doi: 10.12068/j.issn.1005-3026.2018.07.012

基于样本选择与PSO-ANN的葡萄酒酒精浓度预测

王巧云, 郑念祖
(东北大学 信息科学与工程学院, 辽宁 沈阳 110819)

摘 要: 为了提高拉曼光谱定量分析模型的准确性以及稳健性,提出了一种新的样本选择算法——KM法.实验中以40组葡萄酒光谱为分析对象,将KM法与传统的RS,KS,SPXY样本选择算法相比较.实验结果表明: KM法获得的|RMSEP-RMSEC|要优于其他三种方法,剩余预测偏差(RPD)存在显著性差异,说明KM法具有很好的预测准确度.同时,针对BP神经网络易陷入局部极值的问题,将粒子群优化算法用于优化人工神经网络的参数(PSO-ANN),通过与遗传算法、人工鱼群算法及混合蛙跳算法比较,发现PSO-ANN较之于其他三种方法,能够提高BP神经网络泛化性能,具有收敛速度快、稳健性强及预测精度高等优势.

关 键 词: 样本选择算法;群体智能算法;BP神经网络;拉曼光谱;葡萄酒;粒子群优化
中图分类号: O 657.37 **文献标志码**: A **文章编号**: 1005-3026(2018)07-0970-06

Prediction of Wine Alcohol Concentration Based on Sample Selection and PSO-ANN

WANG Qiao-yun, *ZHENG Nian-zu*
(School of Information Science & Engineering, Northeastern University, Shenyang 110819, China. Corresponding author: WANG Qiao-yun, E-mail: wangqiaoyun@neuq.edu.cn)

Abstract: In order to improve the accuracy and robustness of the quantitative analysis model, a new sample selection algorithm named KM was proposed. In the experiment, 40 samples of wine were used as the analysis objects, and the KM algorithms was compared with traditional sample selection algorithms, such as RS, KS and SPXY. The experimental results show that |RMSEP-RMSEC| obtained by KM algorithm is superior to the other three algorithms, and there are significant differences in RPD, which indicates that KM method has good prediction accuracy. In order to overcome the neural network training algorithms drawbacks that BP neural networks converge slowly and is easy to fall into local optimum, the particle swarm optimization algorithm was used to optimize the parameters of artificial neural network (PSO-ANN). The results show that PSO-ANN algorithm can improve the convergence velocity of training, robustness and the accuracy of classification than genetic algorithm, artificial fish swarm algorithm, and shuffled frog-leaping algorithm.

Key words: sample selection algorithm; swarm intelligence algorithm; BP neural network; Raman spectroscopy; wine; particle swarm optimization

拉曼光谱由于具有检测快速、无损、信息丰富、适用于水溶液、分辨率高以及可在线监测等优点,使得其在食品药品、水质检测、化工等众多领域得到了广泛的应用[1-2].但是在外界因素以及物质复杂组分的干扰下,使得拉曼光谱与待测属性之间存在较强的非线性,因此常见的线性建模方法,如多元线性回归(multiple linear regression, MLR)、主成分回归(principal component

收稿日期: 2017-03-23
基金项目: 国家自然科学基金青年基金资助项目(11404054,61601104); 河北省自然科学基金青年基金资助项目(F2017501052); 中央高校基本科研业务费专项资金资助项目(N142304003,N172304032).
作者简介: 王巧云(1980-),女,河北秦皇岛人,东北大学副教授.

regression，PCR）和偏最小二乘法（partial least squares，PLS）等方法很难得到满意的效果. 而近几年发展的非线性建模方法如人工神经网络（artificial neural network，ANN）以及支持向量机（support vector machine，SVM）等往往无法适用于实际中存在的非线性关系，虽然具有较好泛化能力的 BP 神经网络可以有效地建立光谱信息与待测组分浓度的定量模型[3]，但模型建立的前提是有一个具有代表性的训练样本集，这对模型的稳定性有至关重要的作用. 目前常用的样本选择算法有随机样本选择法（random sampling，RS），Kennard – Stone（KS）法，SPXY（sample set partitioning based on joint $x-y$ distance）法等方法. 不少学者对样本选择及建模方法有所研究：詹雪艳等[4]基于 SPXY 样本选择算法，采用偏最小二乘回归（PLSR）算法建立近红外定量模型，用于积雪草苷含量的快速预测，取得了较好的预测效果；靳召晰等[5]提出 K 近邻 – 密度样本选择方法，建立近红外光谱的定性分析模型，用于小麦的多分类问题，该方法有效增强了模型的识别效果，并降低建模样本量；Zhao 等[6]则提出了子空间分离与多光谱独立成分回归建模算法，通过放大不同的较小特定段，从而使得模型具有更好的预测能力. 而本文提出一种新的样本选择算法 K – Means（KM）法，实践证明 KM 法选出的样本集具有很好的代表性.

