基于 ABC-SVR 算法的拉曼光谱检测混合油脂肪酸含量

张燕君[1,2]，张芳草[1]，付兴虎[1*]，金培俊[1]，侯姣茹[1]

1. 燕山大学信息科学与工程学院，河北省特种光纤与光纤传感重点实验室，河北 秦皇岛 066004
2. Department of Electrical and Computer Engineering，Missouri University of Science and Technology，Rolla，Missouri 65401，USA

摘 要 提出了一种基于激光拉曼光谱和人工蜂群智能优化支持向量回归机(ABC-SVR)算法的快速定量检测三组分混和油中3种脂肪酸含量的方法。该方法针对光谱数据信息与样本之间非线性、高维度的关系，建立了预测精度及建模效率均高于同类对比算法的数学模型，同时避免了气相色谱法、液相色谱法等对混合油脂肪酸含量的检测方式，根据纯种油中3种脂肪酸含量的国际标准，由油品配置体积得到脂肪酸质量，有效降低了检测成本与实验复杂程度，提高了检测工作的实用价值。首先根据一定梯度配置66组混合油检测样品，使用便携式拉曼光谱仪采集样本的拉曼光谱信息，扣除背景噪声；观察多组样本的拉曼光谱图可知，由于官能团浓度的差异，食用油的拉曼特征峰位移基本相同，特征峰的峰值明显不同，因此基于特征峰信息可以区分食用调和油的不同混合物；其次对拉曼光谱做背景扣除、光谱平滑、最大值谱线归一化三步预处理，以降低实验中不可控的外界因素及背景荧光的影响，准确提取光谱特征峰强度信息；然后根据纯种油中3种脂肪酸的国际标准含量，结合国家食品法典委员会标准CODEX STAN210—1999《指定的植物油法典标准》中规定的纯种油密度中值，由油品体积得到脂肪酸质量数；随机选取56组样本数据作为训练集，剩余10组样本数据作为预测集；以训练集光谱特征峰强度和脂肪酸质量分别作为回归模型的输入及输出值，建立SVR和PSO-SVR，ABC-SVR三种混合优化算法对比的定量分析模型，对测试集的3种脂肪酸含量分别进行预测；最后通过均方误差(MSE)、相关系数(r)及建模时间(Elapsed time)分别进行对比，建立数据表对模型精准度进行检验。实验结果表明，通过ABC-SVR定量分析模型效果最佳，3种脂肪酸含量预测值与真实值的均方差分别为0.88×10^{-4}，16×10^{-4}和8×10^{-4}，均低于0.002；相关系数分别为93.43%，99.65%和99.43%，均高于93%；预测时间(Elapsed time)分别为1.26，2.42和2.14 s。因此，所提出的检测方法，具备较高的精确度、较快的建模时间，且在理论上的类似条件下可适用于其他样品检测工作，可为振动光谱学对食用油掺伪分析的进一步工作提供可行的理论依据。

关键词 激光拉曼光谱；人工蜂群；支持向量回归机；脂肪酸；混合油

中图分类号：TN247 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2019)07-2147-06

引 言

中国营养协会对脂肪酸的摄入量给出建议，饱和脂肪酸、单不饱和脂肪酸、多不饱和脂肪酸3种关键脂肪酸的的供能比应符合一定的营养比例[1]。营养调和油因其科学控制的脂肪酸比例受大众青睐，对市面食用调和油的脂肪酸比例进行定量分析从而识别掺假及不合格调和油的工作尤为重要。

目前，应用于食用油脂肪酸含量检测的实验室分析法主要是气相[2]、液相色谱法、质谱法等[3]，需要严格控制实验条件，时间成本极高；振动光谱学结合化学计量方式在食用油检测工作中因其独特的优势广受研究者青睐：邓之银等将支持向量机与拉曼光谱技术结合起来，对纯种食用油及混合油的脂肪酸含量进行了检测；Carmona等[4]利用拉曼光谱探索了食用油的不饱和度及抗氧化性等特性；Gouvinhas等[5]根据拉曼光谱法判断不同成熟期橄榄果榨取的油脂的化学特性。以上研究大多限于单组份食用油检测，且相应的数学模

