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，pp3916-3920
December，2019

最小二乘支持向量机的核桃露饮品中脂肪成分的定量分析

李子文[1]，李宗朋[1]，买书魁[1]，盛晓慧[1]，尹建军[1]，刘国荣[2]，
王成涛[2]，张海红[3]，辛立斌[4]，王 健[1*]

1. 中国食品发酵工业研究院有限公司，北京 100015
2. 北京工商大学北京市食品添加剂工程技术研究中心，北京 100048
3. 宁夏大学农学院食品科学系，宁夏 银川 750021
4. 上海普丽盛包装股份有限公司，上海 201514

摘 要 利用近红外光谱对核桃露中的重要指标脂肪含量进行定量分析，同时进行建模变量优化、建模方法比较以优选最佳模型。为消除散射对光谱造成的影响，采用标准正态变换(SNV)方法对数据进行预处理，采用遗传算法(GA)结合向后间隔偏最小二乘法(BiPLS)优选的特征波长分别作为偏最小二乘法(PLS)及最小二乘支持向量机(LS-SVM)的输入变量，建立核桃露中脂肪含量的近红外定量模型，采用决定系数(R^2)、预测标准偏差(RMSEP)以及性能偏差比(RPD)对各模型进行评价，探究光谱波段选择方法对于核桃露中脂肪指标模型构建的影响，同时确定最佳建模方法。结果表明：进行变量筛选能够对模型起到优化作用，Bi-PLS及GA-BiPLS方法分别选择了150及30个变量点，占全光谱的10%及2%，对应了核桃露样品中脂肪成分的特征吸收峰，使得PLS模型的RMSEP值从0.049分别下降到0.043和0.040，同时模型的相关系数R^2从0.964提高到0.973及0.974，性能偏差比RPD从4.88增长到5.62及6.00，主成分数也有不同程度的减少，降低模型复杂程度的同时提高了模型准确性。相比于PLS模型，核桃露脂肪指标的LS-SVM模型的R^2，RMSEP及RPD值均表现出了更好的效果，分别达到0.986，0.036及6.52。说明基于最小二乘支持向量机建立的分析模型有较高的准确度及稳定性，可能是由于PLS作为一种经典的线性建模方法，在建立模型的过程中忽略了样品数据集中的非线性因素，而核桃露样品光谱测量过程中噪声、背景等因素的干扰，以及各指标成分间的相互影响，使得脂肪含量与近红外光谱信息间存在复杂非线性的变化关系，LS-SVM方法能够更为有效地对其进行描述，增强了光谱变量与指标浓度间的相关性，使得建立的模型有着更好的准确度以及普适性，说明了在实际生产中，LS-SVM方法具备优良的可行性，体现了其在核桃露饮品品质分析方面的巨大潜力。基于最小二乘支持向量机方法所建立的核桃露脂肪含量的定量分析模型，具有准确、稳定的特点，能够为核桃露生产的质量监控提供技术借鉴，同时为饮品品质的分析方法研究提供了新的思路。

关键词 核桃露；近红外光谱技术；最小二乘支持向量机；波段筛选
中图分类号：O657.3 **文献标识码**：A **DOI**：10.3964/j.issn.1000-0593(2019)12-3916-05

引 言

核桃露饮品凭借其柔和的口感以及均衡丰富的营养价值受到了消费者广泛的喜爱，但是一些小生产企业为牟取私利，在生产加工过程中偷工减料，使得产品的关键指标达不到国家的相关标准要求，严重影响了核桃露饮品行业的良性发展。核桃露饮品的质量检测受到了市场及消费者越来越多的重视[1]。脂肪作为核桃露饮品原材料核桃的主要组成成分，是衡量核桃露品质的重要指标之一[2]。目前，国标所规定的测量方式为酸水解法，检测过程较为复杂耗时，无法满足快速检测的迫切需求。

近红外光谱分析技术作为一种能够实现快速检测的绿色分析技术，具有操作简单、样品无需前处理、检测效率高等优点[3]。近年来已经在饮料分析方面得到了较为广泛的应用[4-6]，但目前多是针对果汁及茶饮料等相关方面的研究，