为了克服 BP 神经网络鲁棒性差等局限性，将群体智能算法（swarm intelligence algorithm）引入 BP 神经网络成为研究热点[7]. 常见的群体智能算法有：遗传算法、粒子群算法、混合蛙跳算法、人工鱼群算法等，这些群体智能算法优化神经网络，能提高其精度和稳健性，然而对它们之间进行系统比较的报道很少.

本文以葡萄酒发酵过程中的拉曼光谱为研究对象，比较了不同校正集样本选择方法对拉曼光谱模型的预测能力和稳健性的影响，并讨论基于不同群体智能算法的神经网络的性能.

1　基本原理与方法

1.1　常规样本选择算法

1）RS 法. 无重复随机地选取所需数量的样本作为训练集，由于实际的样本集大多是不平衡数据，导致选择的样本集无法具有足够的代表性，因此该算法所选样本集建立的模型性能较差.

2）Kennard – Stone（KS）法. KS 法是根据样本光谱间距离进行选择样本集的方法[8]. 其运行过程为：首先计算两两样本之间的距离，将相距最远的两个样本加入训练集；之后计算训练集中每个已选样本与剩余候选样本之间的距离，在距已选样本最近的样本集合中选择距离最远的样本加入训练集中；以此类推，不断选择直到满足训练样本数目的要求为止，任意两个样本之间的距离多采用欧氏距离来衡量. 该算法选出的训练集样本多分布在样本空间的边缘，会有不同程度的区域集中现象.

3）SPXY 法. 在 KS 算法基础上，将其距离函数引入待测成分变量 y，以光谱与待测成分共生距离作为样本选择的依据[9]. SPXY 算法的选择过程与 KS 算法相似. 对样本所在光谱高维空间 X 与待测组分空间赋予相同的权重，可以保证样本选择能够更有效地覆盖各个空间. SPXY 法样本选择改善了模型的稳定性与预测能力，同时由于待测组分浓度的引入，也带来其他方面的误差，无法根本上克服 KS 算法所具有的问题.

1.2　KM 法

通过对代表整个光谱高维空间的主成分空间进行不完全搜索，将该空间聚类为若干个范围，之后再从每个范围按照一定比例随机选择所需数量的训练样本. KM 算法可以描述为：根据样本特征选定 k 类，并通过 PCA 得到得分矩阵 $\boldsymbol{S}$，然后通过迭代更新 k 个聚类中心使目标函数 F 取得最小值，其评价函数为

$$F=\sum_{s\in \boldsymbol{s}}\min_{c\in C}\|\boldsymbol{s}-c\|^2. \tag{1}$$

KM 算法具体步骤如下：

1）将获得的原始光谱数据 $\boldsymbol{X}\in\mathbf{R}^{n\times p}$ 标准化，计算其相关系数矩阵：

$$\boldsymbol{R}=\boldsymbol{X}^{\mathrm{T}}-\boldsymbol{X}/(n-1)\ . \tag{2}$$

2）求解相关矩阵 $\boldsymbol{R}$ 的特征方程式（式（3）），获得 p 个特征值，并使得特征值 λ 满足 $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p$.

$$|\boldsymbol{R}-\lambda\boldsymbol{I}_p|=0\ . \tag{3}$$

3）根据累计贡献率 $\sum_1^m\lambda_1/\sum_1^p\lambda_1$ 确定 m 值，并用式（3）解得前 m 个特征值 $\lambda_1,\lambda_2,\cdots,\lambda_m$ 对应特征向量，并组成特征向量矩阵 $\boldsymbol{P}\in\mathbf{R}^{m\times p}$.

$$|\boldsymbol{R}-\lambda\boldsymbol{I}_p|P_i=0,i=1,2,\cdots,m\ . \tag{4}$$

4）将标准后的光谱数据与步骤 3）求得的特征向量矩阵 $\boldsymbol{X}\in\mathbf{R}^{p\times m}$ 相乘，从而获得主成分矩阵：$\boldsymbol{S}\in\mathbf{R}^{n\times m}$.

