DOI:10.13652/j.issn.1003-5788.2016.01.009

近红外光谱技术结合支持向量机对食用醋品牌溯源的研究

Research on vinegar brand traceability based on near infrared spectrum technology combined with support vector machine

刘 静[1] 管 骁[2] 易翠平[3]

LIU Jing[1] *GUAN Xiao*[2] *YI Cui-ping*[3]

(1. 上海海事大学信息工程学院,上海 201306;2. 上海理工大学医疗器械与食品学院,上海 200093;
3. 长沙理工大学化学与生物工程学院,湖南 长沙 410114)

(1. *College of Information Engineering, Shanghai Maritime University, Shanghai* 201306, *China*;
2. *School of Medical Instrument and Food Engineering, University of Shanghai for Science and Technology, Shanghai* 200093, *China*; 3. *School of Chemistry and Biology Engineering, Changsha University of Science and Technology, Changsha, Hunan* 410114, *China*)

摘要:研究近红外光谱技术对食用醋品牌进行快速无损溯源。收集市场上保宁、东湖、恒顺、镇江 4 个品种共 152 份具有代表性的食用醋样品,采集它们的近红外光谱数据,对原始光谱数据进行多元散射校正(multip,licative scatter correrrtion,MSC)预处理,对预处理后的光谱数据利用主成分分析法(principal component analysis,PCA)进行聚类分析,根据主成分的累计贡献率选取主成分数,然后用支持向量机(support vector machine,SVM)建立预测模型,选取合适的 SVM 核函数,并利用粒子群优化算法(particle swarm optimization,PSO)优化模型参数。结果表明,近红外光谱技术结合支持向量机对食用醋品牌分类正确率可达 100%。

关键词:近红外光谱;主成分分析法;支持向量机;粒子群优化算法;品牌溯源

Abstract: Presented a fast and non-destructive method for the discrimination of vinegar brands by near-infrared spectroscopy technology. One hundred and fifty-two representative samples of vinegar including Bao Ning, East Lake, Heng Shun, Zhenjiang were collected from market. Multiplicative Scatter Correction (MSC) was used to handle the original near infrared spectrum (NIR) data and Principal Component Analysis (PCA) was used to process the spectral data after pretreatment according to the accumulative contribution rate of principal components to select principal components. Support Vector Machine (SVM) was then applied to build the brand traceability model with proper kernel function. Particle Swarm Optimization was applied to optimize the parameters of the model. The experiments indicated that the method combing near infrared spectroscopy with support vector machine could classify the vinegar brand with 100% accuracy.

Keywords: near-infrared spectroscopy; principal component analysis; support vector machine; particle swarm optimization; brand traceability

根据原料和工艺的不同,不同品牌的食用醋在营养价值、价格等方面有较大差别。目前中国食用醋品牌很多,其中不乏假冒伪劣产品,而通过常规的味道、色泽等方面的分析很难分辨出来[1]。

近红外光谱分析技术是一种快速无损的检测分析技术,具有分析效率高、成本低、速度快、重现性好等优点[2],已经被越来越多地应用于环境科学[3]、制药与临床医学[4]、生物化工[5]、高分子化工[6]等领域。传统的食用醋检测基本上采用感官评定法、卫生检验法和理化分析法,这些方法需要专业人员进行试验操作,且存在主观性强和检测时间长等缺点。已有研究表明,近红外光谱结合适当的化学计量学方法,可以成功应用于食用醋品牌分类和质量的定性和定量分析,如基于近红外光谱技术对果醋有机酸和 pH 值的检测[7]、基于近红外光谱技术的瓶装醋定性检测[8]等。这些研究多利用偏最小二乘或谱区筛选方法来建立校正模型,虽也获得了较好的结果,但仍有提升的空间。支持向量机(support vector machine,SVM)是近年来发展出的一种新的机器学习技术,通常在处理小样本、非线性、模糊性等特征的数据

基金项目:上海市自然科学基金(编号:14ZR1419200)
作者简介:刘静(1979—),女,上海海事大学副教授,博士。
E-mail:jingliu@shmtu.edu.cn
收稿日期:2015-09-18

时表现出很强的高维辨识能力[9-11]。本研究拟将近红外光谱技术与支持向量机结合，对市场上 4 种知名的食用醋品牌保宁醋、东湖醋、恒顺醋、镇江醋采集近红外光谱图数据，运用多元散射校正方法进行数据预处理后，利用支持向量机建立具有高稳定性和高精度的近红外光谱数据的品牌溯源模型，旨为后续研究提供参考。

