 2.3 光谱预处理

采用一阶导数加 Savitzky-Golay 平滑的方法进行光谱预处理[14]，移动窗口宽度为 9，多项式数为 3。图 3 为采用 2.2 中方法采集的原始光谱图，图 4 为预处理后光谱图。

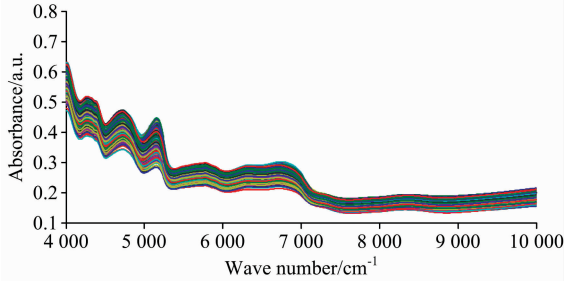


**图 3 原始光谱图**

**Fig. 3 Original spectra**

## 3 结果与讨论

为充分验证算法的有效性，选取烟草化学分析中比较重要且常用的烟碱、总糖、总氮三个指标，分别开展对应的光谱特征变量选择和预测建模实验。

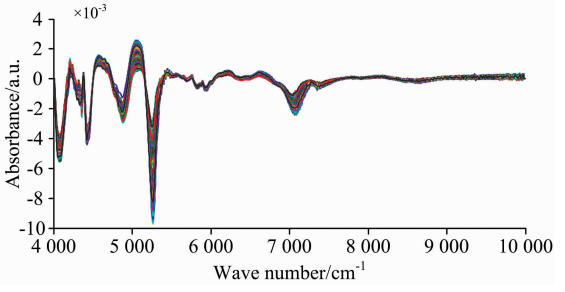


**图 4 预处理后光谱图**

**Fig. 4 Pre-processed spectra**

### 3.1 光谱变量特征贡献度计算

在 600 个样品的训练集上，采用近红外光谱的全谱段和各指标检测结果构建 PLS 模型，得到第一主成分下各光谱点的载荷矢量，然后求得各光谱变量的特征贡献度。不同预测指标对应光谱变量的特征贡献度均不相同，各指标光谱变量贡献度如图 5 所示。

### 3.2 光谱特征变量选择结果

改进和声搜索算法根据各训练子集 RMSE 均值最小化作为约束标准进行光谱特征变量选择，以向量中取值为 1 的分量作为筛选出的变量。算法在 3 个训练子集上各运行 100 次。以烟碱指标为例，图 6 表示 1 555 个变量在总计 300 次训练中被选择的累积频次，频次越高说明该光谱变量对预测指标越重要，可以选做特征变量。将各变量被选择的累积频次从高到低排序，以 25 为下降梯度不断增加变量数，分别计算训练集上 PLS 模型预测结果的 RMSEC。图 7 为随变量数的增加 RMSEC 变化趋势图。由图可知，开始时，随着变量数增加，RMSEC 逐渐降低，当选择变量频次为 200 时达到最小值。如果变量继续增加，会引入噪声和冗余信息，