收稿日期：2018-06-03，**修订日期**：2018-10-15

基金项目：国家自然科学基金项目(11673040，61675176)，国家公派访问学者项目(201708130010)，燕山大学“新锐工程”人才支持计划项目资助

作者简介：张燕君，女，1973年生，燕山大学信息科学与工程学院教授 e-mail：yjzhang@ysu.edu.cn

*通讯联系人 e-mail：fuxinghu@ysu.edu.cn

型存在局部最优等问题，基于此，本文将人工蜂群智能优化支持向量回归机(artificial bee colony-support vector machine for regression，ABC-SVR)算法与拉曼光谱相结合，对混合油的脂肪酸进行定量检测，以期为振动光谱学对食用油掺伪分析提供可行的理论依据。

1 混合优化算法基本原理

1.1 回归支持向量机

支持向量回归机(SVR)是在支持向量机分类(SVM)的基础之上发展起来的、用于处理回归预测问题的机器学习算法[6-7]。设样本集$\{(x_i, y_i), i=1, 2, \cdots, l\}$，其中$x_i \in R^n$是第$i$个样本的输入值，$y_i \in R$为第$i$个样本的输出值。假定函数

$$f(x) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b \tag{1}$$

式中，$\langle \boldsymbol{w}, \boldsymbol{x} \rangle$表示对同维向量$\boldsymbol{w}$，$\boldsymbol{x}$求内积，$b \in R$为调节参数。

引入拉格朗日函数

$$L(\boldsymbol{w}, b, \xi^{(*)}, \alpha^{(*)}, \eta^{(*)}) = \frac{1}{2} \| \boldsymbol{w} \|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) - \sum_{i=1}^{l}(\eta_i\xi_i + \eta_i^*\xi_i^*) - \sum_{i=1}^{l}\alpha_i(\varepsilon + \xi_i - y_i + \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) - \sum_{i=1}^{l}\alpha_i^*(\varepsilon + \xi_i^* - y_i - \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle - b) \tag{2}$$

式中，$\xi^{(*)}$，$\alpha^{(*)}$和$\eta^{(*)}$分别为ξ，α，η有$*$和无$*$两种情况，$\alpha^{(*)}$，$\eta^{(*)}$是拉格朗日乘子。引入松弛变量ξ_i，$\xi_i^* \geqslant 0$，$i=1, 2, \cdots, l$，可以将优化问题表示为如下对偶形式

$$\max \frac{1}{2}\sum_{i=1}^{l}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle + \varepsilon\sum_{i=1}^{l}(\alpha_i + \alpha_i^*) - \sum_{i=1}^{l}y_i(\alpha_i - \alpha_i^*)$$
$$\text{s.t.}: \begin{cases} \sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0 \\ \alpha_i, \alpha_i^* \in [0, C] \end{cases} \tag{3}$$

式中，$C>0$为惩罚参数[8]。

根据最优KKT条件[9]，$\alpha^{(*)}$和$\eta^{(*)}$与约束项之积在最优点处应为0，即

$$\begin{cases} \alpha_i(\varepsilon + \xi_i - y_i + \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) = 0 \\ \alpha_i^*(\varepsilon + \xi_i^* - y_i - \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle - b) = 0 \\ \eta_i\xi_i = 0 \rightarrow (C - \alpha_i)\xi_i = 0 \\ \eta_i^*\xi_i^* = 0 \rightarrow (C - \alpha_i^*)\xi_i^* = 0 \end{cases} \tag{4}$$

当$0<\alpha_i<C$时，$\xi_i=0$；当$0<\alpha_i^*<C$时，$\xi_i^*=0$，此时对应的样本为标准支持向量，为计算的可靠性，通常对标准支持向量分别求b值，再求平均值，即

$$b = \frac{1}{N_{\text{NSV}}}\left\{\sum_{0<\alpha_i<C}\left[y_i - \sum_{x_j \in SV}(\alpha_j - \alpha_j^*)\langle \boldsymbol{x}_j, \boldsymbol{x}_i \rangle - \varepsilon\right] + \sum_{0<\alpha_i^*<C}\left[y_i - \sum_{x_j \in SV}(\alpha_j - \alpha_j^*)\langle \boldsymbol{x}_j, \boldsymbol{x}_i \rangle + \varepsilon\right]\right\} \tag{5}$$