1 材料与方法

1.1 材料与仪器

试验样品：市售的 4 种品牌（保宁醋、东湖醋、恒顺醋、镇江醋）共 152 组不同的食用醋样品，其中保宁醋 24 组，东湖醋、镇江醋各 40 组，恒顺醋 48 组，所有食用醋的生产日期均在 2012 年 1～12 月；

多通道傅里叶变换近红外光谱仪：MPA 型，德国 BRUKER 公司。

1.2 方法

1.2.1 光谱采集 采用 MPA 型多通道傅里叶变换近红外光谱仪。光谱采集为漫反射模式分辨率 8 cm^{-1}，扫描次数 32，每个样品测量 5 次，取平均值作为样品典型光谱。光谱采集过程中保持环境温度 25 ℃，湿度 72.5%。

1.2.2 数据预处理 多元散射校正（MSC）是由 Martens 等[9]提出的一种用于消除由于样品颗粒大小不同和分布不均匀产生的散射对其光谱的影响的预处理方法。多元散射校正（MSC）的算法如下：

（1）计算所有样品光谱的平均值：

$$\overline{\boldsymbol{A}_j} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{A}_{i,j} \tag{1}$$

式中：

$\boldsymbol{A}$——$n\times p$ 维定标光谱数据矩阵（n 为样品数，p 为光谱采集所用波长点数）；

$\bar{\boldsymbol{A}}$——所有样品的原始光谱在各个波长点处求平均值得到的平均光谱矢量。

（2）一元线性回归：

$$\boldsymbol{A}_i = m_i\bar{A} + b_i \tag{2}$$

式中：

$\boldsymbol{A}_i$——$1\times p$ 维矩阵，表示单个样品光谱矢量；

m_i——各样品近红外光谱 A 与平均光谱 $\bar{A}$ 进行一元线性回归后得到的相对偏移系数；

b_i——各样品近红外光谱 A 与平均光谱 $\bar{A}$ 进行一元线性回归后得到的平移量。

（3）计算校正后的光谱：

$$A_{i(MSC)} = \frac{(A_i - b_i)}{m_i} \tag{3}$$

其中 $i = 1,2,\cdots,n$，n 为样品数；j 表示第 j 个波数。

1.2.3 建模方法

（1）主成分分析法（PCA）：PCA 是多元统计中的一种利用降维的思想，将原来多个变量转化为数量较少的几个综合变量的数学变换方法，这些综合变量保留了原来多个变量中绝大部分信息，降低了数据处理的复杂程度，并且最大化地反映出原来多个变量所包涵的内容，减小了误差因素的干扰[10]。文中运用主成分分析法，在不丢失光谱主要信息的前提下，选择几个能代替原始光谱数据的特征变量，解决因谱带重叠带来分析困难的问题。

（2）支持向量机（SVM）：SVM 最初于 20 世纪 90 年代由 Vapnik[11]提出的一种结构风险最小化原理的机器学习方法。SVM 包括非线性支持向量机和线性支持向量机两类，当样品线性不可分时，通过非线性映射 ϕ，将原样本映射到高维特征空间，即定义映射 $\phi: x\in R^m \to \phi(x)\in R^p\ (p>m)$；当样品线性可分时，通过构造线性最优分类超平面 $w*x+b=0$，将两类样本完全精确地分开。选择不同的核函数，可以得到不同的 SVM，常用的核函数有以下 4 种：① 线性核函数：$K(X,X_i) = X\bullet X_i$；② 多项式核函数：$K(X,X_i) = ((X\bullet X_i)+c)^d$；③ RBF 核函数：$K(X,X_i) = \exp(-\sigma||X-X_i||^2)$；④ sigmoid 核函数：$K(X,X_i) = \tanh(\kappa*(X\bullet X_i)+v)$。