5）随机选取 k 个初始聚类中心 c .

6）对于每个样本的得分向量 $\boldsymbol{s}\in\boldsymbol{S}$，若与聚

类中心 c_i, $i=1,2,\cdots,k$ 距离最近,则将该样本放入聚类簇 C_i 中,有 $C_i=\underset{i}{\operatorname{argmin}}\|s-c_i\|$.

7) 重新计算聚类簇 C_i 的中心,见式(5):

$$c_i=\frac{1}{|C_i|}\sum_{s\in C_i}s,i=1,2,\cdots,k. \tag{5}$$

8) 不断执行步骤 6)和 7),直到聚类中心趋于收敛.

9) 从聚类簇 C_i 中按一定比例随机选取样本组成训练集,其余的组成预测集.

实验表明,该样本集具有很强的代表性,能够实现训练样本集有足够宽的背景信息及训练集在光谱空间均匀分布的基本要求,因此,KM 算法是一个很有潜力的研究方向.

1.3 群体智能算法优化 BP 神经网络

对于现实情况中存在复杂的映射关系,随机地初始化网络结构参数,由于很难获得具有全局收敛性的初始位置,所以无法取得全局最优解. 而群体智能算法利用群体的优势,以全局并行搜索的方式进行搜索,不断迭代寻优,能有效解决局部极小值的问题. 将两者结合,既可以发挥神经网络强大的非线性映射能力,又能借助群体智能算法的全局寻优的特性以提高神经网络的泛化性能,能够解决神经网络自身所存在的不足.

其基本思想如下:首先群体智能算法初始化群体,其中群体中每个个体都是由神经网络的权值阈值组成,这些个体可以随机产生,也可以通过特定的方式产生;之后内部个体依照各自觅食进化模式和协同优化机制移动,不断迭代趋近全局最优解,最后将获得的最好个体解码赋给神经网络. 这种群体搜索突破了邻域搜索的限制,可以直接对求解空间的分布式信息进行采集和搜索,具有简单、快速、稳定性强的特点.

不同群体智能算法有不同的搜索模式:粒子群算法(PSO)根据速度 - 位移方式,并参考自身和全局最优解来实现搜索位置的变动[10];遗传算法(GA)通过染色体的交叉、变异、选择算子来搜索最优解[11];人工鱼群算法(AFSA)基于三种搜索行为即聚群行为、追尾行为、觅食行为来实现搜索过程[12];混合蛙跳算法(SFLA)通过模因分组、局部位置更新实现寻优[13]. 将这些算法优化 BP 神经网络的模型分别简称为 PSO - ANN, GA - ANN, AFSA - ANN, SFLA - ANN.

BP 神经网络随着隐含层数的增加更能够刻画现实世界的复杂关系,同时对数据有更深度的表达[14]. 但是传统的梯度下降法具有随着层数的增加存在梯度消失(gradient vanish)的现象[15],这对于多层神经网络的学习能力有所限制,因此有必要引入群体智能算法避免陷入局部最优解,本文选取 4 层 BP 神经网络(2 个隐含层)建立模型,各隐含层神经元数采用构造法予以确定.

2 实验和方法

2.1 实验材料与仪器

实验所用的标准溶液均是通过商业途径购买纯酒精、甘油和葡萄糖加入纯水稀释得来的. 发酵过程中的葡萄酒溶液由实验室的发酵罐获得. 光谱数据来源于 40 个样本,其中酒精体积分数范围为 0.01% ~0.09%;葡萄糖体积分数为 0.01% ~0.19%,甘油质量浓度范围为 0.003 ~0.05 g/mL. 所有的样本均储存在 5 ℃以下冷藏环境,以防止在光谱采集和实验室测试中发生任何特性的变化. 所有样本的标准浓度值均由安捷伦高效液相色谱仪获得.