式(5)中，N_{NSV}是标准支持向量个数，因此可得SVR的估计函数为

$$f(\boldsymbol{x}) = \sum_{x_i \in SV}(\alpha_i - \alpha_i^*)(\boldsymbol{x}_i\boldsymbol{x}) + b \tag{6}$$

1.2 人工蜂群智能算法优化支持向量回归机

ABC作为仿生算法用于模拟蜜蜂的搜索行为，具备独特的全局和局部搜索方式。食物源的位置等同于数学问题中的最优解更新公式

$$v_{ij} = x_{ij} + r_{ij}(x_{ij} - x_{kj}) \tag{7}$$

式(7)中，$k \in \{1, 2, \cdots, SN\}$，$k \neq i$，是$i$邻域的一个值，$j \in \{1, 2, \cdots, d\}$，$d$是样本的维数，$k$，$j$是随机数，$r_{ij} \in [-1, +1]$为随机数，用于限制$x_{ij}$的邻域范围。

ABC将对找到的食物收益进行估计，决定是否定义其为目标

$$P_i = \frac{fit_i}{\sum_{i=1}^{SN} fit_i} \tag{8}$$

式(8)中，fit_i代表第i个解的适应度值，SN表示解的数量。

通过控制参数limit改变某个解的更新次数，若通过limit次更新之后仍然需改进求解精度，则预示算法陷入了局部最优，此时相应的解将更新

$$x_{ij} = x_{\min j} + \text{rand}(0, 1)(x_{\max j} - x_{\min j}) \tag{9}$$

基于以上基本理论，首先对ABC的基本参数初始化，再对SVR的关键参数进行寻优，适应度函数为

$$fit_i = \frac{1}{SN}\sum_{i=1}^{SN}(y_i - \hat{y}_i)^2 \tag{10}$$

式(10)中，y_i为实际值，$\hat{y}_i$为预测值，当fit_i达到最小时得到最优解。

2 实验部分

2.1 激光拉曼光谱技术

拉曼光谱技术是印度科学家发现的散射光谱，谱峰强度随着样品中化学键及官能团的浓度发生变化，可用于混合物中各组分的定性、定量无损检测[10]。

2.2 仪器相关参数及样品制备

实验使用的上海必达泰克公司生产的型号为BWS465-785S的便携式拉曼光谱仪，其光谱范围为0～3 500 cm^{-1}。实验中选择激发光源波长为785 nm，经多次测试，设置拉曼光谱仪激光功率百分比为60%(最大激发功率为300 mW)，每个样本扫描10次并取均值作为该样本光谱数据，积分时长设置600 ms。通过BWRam4TM软件进行光谱预处理，获取光谱特征值，根据获得的特征值建立模型，并测试其预测功能，最后得到结果并分析。实验中使用的纯种油购自本地大型超市，按照一定体积梯度配置成66组混合样品待测。调和油产品标识的3种脂肪酸营养比例为质量比，因而需将体积比转化成质量比。通过查询各纯种油产品营养成分，得出3种脂肪酸的含量如表1所示。

国家食品法典委员会标准CODEX STAN210—1999《指定的植物油法典标准》中规定纯大豆油相对于水的密度为0.919～0.925，纯花生油：0.914～0.917，纯葵花籽油：0.918～0.923，本实验中选取各范围的中值进行后续研究。

通过 $m=\rho v$ 得到样品中各油分的质量数，再结合表 1 计算得到 3 种脂肪酸的含量。

表 1　各纯种油营养成分

Table 1　The nutrient content of each pure oil

	SFA (per 100 g)	MUFA (per 100 g)	PUFA (per 100 g)
Soybean oil/g	16	25	59
Peanut oil/g	20	44	36
Sunflower oil/g	13	21	66

