

文章编号:1671-1513(2011)01-0075-04

基于近红外光谱的纯花生油掺伪快速鉴别方法研究

吴静珠, 刘翠玲, 李 慧, 孙晓荣

(北京工商大学 计算机与信息工程学院, 北京 100048)

摘 要: 针对目前国内缺乏快速鉴别花生油掺伪鉴别技术的现状, 提出基于近红外光谱的纯花生油掺伪快速鉴别方法. 实验分别配制了掺入大豆油、菜籽油、棕榈油和调和油的4类掺伪花生油样品共40个, 纯花生油样品5个, 采集样品近红外全谱, 通过支持向量机技术建立纯花生油掺伪鉴别模型. 结果表明, 选取径向基函数为支持向量机核函数, 通过网格搜索和 k 折校验法确定核参数 γ 为1, 惩罚参数 c 为1 024, 建立纯花生油掺伪鉴别模型的识别率和预测率均达到100%, 基于近红外光谱的花生油掺伪快速检测技术具有较好的可行性和实用性.

关键词: 纯花生油; 掺伪鉴别; 近红外光谱; 支持向量机

中图分类号: TS207.3

文献标志码: A

花生油是我国第三大食用油, 以其营养丰富著称, 在市场上属于价格较高的食用油之一. 目前市场上频频出现不法商贩为谋取暴利, 将价格相对较低的油脂如菜籽油、大豆油和棕榈油等掺入花生油进行销售的现象. 为了保护消费者权益以及维护市场秩序, 在加强对花生油生产和销售监管的同时, 研究花生油掺伪快速鉴别方法在目前形式下尤为重要和迫切.

目前国内检测花生油掺伪的方法主要有常规理化检测方法和气相色谱法等. 这些方法具有较高的准确度和可靠性, 但需借助于昂贵的设备或是严格的实验室条件对样品进行复杂的前处理, 且分析速度慢, 不能满足市场快速检测的需要. 20世纪70年代迅速发展起来的近红外光谱分析技术(Near Infra-red Spectroscopy, 简称NIR)是利用近红外谱区包含的物质信息, 用于有机物质定性和定量分析的一种分析技术^[1]. 该方法具有快速、高效、无污染、无需前处理、无损分析及多组分同时测定等优点. 西方国家最早将近红外技术应用于油脂检测, 在橄榄油、坚果油及其他植物油的检测中已得到成功应用. 我

国刘福莉^[1]以8种食用油纯油的43个样品为对象, 采用近红外透射光谱结合聚类分析法快速鉴别食用油种类, 判别模型对预测集样品的准确率达到100%. 范璐^[2]等人利用气相色谱和傅里叶变换红外吸收光谱, 对21种花生油、20种棕榈油及两者的4种调和油进行分析, 对花生油和棕榈油做了科学的识别分析.

现有的基于近红外食用油定性鉴别研究主要是针对食用油种类, 而对食用油掺伪鉴别的研究甚少. 但是目前国内将低价油掺入高价油的现象频频出现, 屡禁不止. 基于国内实际情况, 本文研究基于近红外光谱技术的花生油掺伪快速鉴别方法, 探索其作为一种快速无损鉴别食用油掺伪方法的可行性.

1 材料与方法

1.1 样品制备

综合市场上目前出现的在花生油中掺伪的主要是大豆油、菜籽油、棕榈油等植物油, 因此购买了如下样品: 花生油(超市, 鲁花压榨一级, 非转基因), 大豆油(超市, 福临门一级浸出, 转基因), 菜籽油

收稿日期: 2010-12-31

基金项目: 北京市优秀人才培养资助项目(20081D0500300130).

作者简介: 吴静珠, 女, 副教授, 博士, 主要从事基于近红外光谱的农产品及食品检测技术方面的研究.