收稿日期：2018-10-22，**修订日期**：2019-02-16
基金项目：国家重点研发计划项目(2018YFD0400905)，国家自然科学基金项目(31671937)资助
作者简介：李子文，1992年生，中国食品发酵工业研究院有限公司工程师 e-mail：1339299142@qq.com
*通讯联系人 e-mail：81214112@qq.com

对于含乳饮料尤其是核桃露饮品重要指标的近红外相关报道更是较为少见。

本研究拟对核桃露中的重要品质指标—脂肪含量进行快速分析，采用遗传算法结合向后间隔偏最小二乘法对整个谱区进行波段选择，并分别采用偏最小二乘及最小二乘支持向量机建立快速检测模型，同时进行对比分析，以此降低模型复杂程度，提高模型运算速度及预测能力，为核桃露饮品品质的快速评价提供一定参考依据。

1　实验部分

1.1　仪器

使用 N500 傅里叶变换近红外光谱仪(瑞士步琦有限公司)，光源为卤钨灯，检测器为温控铟镓砷，配有固体测量池及透反射盖。光谱范围为 10 000～4 000 cm^{-1}，分辨率为 8 cm^{-1}，扫描次数为 32 次；利用配套软件 NIRWare Operator 采集核桃露样品的近红外光谱。

1.2　材料

核桃露样品共 372 个，由某饮料公司提供，采用透反射方式采集核桃露样品的近红外光谱，同时为消除散射对光谱造成的影响，采用标准正态变换(SNV)方法对数据进行预处理。核桃露样品脂肪含量根据 GB 5009.6—2016《食品中脂肪的测定》，采用酸水解法测定。

1.3　方法

1.3.1　样品集的划分

在剔除 3 个异常点的基础上，将 60 个核桃露样品随机进行保留，以作为独立预测集样品不参与模型构建，来对模型的预测能力进行独立验证。同时采用 Kennard-Stone(K-S)法[7]根据变量间的欧氏距离以 2∶1 的比例均匀地对余下 309 个样品进行校正集和验证集划分。最终分别得到校正集及验证集样品 206 和 103 个。校正集与验证集样品脂肪含量的分布情况如表 1 所示。

表 1　校正集与验证集脂肪含量统计结果

Table 1　Statistical results of fat in calibration set and validation set

	样品数	平均值/%	最大值/%	最小值/%	标准差
校正集	206	1.97	2.35	1.26	0.28
验证集	103	2.00	2.29	1.30	0.24

1.3.2　特征波长选取及模型建立方法

分别采用广泛应用的线性建模方法-偏最小二乘法(partial least squares，PLS)以及最小二乘支持向量机(least squares-support vector machine，LS-SVM)算法建立校正模型，优选出适宜核桃露脂肪含量的建模方法。

LS-SVM 是对经典支持向量机算法的一种改进，可用于线性及非线性建模[8]。但当输入的变量过多且数据含有一定噪声时，会降低模型的性能及计算速度[9]。因此先应用遗传算法(genetic algorithms，GA)[10]结合向后间隔偏最小二乘法(backward interval PLS，BiPLS)[11]对全谱变量进行优化筛选，再结合 PLS 及 LS-SVM 算法建立核桃露脂肪含量的快速分析模型。

1.3.3　数据处理与分析

SNV 光谱预处理及 PLS 计算采用 UnscramblerX10.3 光谱分析软件(挪威 CAMO 公司)实现，BiPLS，GA，LS-SVM 等程序均在 MATLAB 环境下运行，模型的准确度与稳定性通过决定系数 R^2、预测标准偏差(root mean square error of prediction，RMSEP)及性能偏差比(ratio of performance to standard deviate，RPD)进行评价。R^2 越接近 1，RMSEP 越接近 0，RPD 越大，表明模型预测效果越好[12]。

2　结果与讨论

2.1　光谱波段优选

2.1.1　向后间隔偏最小二乘波段选择法(backward interval partial least squares，BiPLS)