2.3　数据采集及预处理

测试系统的暗电流噪声，每次检测拉曼光谱后需对相应的暗电流进行扣除。随机选取 10 组测得的光谱进行观察，如图 1 所示。

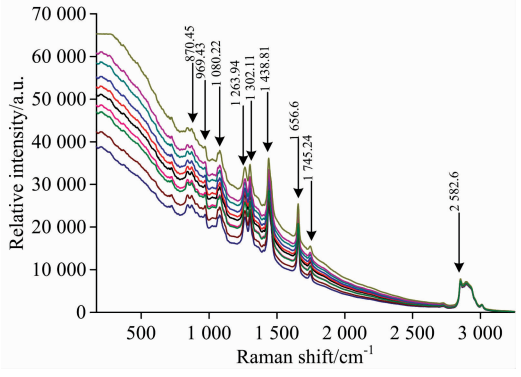

图 1　10 组样本的拉曼光谱

Fig. 1　Raman spectra of 10 samples

由图 1 可知，不同成分混合油的特征峰位移基本相同，最高峰位强度明显不同；选用图中标识的九个峰值作为关键特征峰进行采集待下一步数据处理；由于外界因素及荧光影响，原始光谱出现了峰值淹没的情况导致光谱信噪比降低，故需要进行预处理工作。本文选取背景扣除、光谱平滑、最大值谱线归一化对光谱进行预处理效果分别如图 2 所示。

图 2 中，借助惩罚最小二乘法线性拟合光谱背景噪声，再于原始光谱中对噪声做背景扣除得所需信号；选用 Savitzky-Golay Filters 法进行光谱平滑，利用局部多项式时域最小二乘法拟合，可滤除部分噪声并保护原信号的形状、强度及谱宽[11]；选取拉曼峰值最强(1 438.81 cm^{-1}对应的谱峰峰值强度)处作为归一化因子对整个拉曼光谱进行归一化处理，以此缩小量值。

3　结果与讨论

随机选取 56 组样本作训练集，剩余作预测集。采集光谱数据进行处理并提取特征峰 X 作为模型输入，已知脂肪酸质量样本信息 Y 作为模型输出。预测模型流程如图 3 所示。