(超市,洪井源,非转基因),棕榈油(市场),调和油(超市,福临门天然谷物调和油)。

分别将大豆油、菜籽油、棕榈油、调和油以一定比例掺入纯花生油。以大豆油为例,掺伪比例如表 1。菜籽油、棕榈油、调和油的掺伪比例同大豆油,样品编号分别以 hcXX,hzXX,htXX 命名。实验共配置掺伪样品 40 个。纯花生油样品以 hsXX 编号,共采集 5 份纯花生油样品。

表 1 大豆油掺入花生油的配比

Tab. 1 Mix proportion of soybean oil and peanut oil

编号	$V_{花生油}/\text{mL}$	$V_{大豆油}/\text{mL}$	掺伪比例/%
hd01	45	5	10
hd02	40	10	20
hd03	35	15	30
hd04	30	20	40
hd05	25	25	50
hd06	20	30	60
hd07	15	35	70
hd08	10	40	80
hd09	5	45	90
hd00	0	50	100

1.2 光谱采集

实验采用德国 Bruker 公司的 VERTEX 70 红外光谱仪。采集光谱波数范围:4 000 ~ 12 500 cm^{-1} ,分辨率:8 cm^{-1} ,每个样品重复扫描次数:32 次,采样点:1 102。使用液体光纤探头采样,光程:2 mm。

全部样品未经任何化学处理,将光纤探头插入装有样品的小瓶中,逐一扫描样品,每次测量前均用石油醚清洗探头,避免样品间交叉污染。以表 1 中的样品为例,其近红外谱图如图 1。

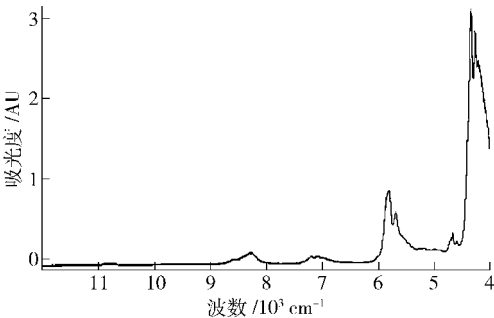


图 1 掺伪(大豆油)花生油样品近红外谱图

Fig. 1 Near-infrared spectrum of adulteration peanut oil mixed with soybean oil

1.3 支持向量机分类器

基于近红外光谱的快速定性检测技术目前被逐渐应用于质量控制中,但是,样本少以及近红外光谱的高维数给近红外定性识别模型的建立带来很大困难。支持向量机^[3](Support Vector Machines,简称 SVM)是一种新的通用的机器学习方法,以其小样本下良好的推广能力而被广泛用于各种模式分类问题。

设训练集为 $\{x_i, y_i\}, i = 1, \cdots, n, x_i \in R^n, y_i \in \{-1, 1\}$, 则 SVM 分类器的一般形式如下所示:

$$f(x) = \text{sgn} \left\{ \sum_{i=1}^n a_i^* y_i K(x_i, x) + b \right\}, \quad (1)$$

其中, $K(x_i, x)$ 是核函数, a_i^* 通过在约束条件 $\sum_{i=1}^n y_i a_i = 0$ 和 $0 \leq a_i \leq c, i = 1, \cdots, n$ 下最大化下式(2)求得。其中, c 为惩罚参数。

$$Q(a) = \sum_{i=1}^n a_i - \frac{1}{2} \sum_{i,j=1}^n a_i a_j y_i y_j K(x_i x_j). \quad (2)$$

2 结果与分析

实验采用 MATLAB 7.0 和 LIBSVM 模式识别与回归的软件包(<http://www.csie.ntu.edu.tw/~cjlin/libsvm> 免费下载)实现 SVM 的分类算法,目前的最新版本为 libsvm-2.91。LIBSVM 使用的一般步骤如下。

- 1) 按照 LIBSVM 软件包所要求的格式准备数据集;
- 2) 对数据进行简单的缩放操作;
- 3) 考虑选用 RBF 核函数;
- 4) 采用交叉验证选择最佳参数 c 与 r ;
- 5) 采用最佳参数 e 与 r 对整个训练集进行训练获取支持向量机模型;
- 6) 利用获取的模型进行测试与预测。