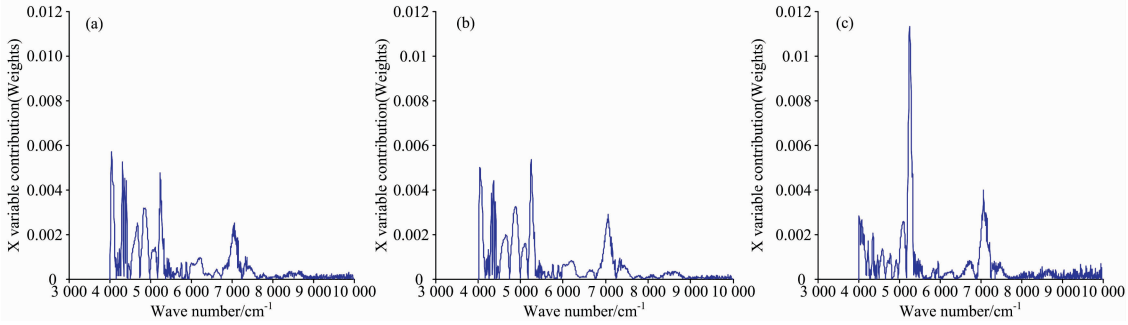


**图 5　不同指标对应的光谱变量贡献度**

(a)：烟碱；(b)：总糖；(c)：总氮

**Fig. 5　Contributions of each spectral variable**

(a)：Nicotine；(b)：Total sugar；(c)：Total nitrogen

RMSEC 随之变大，对模型效果产生不利影响。因此以累积频次 200 作为临界点，最终得到烟碱指标光谱的特征变量个数为 199。对于总糖和总氮指标，临界点分别 200 和 150，光谱特征变量个数分别为 213 和 276。

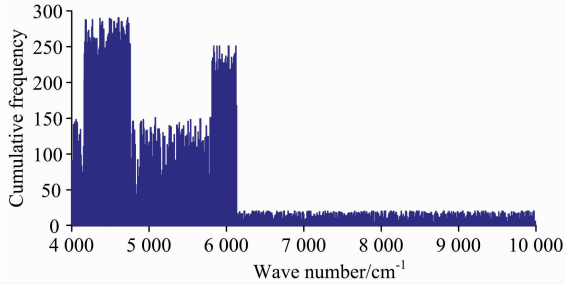


**图 6　光谱变量被选择的累积频次**

**Fig. 6　Cumulative frequency of spectral variables selected**

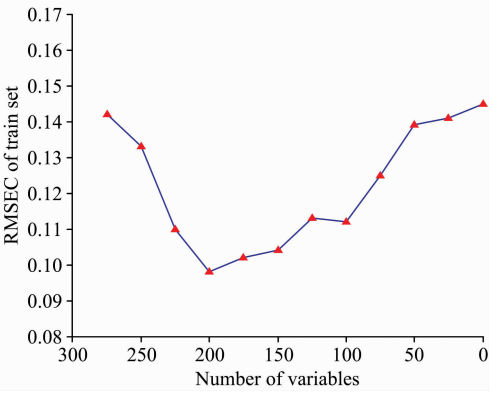


**图 7　RMSEC 随变量增加变化趋势**

**Fig. 7　RMSEC trends with variables increasing**

### 3.3　预测结果对比

为了验证算法的有效性，将改进和声搜索算法与全光谱以及 UVE 和 PSO 等光谱特征变量选择方法进行对比，分别构建烟碱、总糖、总氮三个指标的 PLS 预测模型，以选择的光谱特征变量数、训练集和测试集的决定系数及均方根误差作为评价标准，结果如表2所示。选择多次实

**表 2　各类型特征光谱在三个指标训练集和预测集上的性能**

**Table 2　Performance of training and prediction sets using various types of characteristic spectra of three indicators**

| 预测指标 | 变量选择方法 | 光谱特征变量个数 | 主因子数 | 训练集 | | 测试集 | |
|---|---|---|---|---|---|---|---|
| | | | | $R^2$ | RMSEC | $R^2$ | RMSEP |
| 烟碱 | Full spectrum | 1 555 | 6 | 0.905 2 | 0.145 0 | 0.887 5 | 0.148 1 |
| | UVE | 797 | 6 | 0.925 4 | 0.122 5 | 0.906 4 | 0.132 2 |
| | PSO | 882 | 6 | 0.918 1 | 0.131 2 | 0.904 9 | 0.145 3 |
| | HS | 199 | 6 | 0.954 5 | 0.098 3 | 0.921 1 | 0.102 3 |
| 总糖 | Full spectrum | 1 555 | 7 | 0.909 1 | 1.133 5 | 0.874 3 | 1.145 7 |
| | UVE | 702 | 7 | 0.934 8 | 1.097 2 | 0.915 9 | 1.116 5 |
| | PSO | 743 | 7 | 0.915 3 | 1.126 9 | 0.890 2 | 1.138 4 |
| | HS | 213 | 7 | 0.942 8 | 0.902 1 | 0.925 7 | 1.034 6 |
| 总氮 | Full spectrum | 1 555 | 9 | 0.948 5 | 0.065 4 | 0.9149 | 0.076 8 |
| | UVE | 849 | 9 | 0.963 2 | 0.056 9 | 0.932 4 | 0.066 4 |
| | PSO | 991 | 9 | 0.957 7 | 0.062 1 | 0.9197 | 0.071 3 |
| | HS | 276 | 9 | 0.968 5 | 0.045 2 | 0.941 2 | 0.053 1 |