测试集样本数据均方误差(mean square error，MSE)及 R 的计算公式为

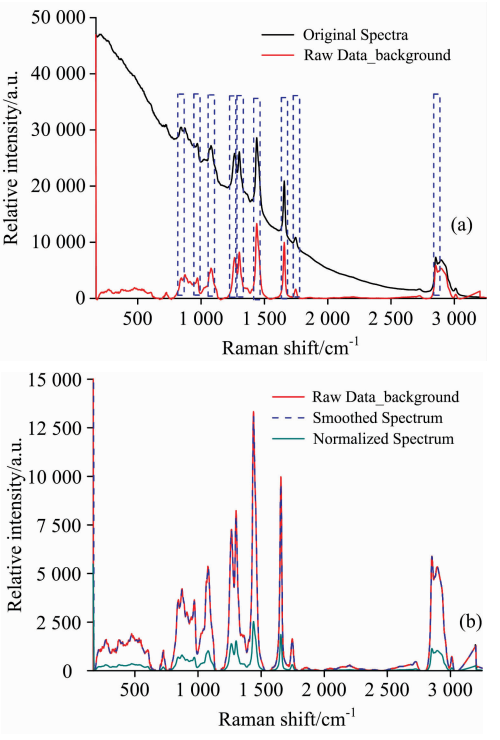

图 2　光谱预处理

Fig. 2　Spectre pretreatment

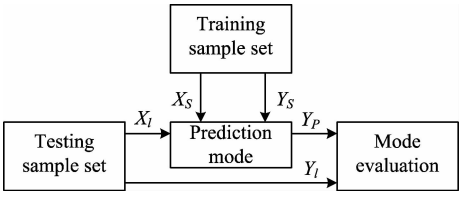

图 3　预测模型流程图

Fig. 3　Flow chart of prediction model

$$\mathrm{MSE}=\frac{\sum_{i=1}^{M}(y_i-\bar{y}_i)^2}{M} \tag{12}$$

$$R=\frac{M\sum_{i=1}^{M}y_i\hat{y}_i-\sum_{i=1}^{M}y_i\hat{y}_i}{\sqrt{\frac{1}{M}\sum_{i=1}^{M}y_i^2-\left(\sum_{i=1}^{M}y_i\right)^2}\times\sqrt{\frac{1}{M}\sum_{i=1}^{M}\hat{y}_i^2-\left(\sum_{i=1}^{M}\hat{y}_i\right)^2}} \tag{13}$$

其中，M 为样本个数，y_i 为实际值，$\hat{y}_i$ 为预测值。

ABC-SVR 数学分析模型对混合油中 3 种脂肪酸预测的评价结果如表 2 所示。

表 2　ABC-SVR 模型评价结果

Table 2　Model evaluation results of ABC-SVR

	MSE	R^2	Elapsed time/s
SFA	0.88×10^{-4}	0.934 3	1.26
MUFA	16×10^{-4}	0.996 5	2.42
PUFA	8×10^{-4}	0.994 3	2.14

表 2 中，由 ABC-SVR 数学分析模型得到的 MSE 均在 0.002 以内；R^2 均高于 93%，相关性拟合如图 4 所示，可见相关性极高；建模时间在 2.5 s 以内，具备良好的实时性。

同时，训练集的训练结果也较好，拟合图如图 5 所示，其中，图中空心圆圈代表训练集实际值，实心方块代表训练集各组分对应的训练值。

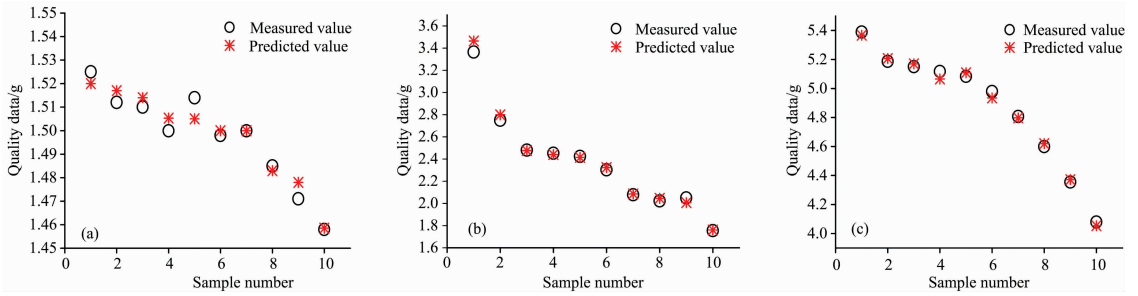

图 4 脂肪酸预测与真实值相关性拟合

(a)：饱和脂肪酸；(b)：单不饱和脂肪酸；(c)：多不饱和脂肪酸

Fig. 4 The relationship between the predicted and the measured values of three fatty acids