本实验采用上述步骤逐步建立基于近红外-支持向量机的花生油掺伪鉴别模型。

2.1 数据的归一化

归一化有助于提高支持向量机的性能,可以避免数值范围较大的属性控制数值范围较小的属性,并且可以避免计算时的数值困难问题。本文采用的归一化的计算公式如下式所示:

$$z_{ij} = (OUT_{\max} - OUT_{\min}) \frac{x_{ij} - x_{j,\min}}{x_{j,\max} - x_{j,\min}} + OUT_{\min}, \quad (3)$$

其中, $x_{j, \max}, x_{j, \min}$ 是光谱矩阵 x 中第 j 列的最大值和最小值, $OUT_{\max}, OUT_{\min}$ 是归一化后输出范围设定的最大最小值, x_{ij}, z_{ij} 为归一化前后光谱矩阵. 本文设定归一化后光谱矩阵的数值在 $[-1, 1]$ 内, 则归一化公式如下式所示.

$$z_{ij} = 2 \frac{x_{ij} - x_{j, \min}}{x_{j, \max} - x_{j, \min}} - 1.$$

(4)

需要注意对预测集也要进行与训练集相同的归一化, 即对预测数据归一化时用的最大最小值是训练数据的最大最小值. 本文采用 MATLAB 提供的 premnmx 函数进行归一化.

2.2 核函数和参数的选择

本文采用 RBF 径向基函数为支持向量机的核函数, 通过网格搜索和 k 折交叉验证方法搜索最佳惩罚参数 c 和 RBF 核参数 γ 的最佳组合, 设定初步搜索范围 $c = [2 - 10, \cdots, 210]$, $\gamma = [2 - 10, \cdots, 210]$. 网格搜索过程在 Matlab 7.0 中编程实现. 实验设定参数如下.

纯花生油样品用 1 标记, 各种掺伪样品用 -1 标记.

测试集样品的选择: 纯花生油 hs02, hs04, 根据掺伪浓度的梯度选择掺大豆油样品 hc02, hc04, hc06, hc08, 掺菜籽油, 棕榈油, 调和油样品也同上类似选择; 其他样品为训练集样品.

SVM 类型采用 C-SVC, 交互校验采用 K 重交叉验证, $K = 5$.

根据上述设定的参数进行最佳参数的网格搜索, 结果如图 2. 图中所示曲线为参数组合 $\log_2(c)$ 和 $\log_2(\gamma)$ 下, SVM 分类模型的交叉校验预测率的等高线图, 等高线上标注的为预测率. 选取 SVM 模型交叉校验预测率为 100% 时的一组最佳参数 $c = 1024, \gamma = 1$.

2.3 模型建立与预测

在 Matlab 7.0 中调用 SVM 模式识别与回归的软件包 LIBSVM. “svmtrain” 和 “svmpredict” 是软件包中两个重要函数, 分别用来实现 SVM 的训练和预测过程. 根据网格搜索得到的最佳参数训练 SVM 模型, 在 MATLAB 中调用形式:

model = svmtrain (ytrain, xtrain, ' - c 1 024 - g

1');

模型参数为, 支持向量数 11, rho(即式 1 中的参

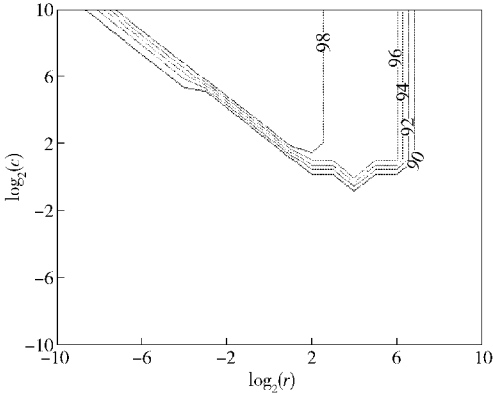


图 2 SVM 模型交叉校验预测率等高线图
Fig. 2 Contour map of cross validation prediction rate of SVM model

数 b) 为 0.683 9.
采用模型预测训练集和测试集, 调用形式为,
[predicte, accuracy] = svmpredict (ytrain, xtrain, model).

