基于近红外光谱和随机森林方法鉴别蜂蜜真伪

莫菲凡[1], 范 伟[2], 周冀衡[2], 梁逸曾[3*]

(1. 湖南师范大学附属中学, 长沙 410014; 2. 湖南农业大学生物科学与技术学院生物品质安全联合实验室, 长沙 410128; 3. 中南大学化学化工学院, 长沙 410083)

摘 要: 目的 建立蜂蜜样品真伪鉴别的近红外光谱快速检测方法, 为今后蜂蜜检验工作提供可靠参考依据。**方法** 采用积分球透反射模式采集样品近红外光谱数据, 以 Savitzky-Golay 1 阶微分方法对原始光谱进行预处理, 以随机森林方法建立光谱数据与蜂蜜真伪的定性判别模型。**结果** 所建立的判别模型中训练样本判别正确率为 100%, 测试样本判别正确率为 95%。**结论** 近红外透反射光谱技术应用于蜂蜜真伪鉴别的可行性, 同其他分析方法相比具有操作简单、速度快、效率高、无污染、费用低、无需复杂前处理等优点。

关键词: 近红外光谱法; 随机森林法; 蜂蜜; 掺伪鉴别

Detection of honey adulteration by near infrared spectroscopy coupled with random forest method

MO Fei-Fan[1], FAN Wei[2], ZHOU Ji-Heng[2], LIANG Yi-Zeng[3*]

(1. *The High School Attached to Hunan Normal University, Changsha* 410014, *China*; 2. *Joint Lab for Biological Quality and Safety, College of Bioscience and Biotechnology, Hunan Agriculture University, Changsha* 410128, *China*; 3. *College of Chemistry and Chemical Engineering, Central South University, Changsha* 410083, *China*)

ABSTRACT: Objective To study honey adulteration by near infrared spectroscopy-chemometrics method. **Methods** Transflectance mode with integrating sphere was used to collect NIR spectra. Savitzky-Golay first derivative method was employed to preprocess the raw NIR spectra. A Random Forest model was established based on the correlation between the NIR spectra and reference value. **Results** The prediction accuracy of the proposed method was 100% for training samples and 95% for test samples. **Conclusion** The proposed method is fast and non-destructive, and it provides a novel efficient and environmental friendly approach for the rapid qualitative detection of honey adulteration.

KEY WORDS: near infrared spectroscopy(NIR); Random Forest method; honey; adulteration detection

1 引 言

我国是世界最大的养蜂国, 蜜蜂饲养量为 750 万群, 约占世界蜜蜂饲养量的十分之一, 年产蜂蜜超过 20 万吨, 约占世界总产量的五分之一, 年出口蜂蜜 6~8 万吨, 约占世界贸易量的四分之一。我国传统观点认为蜂蜜具有多种保健功能, 但是由于蜂蜜造假成本低、利润高、危险性较低, 使一些假蜂蜜屡禁不绝, 尤其是在中小城市及县乡一级市场。从每年全国蜂蜜的普查结果来看, 真蜂蜜不足总量的三分之一[1]。

基金项目: 国家自然科学基金项目(21275164)

Fund: Supported by the National Natural Science Foundation of China (21275164)

*通讯作者: 梁逸曾, 教授, 主要研究方向为化学计量学。E-mail: yizeng_liang@263.net

***Corresponding author:** LIANG Yi-Zeng, Professor, College of Chemistry and Chemical Engineering, Central South University, No 932, Lushan South Road, Yuelu District, Changsha 410083, China. E-mail: yizeng_liang@263.net

目前蜂蜜造假手段主要是利用粮食作物加工成糖浆(也叫果葡糖浆)充当蜂蜜。为以假乱真，造假分子还会在假蜂蜜中加入增稠剂、甜味剂、防腐剂、香精和色素等化学物质。这些造假蜂蜜外表与真蜂蜜极为相似，难以辨别，需要复杂的理化实验，才能判定是否掺杂使假。但样品的预处理步骤繁杂，周期长，成本高，易受人为因素的干扰，急需快速无损的检测技术[2]。