验中最小的 PRESS 值对应的主因子数作为最终 PLS 模型的主因子数。

决定系数 $R^2$ 越大，预测均方根误差 RMSEP 越小，代表模型的预测性能越好。由表 2 可见，各种特征变量选择方法相对全光谱来说，均减少了波长点数，同时也提高了预测精度。相较而言，本文提出的改进和声搜索算法对三个预测指标都提取了最少的光谱特征变量(提取的光谱特征变量只占全光谱的 13%～18%)，具有最高的 $R^2$ 和最小的 RMSEP 值。同时，经过专家对筛选谱段对应的官能团分析，所选谱段均能够较好的反应其指标项特征信息，充分表明了本算法提取的光谱特征变量能有效降低冗余信息和噪声、消减变量间的多重共线性，使得模型更加稳健、泛化能力更强。

## 4 结 论

基于改进的和声搜索算法提出了一种近红外光谱特征变量选择方法。首先利用 PLS 载荷系数计算光谱变量对预测指标的贡献度，作为变量权重。然后，利用和声搜索算法进行特征变量筛选，过程中引入变量权重对和声搜索算法的初始化和参数动态调整进行改进。最后针对筛选后的变量通过 PLS 建模在烟碱、总糖、总氮三个指标的训练集和测试集进行验证，并与全光谱和几种常用的特征变量选择方法进行对比。实验结果表明，采用本算法进行光谱特征变量选择对模型性能优于全光谱和其他光谱变量选择方法，由此说明对近红外光谱进行特征变量选择的必要性以及本算法的有效性，既保证了模型预测性能又降低了建模的复杂度，为近红外定量预测模型的构建和优化提供了参考。

## References


[1] CHEN Li-ju, LIU Wei(陈丽菊, 刘 巍). Modern Physics(现代物理知识), 2016, 18(2): 10.

[2] SUN Wen-ping, GONG Hui-li, WANG Mei-xun, et al(孙文苹, 宫会丽, 王梅勋, 等). Microcomputer & Its Applications(微型机与应用), 2015, 34(1): 78.

[3] LI Qian-qian, TIAN Kuang-da, LI Zu-hong, et al(李倩倩, 田旷达, 李祖红, 等). Chinese Journal of Analytical Chemistry(分析化学), 2013, 41(6): 917.

[4] XU Bao-ding, QIN Yu-hua, YANG Ning, et al(徐宝鼎, 秦玉华, 杨 宁, 等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2019, 39(3): 717.

[5] WANG Yong, WANG Li-fu, ZOU Hui, et al(王 勇, 王李福, 邹 辉, 等). Computer Engineering and Design(计算机工程与设计), 2018, 378(6): 127.

[6] Moayedikia A, Ong K L, Boo Y L, et al. Engineering Applications of Artificial Intelligence, 2017, 57(C): 38.

[7] Enayatifar R, Yousefi M, Abdullah A H, et al. Communications in Nonlinear Science & Numerical Simulation, 2013, 18(12): 3481.

[8] ZHAI Jun-chang, GAO Li-qun, OUYANG Hai-bin, et al(翟军昌, 高立群, 欧阳海滨, 等). Control and Decision(控制与决策), 2015, 30(11): 1953.