实验过程中使用的 MultiRAM 傅里叶变换拉曼光谱仪(Bruker Optics, Germany)配有高性能液氮冷却的 Ge 检测器、标准 Nd:YAG (1 064 nm)激光器和 OPUS 7.0 (Bruker Optics, Germany)光谱分析软件程序. 首先将样品分为两份,一份用液相色谱测试,另一份放在石英透明小杯中,用拉曼光谱仪器在室温环境下进行光谱采集,共进行 512 次扫描,扫描速率为 10 kHz,光谱分辨率 6 cm^{-1},扫描范围 400 ~4 000 cm^{-1},重复 3 次测量取均值作为样本光谱,40 个样本获得的拉曼光谱见图 1.

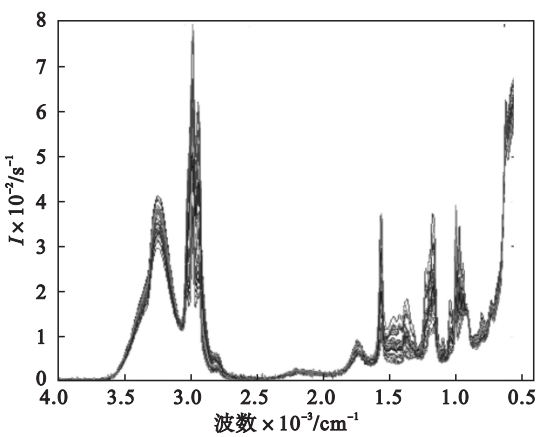

图 1 葡萄酒样本的拉曼光谱

Fig. 1 Raman spectra of grape wine sample

2.2 定量分析模型的评价参数

所建立的拉曼光谱定量分析模型,其性能的好坏决定于该模型的准确度以及稳健性,本文以训练集与预测集相关系数(R_C, R_P)、训练集与预

测集均方根误差(RMSEC,RMSEP)及预测集相对分析误差 RPD 来评价拉曼光谱所建立模型的性能. |RMSEP – RMSEC| 值越大,模型稳健性越差;RPD 越高,模型准确度越高,计算公式见式(6)~式(8).

$$\mathrm{RMSEC} = \sqrt{\frac{\sum_{i=1}^{m}(\hat{y}_{i,\mathrm{c}} - y_{i,\mathrm{c}})^2}{m}}, \quad (6)$$

$$\mathrm{RMSEP} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_{i,\mathrm{p}} - y_{i,\mathrm{p}})^2}{n}}, \quad (7)$$

$$\mathrm{RPD} = \sqrt{\frac{\sum_{i=1}^{n}(\bar{y}_{\mathrm{p}} - y_{i,\mathrm{p}})^2}{\sum_{i=1}^{n}(\hat{y}_{\mathrm{p}} - y_{i,\mathrm{p}})^2}}. \quad (8)$$

式中:m 和 n 分别为训练集与预测集的样本数;$y_{i,\mathrm{c}}$,$\hat{y}_{i,\mathrm{c}}$和 $\bar{y}_{\mathrm{c}}$ 分别为训练集中第 i 个样本待测成分浓度化学值、模型预测值和均值;$y_{i,\mathrm{p}}$,$\hat{y}_{i,\mathrm{p}}$和 $\bar{y}_{\mathrm{p}}$ 分别为预测集中第 i 个样本待测成分浓度化学值、模型预测值和均值.

3　结果与讨论

3.1　样本选择算法的比较

在模型建立前,本文分别采用随机样本选择法、Kennard – Stone(KS)法、SPXY 法与 KM 法进行对比试验. 葡萄酒发酵过程中共有 40 个样本,其中 30 个样本用于训练模型,剩下 10 个样本用于测试,在相同预处理(多元散射校正)条件下,分别对它们得到的样本集建立模型,结果如表 1 所示. 从表 1 可以看出:本文提出的 KM 算法在各个方面均明显优于其他校正集选择算法,具有最小的预测集及训练集误差均方根 RMSEC,RMSEP,且|RMSEP – RMSEP| 值最小,相关系数 R_{P},R_{C} 均优于其他三种方法,RPD 值有显著差异. 综合分析,基于 KM 算法所建立的模型具有很好的精确度和稳健性,其所选样本具有很强的代表性.