近红外光谱(NIRS)分析技术是近年来迅速发展的一项绿色分析技术，因其具有分析速度快、成本低、无污染、无需复杂预处理及多组分同时测定等一系列优点而广泛应用于农业、食品、医药等诸多行业[3,4]。近红外光谱主要为 C-H, N-H, O-H 等含氢基团的倍频和合频吸收，适用于食品化学成分的检测。

近红外光谱的分析应用关键是建立相应的校正模型[5-7]。这个过程需要应用化学计量学的模式识别方法来实现，包括主成分分析(PCA)、聚类分析(CA)和随机森林(RF)等，这些方法被广泛用于鉴别各个类型的样本。其中随机森林是由著名统计学家 Leo Breiman 提出的一种基于分类回归树分类器的集成学习算法，其通过有放回地采样构建多个训练集，最后的预测结果由所有构建分类器进行投票表决得到。随机森林方法具有训练速度快、不易过拟合、对包含奇异值和噪声的数据预测结果比较稳健等优点，目前已广泛应用于多个领域[8]。在本文中，我们以蜂蜜为研究对象，采用近红外光谱结合随机森林方法建立一种蜂蜜真伪的鉴别判定方法，为其监管提供有效手段。

2　材料与方法

2.1　材料与仪器

本次实验的 120 个蜂蜜样本来自于湖南名园蜂业，包括荆条蜜、柑橘蜜、洋槐蜜、荔枝蜜、枣花蜜、油菜蜜 6 个品种，每个品种 20 个样本。每个品种取 10 个样本，共计 60 个作为真实样本，另外 60 个样本人为用果葡混合糖浆进行掺假。掺假比例为 1%, 5%, 10%, 20%, 35%和 50%。每个比例水平掺假 10 个样本。样本处理完毕后按照 SPXY[9]方法划分, 100 个作为训练集, 20 个作为测试集。

Antaris Ⅱ傅立叶变换近红外光谱仪(Thermofisher, 美国)。

2.2　样品光谱采集

蜂蜜样品透明而较为粘稠，难以直接采用漫反射和透射模式进行测量。因此本文采用积分球透反射测量方式进行光谱采集，同时，为避免样品不均匀，产生误差，每次测量样品杯旋转 120 度。光谱数据取 3 次采样的平均值，整个实验过程保持室内温度在 25 ℃左右。仪器参数设定如下：扫描范围：10000~4000 cm^{-1}，分辨率：8 cm^{-1}，扫描次数为 32 次。采集光谱模式示意图如图 1 所示，图 2 为扫描所得光谱。

2.3　随机森林方法

随机森林方法是基于 Bagging(Bootstrap aggregating)方法发展而来。Bagging 算法是由统计学家 Leo Breiman 在 1996 年提出的一种组合分类器算法[10,11]，其基本思想如下：Bagging 算法实际可以理解为有放

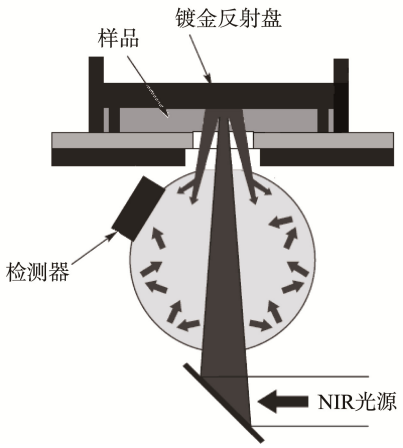

图 1　蜂蜜透反射模式测量示意图

Fig. 1　Transflective mode for honey spectra collection

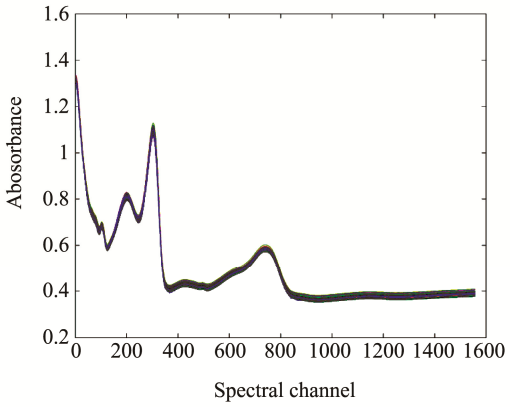

图 2　蜂蜜近红外原始光谱

Fig. 2　Original NIR spectra of honey

回的采用，在确定了元分类器后，从训练样本集 S_{Train} 中，任意提取 S_k 个样本，用于构建单个元分类器。这样 S_{Train} 中的某一样本可能在 S_k 中被多次选择或有可能不会被选择。其采样过程和模型构建过程如图 3 所示。先从训练样本中提取出一部分样本构建子分类器，重复这个过程直到构建出全部的分类器，每个分类器将得到一个分类结果，对所有的结果进行统计分析得到最终结果，并从中得到分类模型的分类规则，进而对独立测试集或者新的数据集进行分析。