考虑到运行遗传算法(GA)进行变量筛选，当筛选的波点数目较多时，可能导致筛选出的变量在后续建模时出现过拟合风险[13]。因此本实验先采用 BiPLS 方法对预处理后的整个谱区进行初步筛选。将核桃露全谱区变量共 1 501 个波长点等距划分为 k 个区间(k=15～30，间距为 5)，在各划分条件下采用 BiPLS 进行处理。当分割数为 20 时，BiPLS 所得交互验证均方差(RMSECV)值最小为 0.042 7，入选区间为[2 7]，所对应的变量区间为 4 304～4 600 和 5 804～6 100 cm^{-1}，共计挑选出 150 个变量点，共占全谱区的 10.0%。

2.1.2　遗传算法(genetic algorithms，GA)

采用 GA 对经 BiPLS 筛选出的核桃露脂肪指标的建模变量进行继续的选择。运行 GA 程序的参数被设置为：初始种群 50，变异概率 Pm=0.01，交叉概率 Pc=0.5，最大因子数 10，遗传迭代次数 100 次，计算每个数据点标识为“1”的概率，图 1 显示了 GA 程序运行过程中各波长变量被选用的频次，以 RMSECV 值确定出最佳的建模变量，选择了被选用 9 次以上的变量点共计 30 个，共占全谱区的 2.0%。

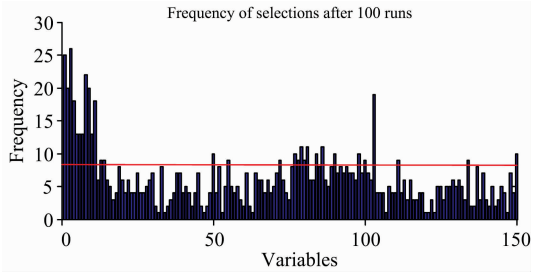

图 1　各变量被选用的频次图

Fig. 1　The frequency of each variable by chosen

2.2　偏最小二乘法(partial least squares，PLS)分析模型的建立与评价

在以上各方法进行波段挑选的基础上，分别建立核桃露饮品脂肪指标的全光谱-PLS，BiPLS-PLS 以及 GA-BiPLS 的快速分析模型，对模型效果进行综合评价分析，具体指标效果如表 2 所示。

表 2 核桃露脂肪指标的不同 PLS 模型及性能分析结果

Table 2 Different PLS models of fat in walnut beverage and the performance evaluation result

模型	模型光谱区间/cm^{-1}	变量数	R^2	RMSEP	RPD	主成分数
PLS(全谱)	4 000～10 000	1 501	0.964	0.049	4.88	10
BiPLS	4 196～4 596，5 800～6 196	150	0.973	0.043	5.62	6
GA-BiPLS	4 304，4 308，4 312，…	30	0.974	0.040	6.00	5

从表 2 中能够看出，相较于全光谱谱区建立的 PLS 模型，经 BiPLS 及 GA-BiPLS 波段优选后建立的模型效果有较为明显的提升。BiPLS 及 GA-BiPLS 模型的 RMSEP 值从 0.049 分别下降到 0.043 和 0.040，分别降低了 12.2% 及 18.4%，同时模型的相关系数 R^2 从 0.964 提高到 0.973 及 0.974，性能偏差比 RPD 从 4.88 增长到 5.62 及 6.00，主成分数也有不同程度的减少，都象征着模型准确度及性能均有较大的改善。同时模型的建模变量数从 1501 分别显著降低到 150 及 30，表明波段筛选方法能够在简化模型、提升模型运算速度的同时有效地对模型的整体性能进行优化。