(a)：SFA；(b)：MUFA；(c)：PUFA

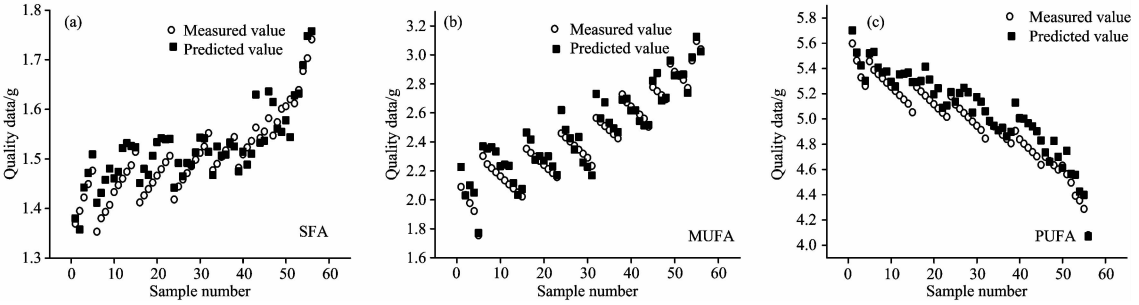

图 5 模型训练集相关性拟合

(a)：饱和脂肪酸；(b)：单不饱和脂肪酸；(c)：多不饱和脂肪酸

Fig. 5 Correlation fitting of model training set

(a)：SFA；(b)：MUFA；(c)：PUFA

将数据通过 SVR 模型与 PSO-SVR 模型进行预测，3 种脂肪酸预测值与真实值间的均方误差对比如表 3 所示。

表 3 均方误差对比

Table 3 Comparison of mean square error

	SFA	MUFA	PUFA
SVR	0.000 930	0.188 400	0.169 000
PSO-SVR	0.244 000	0.146 900	1.033 900
ABC-SVR	0.000 088	0.001 600	0.000 800

表 3 中，传统的 SVR 模型以及 PSO-SVR 模型的均方误差均明显高于 ABC-SVR 模型；为对模型的工作效率进行评估，记录模型建立时间如表 4 所示。

表 4 建模时间对比

Table 4 Comparison of elapsed time (s)

	SFA	MUFA	PUFA
SVR	2.81	0.53	0.48
PSO-SVR	6.96	35.73	29.94
ABC-SVR	1.26	2.42	2.14

综上可知，由 ABC 对 SVR 进行优化，在处理非线性、高维度的基础上，降低了模型复杂度，合理的兼顾了算法全局搜索与局部搜索的能力，因此提高了预测精度、缩短了建模时间，进而提升了模型的工作效率、综合预测性能，更好地完成了对混合食用油中三种脂肪酸的含量的预测工作。

4 结 论

提出利用激光拉曼光谱技术，结合化学计量方式建立数学模型对混合食用油的脂肪酸含量进行快速、无损的定量检测方式。使用便携式拉曼光谱仪进行光谱采集，扣除暗电流后再通过三个步骤对原始光谱进行预处理；将实验中的各油分的体积含量转化成各脂肪酸的质量含量；选取人工鱼群智能优化算法(ABC)对支持向量回归机(SVR)的关键参数进行优化，并结合模型评价标准对本系统的可行性进行确定。结果表明，新方法对于混合油的 3 种脂肪酸无损、快速定量检测的方案可行，具备良好的实时性、极高的预测精度，3 种脂肪酸预测值与真实值的均方误差分别为 0.88×10^{-4}，16×10^{-4} 和 8×10^{-4}，相关度分别为 0.934 3，0.996 5 和 0.994 33，预测时间分别为 1.26，2.42 和 2.14 s，其预测精度及实

时性、实用性均很理想，更可行的完成了对混合食用油脂肪酸含量定量检测的工作，可为振动光谱学对食用油掺伪分析的进一步工作提供可行的理论依据。

References

[1] DONG Jing-jing, WU Jing-zhu, CHEN Yan, et al(董晶晶，吴静珠，陈　岩，等). Imaging Science and Photochemistry(影像科学与光化学), 2017, 35 (2): 147.

[2] ZHENG Yue-ming, FENG Feng, GUO Wei, et al(郑月明，冯　峰，国　伟，等). Chinese Journal of Chromatography(色谱), 2012, 30 (11): 1166.

[3] HE Rong, SHAN Xiao-lin, DONG Fang-yuan, et al(何　榕，山晓琳，董方圆，等). Chinese Journal of Analytical Chemistry(分析化学), 2015, 43(9): 1377.

[4] María Á Carmona, Fernando Lafont, César Jiménez-Sanchidrián, et al. European Journal of Lipid Science & Technology, 2014, 116 (11): 1451.

[5] Irene Gouvinhas, Nelson Machado, Teresa Carvalho, et al. Talanta, 2015, 132: 829.

[6] Balabin R M, Lomakina E I. Analyst, 2011, 136(8): 1703.

[7] Wauters M, Vanhoucke M. Automation in Construction, 2014, 47: 92.