随机森林是在 Bagging 方法基础上的进一步发展[12,13]，其在构建每个独立树分类器的时候并不是使用所有的变量，而是随机的从所有变量中选择一部分进行节点的劈分。基于样本及其分类标识和变量，随机森林可以训练出一系列单个的决策树。因为森林中的每一棵树都是随机选取样本和变量的子集而建成，因而叫做随机森林。

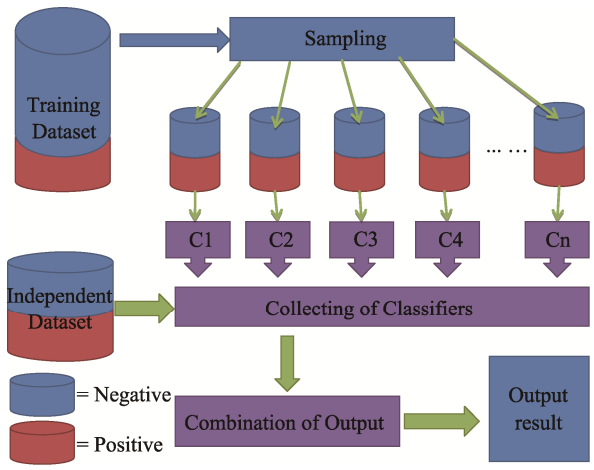

图 3 组合分类器的构建

Fig. 3 Construction of combinatorial classifier

3 结果与讨论

3.1 光谱预处理

光谱测量过程中，样本颗粒大小和杂散光等因素会造成光谱的漂移，因此，在建立校正模型之前，需要对光谱进行预处理。本文中，我们采用 Savitzky-Golay 一阶微分对光谱进行预处理，SG 一阶微分可以有效消除光谱漂移影响，增强与成分含量相关的光谱吸收信息。该方法的使用首先要确立一个合适的窗口，对该窗口内的光谱进行多项式拟合，之后对拟合的多项式求取微分光谱。在本研究中，以训练集光谱作为标准，对所有光谱进行 9 点 2 次 SG 微分处理。结果如图 4 所示，对比图 2 可知，经过预处理之后，光谱的漂移得到一定程度的校正，有效消除了样品影响所导致的基线平移和偏移现象，提高了光谱的信噪比。

3.2 主成分分析模型

主成分分析方法本质是通过一种最优化方法浓缩及综合光谱数据中的信息，从而简化数据、降低维数，进而揭示出光谱数据的内部结构特征，具体看来，主成分分析通过将原来光谱数据进行转换，得到若干由原始光谱数据线性组成的新变量，新变量尽可能多地表征原光谱信息的数据特征而不丢失信息。经转换得到的新变量是相互正交的，即新变量间互不相关，以消除众多信息共存中相互重叠的信息部分。图 5 为预处理之后光谱主成分分析的三维得分图，从