（3）粒子群优化算法（PSO）：PSO 是由 Eberhart 等[12]提出的一种通过群体中个体间信息共享和信息传递来寻找最优解的进化算法技术。粒子群优化算法（PSO）的算法如下：

$$V_i = \omega V_i + c_1\times rand()\times(pbest_i - X_i) + c_2\times rand()\times(gbest_i - X_i) \tag{4}$$

$$X_i = X_i + V_i \tag{5}$$

式中：

ω——值为非负的惯性权重因子，其值的大小对整体的寻优能力有影响；

$i = 1,2,\cdots,M$，M 是该群体中粒子的总数；

$pbest_i$——当前粒子的最优位置；

$gbest_i$——当前群体的最优位置；

V_i——粒子的速度；

c_1，c_2——学习因子，一般取值为 2；

$rand()$——(0,1)之间的随机数；

X_i——粒子的当前粒子位置。

依据目标函数来计算适合每个粒子的应值大小，目标函数有多种选择，一般为方差、标准差、均方误差。根据适应值来确定当前群体最优位置 $gbest_i$ 和当前粒子最优位置 $pbest_i$，接着根据式(4)和式(5)调整粒子位置以及速度，最终搜索到粒子的最优解。

2 食用醋品牌溯源模型

2.1 不同品牌的食用醋的近红外光谱图

4 种品牌食用醋的近红外光谱图见图 1。由图 1 可知，所有样品在 6 740～6 810 cm^{-1} 附近有明显的吸收峰，该区域是 N—H 键伸缩振动的倍频区及伸缩振动和弯曲振动的组合频区，这与食用醋中含有的氨基酸的事实相吻合。同时观察到在 5 538～5 640 cm^{-1} 和 8 330～8 600 cm^{-1} 谱带区域也出现了较小程度的吸收峰，分别为 C—H 键振动的组合频、一级倍频、二级倍频区域，这与食用醋中含有酸类、酯类等有机物有关。从图 1 中可以看出，不同食用醋的原始近红外光谱曲线

重叠交错,无法直接由图谱直观地看出品牌间的差别。故需要结合适当的化学计量学方法对原始光谱数据进行处理。

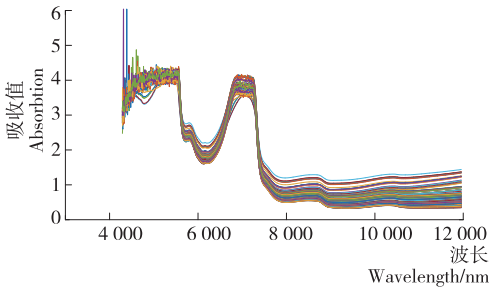


图 1　醋的原始光谱图

Figure 1　Original spectral data of vinegar

图 2 为经过多元散射校正(MSC)处理后的光谱数据,每个样品的原始光谱与平移量相减再除以倾斜偏移量以校准光谱的基线,因此样品的基线偏移和平移都得以修正,而和样品成分含量等相对应的光谱吸收的信息在修正过程中并没有产生任何影响,从而提高了光谱的信噪比。该方法校正了原始光谱数据中由噪音、样品颗粒大小、以及光散射等影响所带来的差异。

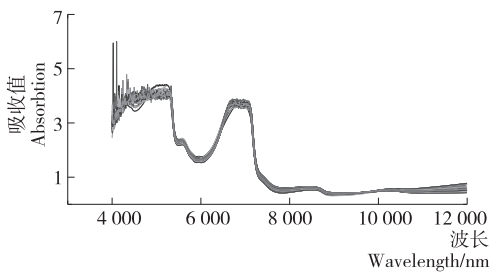


图 2　多元散射校正处理后的光谱图

Figure 2　Spectral after MSC pretreatment

2.2　聚类分析

由图 2 可知,光谱曲线的首端有较大的波动噪音,为消除这些噪音对试验的影响,选取 5 500～11 000 nm 波段的光谱数据进行研究。将经过多元散射校正(MSC)处理后的光谱数据进行主成分分析,表 1 为前 8 个主成分的贡献率,前 4 个主成分的贡献率已经达到 98.14%,之后贡献率的变化较小,说明原有的样品信息可以由前 4 个主成分代替,接下来选用前 4 个主成分进行试验。

2.3　SVM 模型的建立与优化

利用 SVM 工具箱建立 SVM 模型,对训练集样品和测试集样品进行预测。在模型建立过程中使用不同的核函数建立模型,比较不同核函数对模型建立的影响,结果见表 2。

试验结果表明选取径向基函数(radial basis function,RBF)作为核函数时,SVM 模型可得到较高的品牌分类准确率。为进一步提高分类准确率,将利用 PSO 算法对 SVM 模型的参数(惩罚系数 c 和核函数参数 g)进行优化(见图 3),粒子群在迭代次数 100 以内就可快速搜索到模型参数的最优解,其中最优参数 c 和 g 分别为 0.665 22 和 10.475 1,优化后的 SVM 模型的分类准确率达到 100%。根据得到的最优参数 c 和 g,建立 SVM 模型,对外部测试集进行分类预测(见表 3),利用 PSO 算法优化模型参数 c 和 g,结合最适合醋类数据的径向基核函数建立 SVM 模型,对食用醋外部测试样本的鉴别率可达 100%。