[8] WANG Kai, ZHANG Yong-xiang, et al(王　凯，张永祥，等). Computer Engineering and Applications(计算机工程与应用), 2008, 44 (26): 45.

[9] DING Xiao-jian, ZHAO Yin-liang(丁晓剑，赵银亮). Journal of Xi'an Jiaotong University(西安交通大学学报), 2011, 45(6): 7.

[10] XU Bin, LIN Man-man, YAO Hui-lu, et al(徐　斌，林漫漫，姚辉璐，等). Chinese Journal of Lasers(中国激光), 2016, 43(1): 0115003.

[11] ZHAO Fang, PENG Yan-kun(赵　芳，彭彦昆). Chinese Journal of Lasers(中国激光), 2017, 44(11): 1111001.

Detection of Fatty Acid Content in Mixed Oil by Raman Spectroscopy Based on ABC-SVR Algorithm

ZHANG Yan-jun[1, 2], ZHANG Fang-cao[1], FU Xing-hu[1*], JIN Pei-jun[1], HOU Jiao-ru[1]

1. School of Information Science and Engineering, The Key Laboratory for Special Fiber and Fiber Sensor of Hebei Province, Yanshan University, Qinhuangdao　066004, China
2. Department of Electrical and Computer Engineering, Missouri University of Science and Technology, Rolla, Missouri 65401, USA

Abstract　In this paper, a rapid and quantitative detection method combing the laser Raman spectroscopy with Artificial Bee Colony-Support Vector Machine for Regression (ABC-SVR) is proposed for the determination of fatty acids content in three-component blend oil. This method establishes a mathematical model with higher prediction accuracy and higher modeling efficiency than similar comparison algorithms in solving the nonlinear and high-dimensional complex relationship between spectral data information and samples. And it can avoid complicated detection ways like gas chromatography and liquid chromatography, etc. The quality of fatty acids is obtained from the oil configuration volume according to the international standards for the content of three fatty acids in pure oils, which effectively reduces the cost and complexity of the experiment, and increases the practical value of the inspection work. Firstly, 66 groups of mixed oil test samples were arranged according to a certain gradient. The Raman spectroscopic information of the samples was collected from a portable Raman spectrometer and the background noise was subtracted at the time. Through the spectrum of the samples, we could see that the Raman spectra had the same characteristic peak shifts basically, but the intensities of the characteristic peaks were obviously different because of the difference in functional group concentration. Therefore, different components could be distinguished according to the characteristic peak information. Secondly, the spectra were pretreated by background subtraction, spectral smoothing and normalization to reduce the effect of uncontrollable external factors in the experiment. Then the mass of fatty acid was obtained from the oil volume by the international standard content of three kinds of fatty acids in pure oil in National Codex Alimentarius Commission Standard CODEX STAN210—1999. 2/3 of sample data were randomly selected as the training set, and the remaining 1/3 of sample data were used as the prediction set. The characteristic peak intensity and the quality of fatty acid of train set were used as the input and output values of the regression model, and the quantitative analysis model of hybrid optimization algorithms of SVR, PSO-SVR and

ABC-SVR were established to predict the content of fatty acids of test set. The accuracy of the model was tested by using mean squared error (MSE), the correlation coefficient (R^2) and elapsed time. The experimental results showed that the ABC-SVR quantitative analysis model was effective; the MSE of the predicted and true values of three fatty acid contents were 0.88×10^{-4}, 16×10^{-4} and 8×10^{-4}, respectively. The R^2 were 93.43%, 99.65% and 99.43%, respectively. The elapsed time were 1.26, 2.42 and 2.14 s, respectively. Therefore, the proposed method has higher accuracy, faster modeling time than other ways, and it can be applied to other sample detection work under theoretically similar conditions. This method can provide a viable theoretical basis for further study on the analysis of adulterated edible oils by vibration spectroscopy.

Keywords Laser Raman spectroscopy; Artificial bee colony; Support vector regression machine; Fatty acid; Blend oil

(Received Jun. 3, 2018; accepted Oct. 15, 2018)

* Corresponding author