[9] Khalili M, Kharrat R, Salahshoor K, et al. Applied Mathematics & Computation, 2014, 228(9): 195.

[10] Sutskever I, Hinton G E. Neural Computation, 2014, 20(11): 2629.

[11] OUYANG Hai-bin, GAO Li-qun, ZOU De-xuan, et al(欧阳海滨, 高立群, 邹德旋, 等). Control Theory and Applications(控制理论与应用), 2014, 31(1): 57.

[12] JIANG Hong, SU Yang(江 虹, 苏 阳). Laser and Infrared(激光与红外), 2016, 46(1): 119.

[13] Abdelgayed T S, Morsi W G, Sidhu T S. IEEE Transactions on Smart Grid, 2018, 9(2): 521.

[14] LIU Yan, CAI Wen-sheng, SHAO Xue-guang(刘 言, 蔡文生, 邵学广). Chinese Science Bulletin(科学通报), 2015, (8): 704.

# Research on Near Infrared Spectral Feature Variable Selection Method Based on Improved Harmonic Search Algorithm

ZHANG Lei[1], DING Xiang-qian[1], GONG Hui-li[1], WU Li-jun[2*], BAI Xiao-li[2], LUO Lin[2]

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

2. China Tobacco Yunnan Industry Co., Ltd., Technical Research Center, Kunming 650024, China

**Abstract** Near-infrared spectroscopy has been widely used in many fields for detection and analysis because of its advantages of simplicity, speed, efficiency, low cost, and environment protection. However, the NIR spectra also contain interferences such as high variable dimension, multiple collinearities, redundant information, and high frequency noise. The direct construction of the prediction model not only increases the modeling complexity but also affects the prediction performance and generalization. For this purpose, a spectral feature variable selection method based on the improved Harmony Search algorithm (HS) is proposed. HS is often used to solve feature variable optimization problem. When the spectral variable selection is applied by the HS algorithm, the feature contribution of spectra is firstly calculated by the PLS loading coefficient as the disturbance weight of the improved HS. In the process of optimizing the spectral feature variables, the variable feature contribution is introduced as the excitation factor, and the initial solution vectors are generated by the combination of random traversal and excitation factor. When generating the new harmony vector, the feature contribution is applied as a penalty factor, and the parameters of HS are dynamically adjusted with the number of iterations by adding the balance factor, so as to adapt to the search of spectral variables. It enhances the ergodicity of the search process and the diversity of the population. In order to verify the effectiveness of the algorithm, the NIR PLS models of nicotine, total sugar and total nitrogen using tobacco samples are constructed. After pre-processing the original spectra, this method is used to optimize spectral variables. The prediction performance of each model corresponding to the number of different variables is calculated according to the cumulative frequency at which the variables are selected, and the final selected spectral variables are determined by the increasing trend of the Root Mean Square Error of Calibration (RMSEC) with the variables. The three PLS models are established on the training set and the test set respectively, and they are compared with the full spectrum, Uninformative Variables Elimination (UVE) and Particle Swarm Optimization (PSO). The experimental results show that the coefficient of determination ($R^2$) of nicotine, total sugar and total nitrogen models using the selected variables is 0.921 1, 0.925 7 and 0.941 2, respectively; and the Root Mean Square Error of Prediction (RMSEP) is 0.102 3, 1.034 6 and 0.053 1. Compared with other methods, the RMSEP of this study is low, the $R^2$ of these models is more than 0.92, and the spectral characteristic variables are small. It is shown that the improved HS algorithm can effectively filter the feature spectrum, reduce the modeling complexity, improve the model prediction performance and generalization ability.

**Keywords** Near infrared spectroscopy; Feature variables; Harmony search algorithm; Loading factor; Partial least squares

(Received Apr. 15, 2019; accepted Aug. 4, 2019)

* Corresponding author