表 1　训练样本选择算法的比较

Table 1　Comparison of training sample selection algorithm

算法	RMSEC	RMSEP	|RMSEP – RMSEC|	R_{P}	R_{C}	RPD
RS	0.000 96	0.001 23	0.000 27	0.997 91	0.999 36	15.8
KS	0.000 41	0.000 97	0.000 56	0.998 95	0.999 93	22.4
SPXY	0.000 58	0.000 84	0.000 26	0.999 58	0.999 93	28.6
KM	0.000 31	0.000 55	0.000 24	0.999 61	0.999 97	37.9

3.2　隐含层神经元数的选择

为了获得构建 4 层 BP 神经网络的结构参数,本文以 2 个隐含层神经元数为横坐标,通过重复试验 20 次,取预测集相关系数 R_{P} 的均值为纵坐标绘制三维网格图. 通过该图很容易地确定具有较强预测性能的神经网络所需要的结构参数即 2 个隐含层神经元数. 由图 2 可以看出,隐含层层间神经元数比较接近或者递减时,神经网络预测性能较强.

3.3　群体智能算法的选择

为了对比不同群体智能算法对模型的影响,均通过 KM 法选择训练集与预测集,然后分别采用粒子群算法、遗传算法、人工鱼群算法及混合蛙跳算法优化 BP 神经网络建立模型(分别为 PSO – ANN,GA – ANN,AFSA – ANN,SFLA – ANN). 所有模型的群体规模均设置为 20,寻优迭代次数为 15,其中 GA – ANN 设置交叉率、变异率分别为 0.85,0.01;AFSA – ANN 设置 try – number = 4,拥挤因子为 0.11;SFLA – ANN 中石头数量为 5,青蛙尝试跳跃的次数为 3. 在不同主成分数(分别为 7,8,9,10)下,针对每种情况进行 10 组全局寻优实验,得到一组关于模型评价指标(主要为 RPD,|RMSEP – RMSEC| 及时间)的数据,然后对数据组进行统计分析:采用数据的均值表征本组数据的优劣,其标准误差(standard error,SE)则反映该组数据的稳定性.

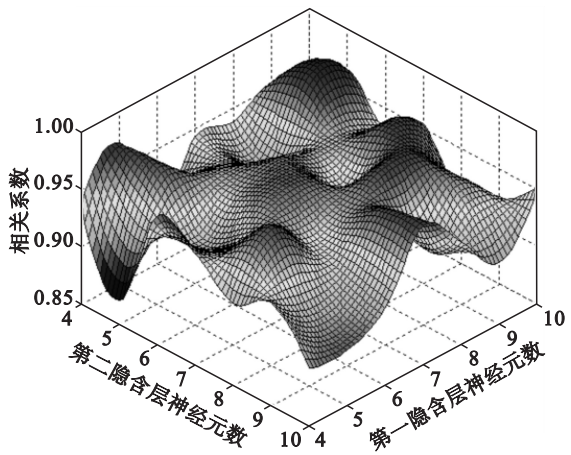

图 2　隐含层节点数对相关系数的影响

Fig. 2　Effect of hidden nodes on correlation coefficient

实验结果见图 3 ~ 图 5,由此得出:PSO – ANN 表现出优异的全局寻优能力,预测分析误差 RPD 在主成分数为 9 时,平均值高达 35,同时|RMSEP – RMSEC|值最小,时间方面也相对较小;GA – ANN 与 SFLA – ANN 性能相当,但后者在时间方面始终小于前者;而 AFSA – ANN 则在所有方面均表现出最差.