而对于 BiPLS 模型而言，GA-BiPLS 进一步优化变量所建立的模型效果相对更为优秀，图 2 显示了 GA-BiPLS 方法筛选的变量在核桃露完整光谱中的分布情况，从图中可以看出，核桃露的近红外全谱在波数为 5 164 及 6 884 cm^{-1} 附近有明显的吸收峰，这些均是由于核桃露样品中的水分吸收所引起的，分别是 O－H 伸缩和 HOH 弯曲的组合频，以及 O—H 伸缩的一级倍频吸收[14]。经波段筛选后的变量排除了这些干扰，避免了全谱建模时水分的吸收对于核桃露脂肪含量的检测。同时 GA-BiPLS 方法筛选的波长与 C—H，O—H 等主要官能团的基频与组合频振动吸收峰位置相对应，如脂肪烃在 4 314 cm^{-1} 处 C—H 的组合频吸收，线性脂肪族分子的亚甲基在 5 800 cm^{-1} 处的组合频吸收，以及 5 872 和 5 905 cm^{-1} 处脂肪族直链甲基及端甲基的 C—H 倍频吸收等[15]，体现出了核桃露样品中脂肪成分的特征吸收，表明了 GA-BiPLS 方法优化的建模变量保留了脂肪的主要吸收谱带，同时避免了 BiPLS 方法在两个相邻波长或是小的区间中存在的共线变量，剔除了大量与脂肪指标无关的干扰信息，减少模型复杂程度的同时提高了预测精度。

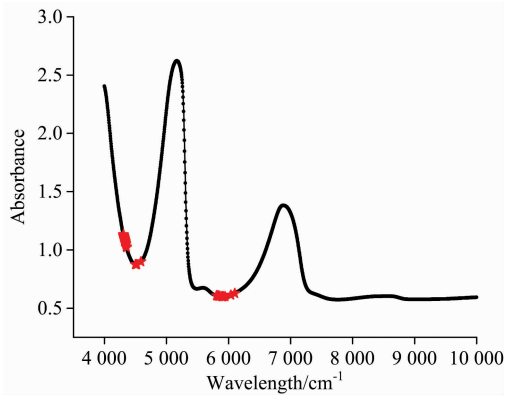

图 2 GA-BiPLS 选择的变量分布情况

Fig. 2 The distribution diagram of variables selected by GA-BiPLS

2.3 最小二乘支持向量机(least squares-support vector machine，LS-SVM)分析模型的建立与评价

依据上述经 GA-BiPLS 筛选出的优化波长作为核桃露脂肪指标的输入变量，建立 LS-SVM 校正模型。实验选用 LS-SVM 建模过程中较为常用的径向基函数(RBF)作为核函数，同时采取耦合模拟退火算法和留一交叉验证的寻优方法对 RBF 核函数中的两个重要调节参数—正则化参数 γ 以及核参数 σ^2 进行优化，以提高模型的泛化性及预测精度。LS-SVM 模型的具体指标效果如表 3 所示。

表 3 核桃露脂肪指标的 LS-SVM 建模结果

Table 3 Result of LS-SVM model for fat in walnut beverage

指标	γ	σ^2	R^2	RMSEP	RPD
总酸	132 304.7	532.148	0.986	0.036	6.52

从表 4 中可以看出，将经 GA-BiPLS 优化的有效波长作为输入变量建立的 LS-SVM 模型有着良好的效果，R^2 达到了 0.986 的同时，RMSEP 及 RPD 为 0.036 及 6.52。为了更直观的对 PLS 及 LS-SVM 的模型效果进行比较，对 GA-BiPLS 模型及 LS-SVM 模型的各项评价指标绘制的柱状图如图 3 所示。

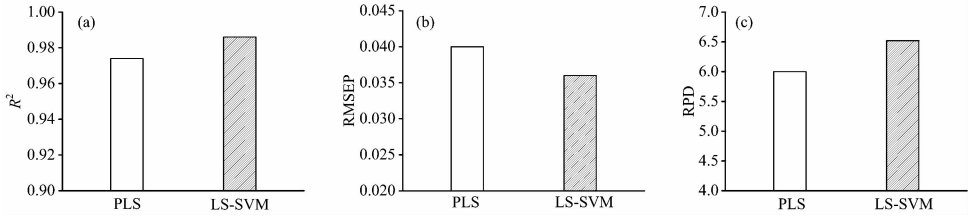

图 3 PLS 与 LS-SVM 模型评价参数比较图

(a)：相关系数；(b)：预测标准偏差；(c)：性能偏差比

Fig. 3 The performance parameters contrast of the PLS model and the LS-SVM model

(a)：R^2；(b)：RMSEP；(c)：RPD

如图 3 所示，可以清晰的看出相较于 PLS 模型，核桃露脂肪指标的 LS-SVM 模型表现出了更好的效果，可能是由于 PLS 作为一种经典的线性建模方法，在建立模型的过程中忽略了样品数据集中的非线性因素，而核桃露样品光谱测量过程中噪声、背景等因素的干扰，以及各指标成分间的相互影响一 使得脂肪含量与近红外光谱信息间存在复杂非线性的变化关系时，LS-SVM 方法能够更为有效地对其进行描述，增强了光谱变量与指标浓度间的相关性，使得建立的模型有着更好的准确度以及普适性。