这里的训练集或是测试集的光谱数据都是经过归一化后的数据. 实验结果如下所示:

Accuracy = 100% (27/27) (classification)

(训练集),

Accuracy = 100% (18/18) (classification)

(测试集).

实验结果表明, 采用近红外光谱-SVM 建立花生油掺伪定性鉴别模型的预测效果较好.

2.4 与系统聚类的比较

目前近红外光谱分析中采用聚类分析法快速鉴别食用油种类研究较多, 效果也较好^[4]. 本文尝试采用系统聚类法对本实验中掺伪样品集近红外光谱进行鉴别分析, 光谱矩阵经标准化预处理后, 采用 OPUS 软件提供的 7 种不同的聚类方法: singl linkage, average linkage, weighted Av, L inkage, median algorithm, centroid algorithm 和 Ward's algorithm, 聚类结果均不理想. 以 Ward's Algorithmn, 为例, 聚类结果如图 3. 因此从本实验来看, 对于掺伪鉴别, 支持向量机较系统聚类法具有更好的性能.

3 小 结

本文根据市场上纯花生油主要的掺假类型, 配置了 4 种不同类型的纯花生油掺伪样品, 通过支持

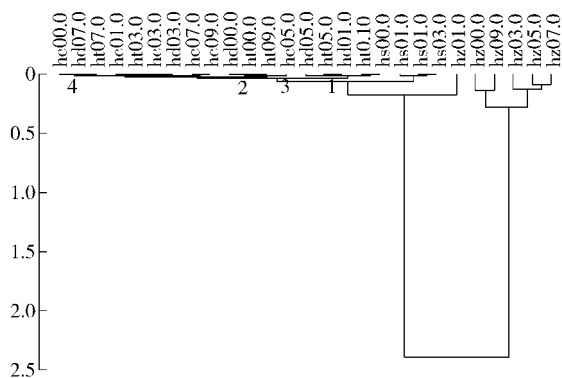


图3 Ward's Algorithmn 聚类结果图

Fig. 3 Cluster result based on Ward's Algorithm

向量机技术,建立了基于近红外光谱的纯花生油掺伪快速定性鉴别模型,对实验中预测集和测试集的预测率均能达到 100%。实验表明基于近红外光谱的纯花生油掺伪快速检测技术具有较好的可行性和

实用性。

参考文献:

- [1] 刘福莉,陈华才,姜礼义,等. 近红外透射光谱聚类分析快速鉴别食用油种类[J]. 中国计量学院学报, 2008,19(3):278-282.
- [2] 范璐,吴娜娜,霍权恭,等. 气相色谱分析和傅里叶变换红外吸收光谱识别花生油中棕榈油[J]. 河南工业大学学报:自然科学版,2008(1):12-16.
- [3] Cristianini N, Shawe-taylor J. 支持向量机导论[M]. 李国正,王猛,曾华军,译. 北京:电子工业出版社, 2004:1-176.
- [4] 吴静珠,刘翠玲,李慧,等. 近红外光谱技术在食用油种类鉴别及脂肪酸含量检测中的应用[J]. 北京工商大学学报:自然科学版,2010(5):56-59.

Study on Identifying Peanut Oil Adulteration Based on Near Infrared Spectroscopy

WU Jing-zhu, LIU Cui-ling, LI Hui, SUN Xiao-rong

(School of Computer Science and Information Engineering, Beijing Technology and Business University, Beijing 100048, China)

Abstract: A new method based on near infrared spectroscopy technology was established to detect peanut oil adulteration quickly. Four kinds of adulteration samples of soybean oil, rapeseed oil, palm oil and blend oil and five samples of pure peanut oil were collected for detection. Near Infrared spectrums of the samples were normalized as the input variables to SVM. The kernel function of SVM was RBF. The best combination of penalty parameter c and RBF kernel parameter γ were selected by grid search and K-fold cross validation ($c = 1\ 024, \gamma = 1$). The recognition ratio and prediction ratio were 100%. The experiments showed that the fast detection technology based on NIR and SVM was practicable for identifying peanut oil adulteration.

Key words: peanut oil; adulteration identification; NIR; SVM

(责任编辑:王 宽)