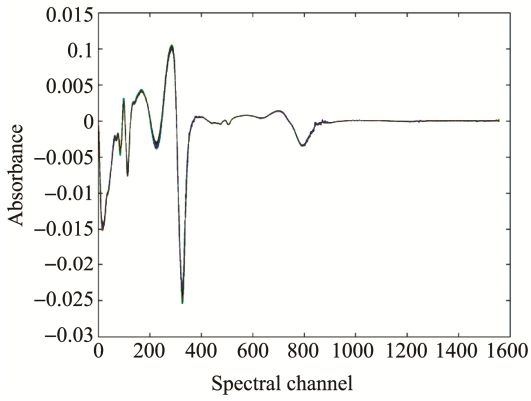

图 4 SG 一阶微分处理后的光谱图

Fig. 4 Spectra of all samples after Savitzky-Golay first derivative processing

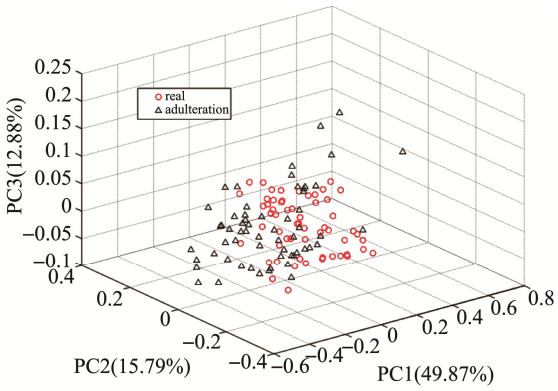

图 5 主成分分析图

Fig. 5 Score plot of principal component analysis

图中可以看出，真伪蜂蜜样本混杂在一起，难以有效区分。因此，需要采用更加有效的随机森林方法。

3.3　随机森林判别模型

在使用随机森林建模的过程中，首先要选择的参数就是训练过程中生长树的数量。在本研究中，我们选定的参数为 500，即在训练的过程中生长 500 棵树，并集成其结果构造森林。我们将 10 折交互验证的误判率对生长树的数目作图，结果如图 6。由图 6 可知，误判率并不随着生长树的数目增多而下降，也就是说，其训练结果不会过拟合，其训练错误率将趋近于某一数值，而不会达到 0。从图 6 可进一步发现，当生长的树的数目达到 120 时，其结果趋近于最小值，且趋于稳定，因而本实验选择 120 颗树作为随机森林模型参数。

确定随机森林方法参数后，建立定性鉴别模型，由表 1 可知，其训练集的样本判别正确率为 100%，预测集的样本判别正确率为 95%。和传统的线性判别方法相比，随机森林方法需要优化的参数少，判别的正确率更高[14]，掺伪低至 1%的样本也可以被鉴别，而且该方法具有较好的抗过拟合的能力，稳定性好，易于推广。

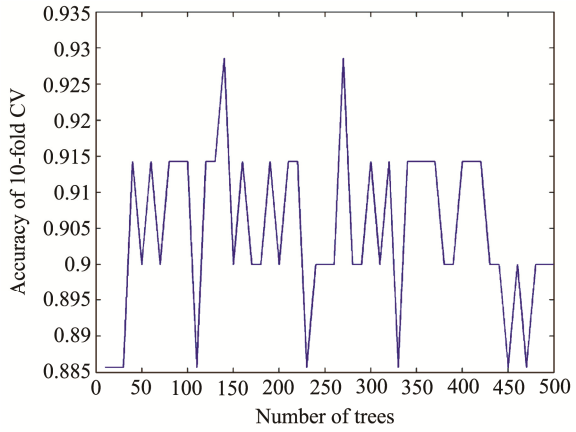

图 6　选择随机森林中树生长的数目

Fig. 6　Trees used in Random Forest

表 1 样本判别结果

Table 1　Results for all samples

	真实样本/判别正确样本个数	掺伪样本/判别正确样本个数	判别正确率 (%)
训练集样本	50/50	50/50	100%
测试集样本	10/9	10/10	95%

随机森林方法除了进行分类判别，其还可以计算分类的变量重要度，重要度的估计对数据集中的近红外波长相关性的解释十分有用。在本研究中，随机森林评估出的重要变量如表 2 所示。

表 2 随机森林方法所评估的重要波长

Table 2　Important wavelengths selected by RF

波数 (cm^{-1})
4180.9 5075.7 5091.2 5183.7 5199.1 5338 5415.1 5430.6 5893.4 7235.6 7312.7 7328.2 7467 9549.8

在这些波长中，9549.8 cm^{-1} 与 R-C-OH 基团中的 O-H 振动相关，而 7328、7312 cm^{-1} 与 $ArCH_3$ 基团的 C-H 振动有关，5893.4 cm^{-1} 则可能与 $COCH_3$ 基团的 C-H 振动有关。5199.1 cm^{-1} 和 5183.7 cm^{-1} 同样来自于 O-H 基团的吸收，5091.2 cm^{-1} 则为 O-H 基团和 C-H 基团的组频吸收，5075.7 cm^{-1} 来自于 N-H 的振动吸收，4180.9 cm^{-1} 可能与芳香环的 C-H 振动相关[15]。所有这些重要波长均与单糖或者氨基酸有关，而这些成份也是真实蜂蜜和掺假糖浆之间的区别。