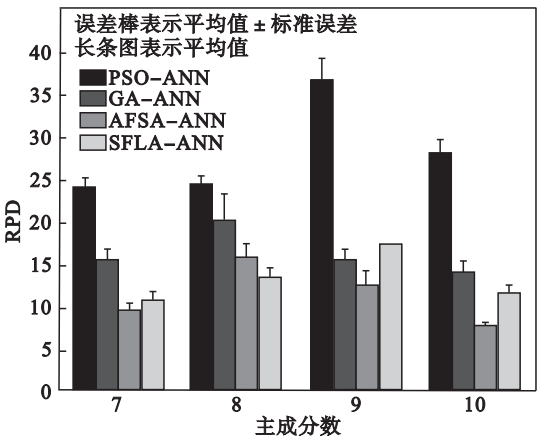

图 3 不同模型 RPD 的统计结果

Fig. 3 Statistics results of RPD of different model

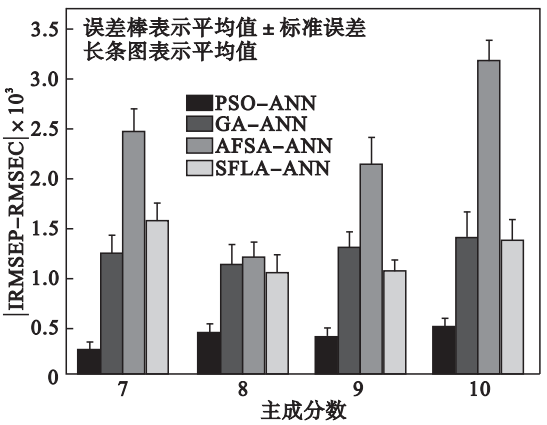

图 4 不同模型|RMSEP – RMSEC|统计结果

Fig. 4 Statistics results of |RMSEP – RMSEC| of different model

所有模型都随主成分数的增加,预测精度先增加后减少,而寻优时间和|RMSEP – RMSEC|变化不大. 4 种模型在最优主成分下的性能指标对比见表 2,综合分析可得,PSO – ANN 能够较好地跳出局部极值,较其他算法,该模型收敛速度快、预测精度高,且搜索具有全局性及较强稳健性,优于其他模型,可广泛应用于实践当中.

因此,首先对光谱进行多元散射校正,然后通过 KM 算法对葡萄酒发酵过程的 40 个样本进行划分以获得建立模型所需的训练集和预测集,之后建立 PSO – ANN 模型,所得模型预测结果见图 6,可以看出,模型很好地预测了葡萄酒发酵过程中酒精的含量,为实际的应用打下良好的模型基础.

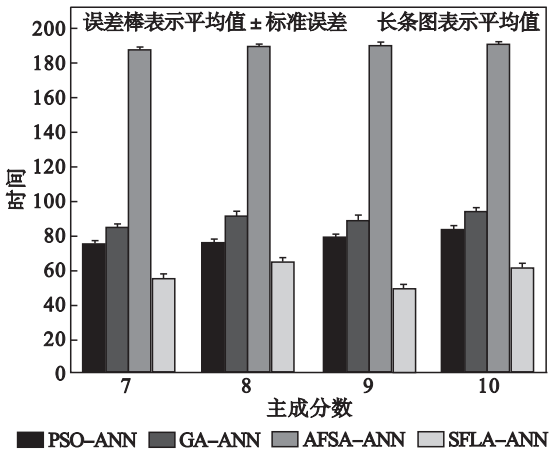

图 5 不同模型时间的统计结果

Fig. 5 Statistics results of time of different model

表 2 训练样本选择方法的比较

Table 2 Comparison of training sample selection

方法	PCs	|RMSEP – RMSEC|	RPD	时间/s
PSO	9	3.6E – 04	36.8	77.7
GA	8	1.0E – 03	20.1	91.0
AFSA	8	1.2E – 03	15.7	190.2
SFLA	9	1.0E – 03	17.3	48.3

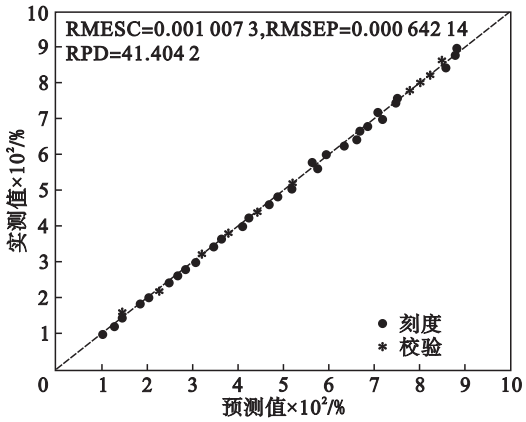

图 6 酒精预测值和实测值之间的相关图

Fig. 6 Correlation plots between predict values and target values

4 结 论

1)本文采用拉曼光谱法测量葡萄酒发酵过程中酒精的含量,提出一种新的样本选择方法——KM 法. 该算法较之于 RS 法、KS 法、SPXY 法可以选出更有代表性的样本集,不仅节约时间,而且能够提高模型的预测精确度及稳健性.

2）针对 BP 神经网络存在易陷入局部最优值的缺陷，分别建立 PSO – ANN，GA – ANN，AFSA – ANN，SFLA – ANN 模型，将其应用于定量分析过程中. 仿真结果表明，粒子群算法较其他算法，优化后的神经网络收敛速度快，具有预测精度高与稳健性强的优势，该研究为神经网络应用于实际提供了解决方法.