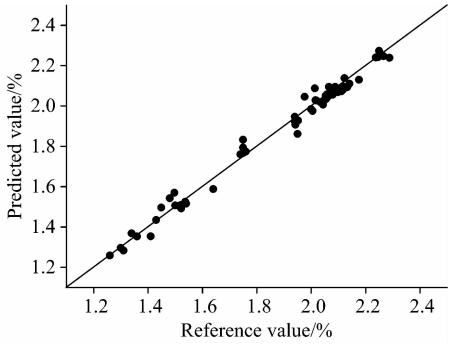

图 4　核桃露脂肪 LS-SVM 模型理化值与预测值分布

Fig. 4　Predicted *vs.* reference values of fat in walnut dew by LS-SVM model

2.4　LS-SVM 模型验证

将 60 个随机进行保留，未参与校正模型建立的核桃露样品光谱带入 LS-SVM 模型中进行验证，如图 4 所示，脂肪指标的实测与预测值数据点总体呈对角线分布，同时通过成对 t 检验，脂肪指标预测值与实测值间无明显差异（$p>0.90$）。LS-SVM 验证模型 R^2 为 0.987，RMSEP 为 0.034，说明模型预测结果较为准确。

3　结　论

采用近红外光谱分析技术结合化学计量学方法对核桃露的脂肪指标进行了定量分析，得出以下结论：

向后间隔偏最小二乘法及遗传算法两种变量优化方法所建模型均表现出了较高的模型精度，同时两种方法相结合所筛选出的关键变量体现了核桃露样品中脂肪成分的特征吸收，充分说明了变量筛选对建模分析的重要性。

与建立的 PLS 模型相比，发现 LS-SVM 所建模型有更为优秀的优化效果，考虑到传统算法忽略复杂非线性关系的这种缺陷以及 LS-SVM 方法广泛的适应能力，说明在实际生产中，LS-SVM 方法具备优良的可行性，体现了其在核桃露饮品品质分析方面的巨大潜力，为核桃露生产的质量监控提供了技术借鉴，同时为饮品品质的分析方法研究提供了新的思路。

References

[1] WEI Xiao-lu, HUANG Xin, FENG Yue, et al(魏晓璐，黄　鑫，冯　悦，等). Science and Technology of Food Industry(食品工业科技), 2014, 35(13): 288.
[2] WU Xiao-ju, XIE Ya-li(吴晓菊，谢亚利). Farm Products Processing(农产品加工), 2013, (8): 66.
[3] Li Zongpeng, Wang Jian, Xiong Yating, et al. Vibrational Spectroscopy, 2016, 84: 233.
[4] WANG Hui, LU Jian-liang(王　会，陆建良). Food Research and Development(食品研究与开发), 2016, 37(24): 132.
[5] GU Ru-xiang, ZHAO Wu-qi, SHI Ke-xin, et al(谷如祥，赵武奇，石珂心，等). Science and Technology of Food Industry(食品工业科技), 2013, 34(20): 75.
[6] TIAN Jing, LI Qiao-ling(田　晶，李巧玲). Food Science(食品科学), 2018, 39(2): 293.
[7] Zhang L, Li G, Sun M, et al. Infrared Physics & Technology, 2017, 86.
[8] ZOU Xiao-bo, CHEN Zheng-wei, SHI Ji-yong, et al(邹小波，陈正伟，石吉勇，等). China Brewing(中国酿造), 2011, 30(3): 63.
[9] ZHANG Hai-liang, LIU Xue-mei, HE Yong(章海亮，刘雪梅，何　勇). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2014, 34(5): 1348.
[10] Aouadni I, Rebai A. Annals of Operations Research, 2017, 256(1): 1.
[11] Zhou S, Yin Q, Lu L, et al. Infrared Physics & Technology, 2017, 80.
[12] Rungpichayapichet P, Mahayothee B, Nagle M, et al. Postharvest Biology & Technology, 2015, 111: 31.
[13] Wu Z, Ma Q, Lin Z, et al. Talanta, 2013, 107(2): 248.
[14] CHEN Li-dan, ZHAO Yan-ru(陈立旦，赵艳茹). Transactions of the Chinese Society of Agricultural Engineering(农业工程学报), 2014, 30(8): 168.
[15] Jerry Workman, Jr Lois Weyer. Practical Guide to interpretive Near-Infrared Spectroscopy(近红外光谱解析实用指南). Translated by CHU Xiao-li, XU Yu-peng, TIAN Gao-you(褚小立，许育鹏，田高友，译). Beijing: Chemical Industry Press(北京：化学工业出版社), 2009. 240.