4　结　论

实验结果表明，近红外透反射光谱结合随机森林判别方法适用于蜂蜜真伪的鉴别，而且能够达到满意的检测精度。

随机森林方法需要优化的参数少，判别的正确率更高，掺伪低至 1%的样本也可以被鉴别，而且该方法具有较好的抗过拟合的能力，稳定性好，易于推广。

综上所述，本研究方法成本低，操作简便，在执行过程中不使用化学试剂可有效减少环境污染和人身危害，在未来食品监管中具有很好的推广前景。

参考文献

[1] 李水芳，文瑞芝，尹永，等. 近红外光谱法定性定量检测蜂蜜中掺入甜菜糖浆的可行性研究[J]. 光谱学与光谱分析，2013, 33(10): 2637–2641.
Li SF, Wen RZ, Yin Y, *et al*. Qualitative and QuantitativeDetection of Beet Syrup Adulteration of Honey by Near-Infrared Spectroscopy:A Feasibility Study[J]. Spectros Spectr Anal, 2013, 33 (10): 2637–2641.

[2] 梁秀英. 基于 NACA 和 LS-SVM 的蜂蜜真伪识别[J]. 湖北农业科学, 2014, 53(2): 430–433.

Liang XY. Identification of Honey Authenticity based on LS-SVM and NACA[J]. Hubei Agric Sci, 2014, 53(2): 430–433.

[3] Bokobza L. Near Infrared Spectroscopy[J]. J Near Infrared Spectrosc, 1998, 6: 3–17.

[4] Batten GD. An Appreciation of the Contribution of NIR to Agriculture[J]. J Near Infrared Spectrosc, 1998, 6: 105–114.

[5] Daszylowski M, Kaczmarek K, Heyden YV, *et al*. Robust Statistics in Data Analysis-A Review: Basic Concepts[J]. Chem Intel Lab Sys, 2007, 85(2): 203–219.

[6] Hubert M, Vanden Branden K. Robust Methods for Partial Least Squares Regression[J]. J Chemometr, 2003, 17(10): 537–549.

[7] Naes T, Isaksson T, Fearn T, *et al*. A User-Friendly Guide to Multivariate Calibration and Classification[M]. Chichester: NIR Publications, 2004.

[8] Ai FF, Bin J, Zhang ZM, *et al*. Application of random forests to select premium quality vegetable oils by their fatty acid composition[J]. Food Chem, 2014, 143: 472–478.

[9] Galvão RKH, Araujo MCU, José GE, *et al*. A method for calibration and validation subset partitioning[J]. Talanta, 2005, 67(4): 736–740.

[10] Breiman L. Bagging predictors[J]. Mach Learning, 1996, (24): 123–140.

[11] Dietterich TG. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization[J]. Mach Learning, 2000, (40): 139–157.

[12] Breiman L. Random Forests[J]. Machine Learning, 2001, (45): 5–32.

[13] Díaz-Uriarte R, Andrés SA. Gene selection and classification of microarray data using random forest[J]. BMC Bioinform, 2006, (7): 3–16.

[14] 马奕颜, 郭波莉, 魏益民, 等. 植物源性食品原产地溯源技术研究进展[J]. 食品科学, 2014, (35): 246–250.
Ma YY, Guo BL, Wei YM, *et al*. An Overview of Analytical Approaches for Determining the Geographical Origin of Plant-derived Foods[J]. Food Sci, 2014, (35): 246–250.

[15] Workman JJ, Weyer L. Practical guide to interpretive near-infrared spectroscopy[M]. Florida: CRC Press, 2008.

(责任编辑: 赵静)

作者简介

莫菲凡，湖南师范大学附属中学，主要研究方向为食品与环境分析。
E-mail: 724184935@qq.com

梁逸曾，教授，主要研究方向为化学计量学。
E-mail: yizeng_liang@263.net