参考文献：

[1] 叶向晖，沈于兰，申兰慧，等. 拉曼光谱法在食品药品分析中的应用与进展[J]. 中国药业，2017，26(1)：1 – 5.
(Ye Xiang-hui，Shen Yu-lan，Shen Lan-hui，et al. Raman spectroscopy in the application of food and drug analysis and progress [J]. *China Pharmaceutical*，2017，26 (1)：1 – 5.)

[2] 文军. 拉曼光谱和表面张力用于水质检测与分析[J]. 物理实验，2014，34(5)：9 – 12.
(Wen Jun. Raman spectroscopy and surface tension for water quality detection and analysis [J]. *Physical Experiment*，2014，34 (5)：9 – 12.)

[3] Rocha R A D，Paiva I M，Anjos V，et al. Quantification of whey in fluid milk using confocal Raman microscopy and artificial neural network [J]. *Journal of Dairy Science*，2015，98(6)：3559 – 3567.

[4] 詹雪艳，赵娜，林兆洲，等. 校正集选择方法对于积雪草总苷中积雪草苷 NIR 定量模型的影响[J]. 光谱学与光谱分析，2014 (12)：3267 – 3272.
(Zhan Xue-yan，Zhao Na，Lin Zhao-zhou，et al. Effect of algorithms for calibration set selection on quantitatively determining asiaticoside content in centella total gluecosides by near infrared spectroscopy [J]. *Spectroscopy and Spectral Analysis*，2014(12)：3267 – 3272.)

[5] 靳召晰，张秀娟，罗付义，等. 近红外光谱建模样本选择方法研究[J]. 光谱学与光谱分析，2016，36(12)：3920 – 3925.
(Jin Zhao-xi，Zhang Xiu-juan，Luo Fu-yi，et al. Study of modeling samples selection method based on near infrared spectrum[J]. *Spectroscopy and Spectral Analysis*，2016，36 (12)：3920 – 3925.)

[6] Zhao C H，Guo F R，Wang F L. Spectra data analysis and calibration modeling method using spectra subspace separation and multiblock independent component regression strategy [J]. *AIChE Journal*，2011，57(5)：1202 – 1215.

[7] Liu F，Xie L，Li B J，et al. Multi-sense swarm intelligence algorithm and its application in feed-forward neural networks training[J]. *Journal of University of Science and Technology Beijing*，2008，30(9)：1061 – 1066.

[8] Saptoro A，Tade M O，Vuthaluru H. A modified Kennard-Stone algorithm for optimal division of data for developing artificial neural network models[J]. *Chemical Product and Process Modeling*，2012，7(1)：1 – 14.

[9] Zhan X M，Zhu X R，Shi X Y，et al. Determination of hesperidin in tangerine leaf by near-infrared spectroscopy with SPXY algorithm for sample subset partitioning and monte carlo cross validation[J]. *Spectroscopy and Spectral Analysis*，2009，29(4)：964 – 968.

[10] Bezborah A. A hardware architecture for training of artificial neural networks using particle swarm optimization[C] // The Third International Conference on Intelligent Systems Modelling and Simulation. Kota Kinabalu：IEEE Computer Society，2012：67 – 70.

[11] Lilichenko M，Kelley A M. Application of artificial neural networks and genetic algorithms to modeling molecular electronic spectra in solution [J]. *Journal of Chemical Physics*，2001，114：7094 – 7102.

[12] Hai M J，Xie K，Wang Y F. Optimization of pump parameters for gain flattened Raman fiber amplifiers based on artificial fish school algorithm [J]. *Optics Communications*，2011，284：5480 – 5483.

[13] Eusuff M，Lansey K，Pasha F. Shuffled frog-leaping algorithm：a memetic meta-heuristic for discrete optimization [J]. *Engineering Optimization*，2006，38：129 – 154.

[14] Diamond P，Fomenko I V. Robustness and universal approximation in multilayer feedforward neural networks [R]. Sydney：Sydney University，1992.

[15] Bengio Y，Simard P，Frasconi P. Learning long-term dependencies with gradient descent is difficult [J]. *IEEE Transactions on Neural Networks*，1994，5(2)：157 – 166.