Determination of Fat in Walnut Beverage Based on Least Squares Support Vector Machine

LI Zi-wen[1], LI Zong-peng[1], MAI Shu-kui[1], SHENG Xiao-hui[1], YIN Jian-jun[1], LIU Guo-rong[2], WANG Cheng-tao[2], ZHANG Hai-hong[3], XIN Li-bin[4], WANG Jian[1]*

1. China National Research Institute of Food and Fermentation Industries Corporation, Beijing 100015, China
2. Beijing Engineering and Technology Research Center of Food Additives, Beijing Technology and Business University, Beijing 100048, China
3. College of Agriculture, Ningxia University, Yinchuan 750021, China
4. Shanghai Precise Packaging Corporation, Shanghai 201514, China

Abstract Near-infrared spectroscopy was used to quantitatively analyze the fat content of walnut beverage. At the same time, modeling variables were optimized and modeling methods were compared to optimize the best model. In order to eliminate the influence of scattering on the spectrum, the data are preprocessed by the standard normal transformation (SNV) method. The preferred characteristic wavelengths of genetic algorithms (GA) combined with backward interval partial least squares (BiPLS) were used as input variables of partial least squares (PLS) and least squares support vector machine (LS-SVM) respectively to establish model of fat content in walnut beverage. The R^2, RMSEP and RPD were used to evaluate the effect of spectral band selection method on the construction of fat index model in walnut beverage and determine the best modeling method. The results showed that the variable selection could optimize the model. 150 and 30 variable points corresponding to the characteristic absorption peaks of the fat components in walnut beverage samples were selected by BiPLS and GA-BiPLS methods, respectively, accounting for 10% and 2% of the full spectrum. The RMSEP value of the PLS model decreased from 0.049 to 0.043 and 0.040, respectively, and the R^2 increased from 0.964 to 0.973 and 0.974. The range error ratio RPD increased from 4.88 to 5.62 and 6.00, and the principal component number also decreased to varying degrees. The method of variable selection could reduce model dimensions and improve model accuracy. Compared with the PLS model, the R^2, RMSEP and RPD values of the LS-SVM model showed better results, reaching 0.986, 0.036 and 6.52, respectively. The LS-SVM model has higher accuracy and stability than the PLS model. Since PLS is a classic linear modeling method, the nonlinear factors in the sample data set are ignored in the process of building the model. However, there was a complex nonlinear relationship between fat content and near-infrared spectral information, which is due to the interference of noise, background and other factors in the spectral measurement process of walnut beverage samples and the interaction between various indicators. The LS-SVM method could enhance the correlation between spectral variables and index concentration, so that the established model has better accuracy and universality. It shows that in the actual production, the LS-SVM method has excellent feasibility, which reflects its great potential in the analysis of the quality of walnut beverage. Based on the LS-SVM method, the quantitative analysis model of walnut fat content has accurate and stable characteristics, which can provide technical reference for the quality monitoring of walnut beverage production, and provide a new idea for the analysis of beverage quality.

Keywords Walnut beverage; Near-infrared spectroscopy; Least squares support vector machines(LS-SVM); Band selection

(Received Oct. 22, 2018; accepted Feb. 16, 2019)

* Corresponding author