表 1　各主成分的贡献率

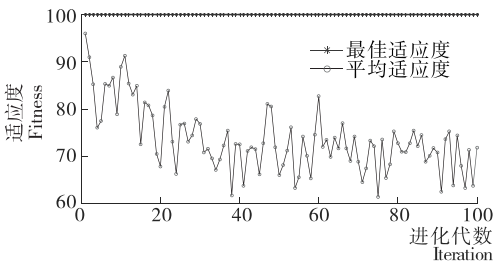
Table 1　The cumulative contribution of principal components

主成分	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
贡献率/%	74.35	87.31	94.60	98.14	98.35	98.49	98.60	98.70

表 2　不同核函数的比较

Table 2　The comparison of different kernel functions

采用的核函数	准确率/%
Linear	71.05
Polynomial	44.74
Radial basis function	84.21
Sigmoid	50.00



参数c_1=1.5,c_2=1.7,终止代数=100,种群数量pop=20
(Best c=1.273 3 g=5.453 bestCVAccuracy=100%)

图 3　粒子群优化算法算法的适应度曲线

Figure 3　Fitness curve of PSO algorithm

表 3　模型的预测结果

Table 3　The prediction results of model

样本种类	外部测试集	误判个数	准确率/%	总准确率/%
保宁	13	0	100	100
东湖	4	0	100	
恒顺	10	0	100	
镇江	11	0	100	

3　结论

近红外光谱技术的关键是从大量重叠的光谱数据中提取有用的信息,建立测试效果优秀的模型。本研究应用多元散射校正(MSC)对原始光谱数据进行预处理,结合主成分分析法(PCA)和支持向量机(SVM)建立测试模型。通过比较不同核函数对 SVM 模型的影响,得出食用醋品牌溯源的 SVM 模型的最佳核函数为 RBF 核函数,并且利用 PSO 算法可以快速寻得最优的参数 c 和 g 进一步提高分类准确率。由此表明,本研究提出的近红外光谱技术结合粒子群优化算法和支持向量机技术,可以高效地达到食用醋品牌溯源的目的。近红外光谱技术以大量具有代表性的样品为研究对象,

(下转第 50 页)

3　结论

本试验研究了酸性饮料三片罐中三聚氰胺的迁移规律,并在此基础上建立了模拟物体系中三聚氰胺的迁移数学模型,进行了效果验证。迁移试验结果表明,酸性饮料三片罐中三聚氰胺的迁移速度随温度的升高而加速,温度越高,迁移至平衡所需的时间越短,因而可通过高温加速试验来预测三聚氰胺最终迁移水平。所建模型对应三聚氰胺的迁移预测值与试验值基本相符,因此说明建立的模型能够对三片罐中三聚氰胺的迁移行为进行较好地预测,为评估三片罐的安全性提供保障。

参考文献

[1] 尹中,李月樵,韦何雯. 反相高效液相色谱法测定金属灌装饮料中三聚氰胺单体迁移量[J]. 中国卫生检验杂志,2011(8):2 085-2 086.

[2] 徐彦辉,陈戈,顾亮,等. 金属罐内层涂料三聚氰胺迁移量的测定[J]. 包装与食品机械,2011(3):66-68.

[3] 王红松,陈烨,刘君峰,等. 金属食品包装罐中三聚氰胺迁移规律的研究[J]. 检验检疫学刊,2014(2):48-51.

[4] 韩冰冰,宋文生,李雪娟. 三聚氰胺及其衍生物的应用[J]. 化学推进剂与高分子材料,2007(6):26-30.

[5] 凌光耀. 食品用金属包装内涂卫生安全标准探讨[J]. 中国包装,2011(6):60-62.

[6] 郝倩,苏荣欣,齐崴,等. 食品包装材料中有害物质迁移行为的研究进展[J]. 食品科学,2014(21):279-286.

[7] 鲁杰,杨大进,宋书锋,等. 食品用蜜胺制品中三聚氰胺迁移量迁移规律的研究[J]. 卫生研究,2013(3):483-485,490.

[8] 商贵芹,王红松,寇海娟,等. 蜜胺仿瓷餐具甲醛和三聚氰胺在食品模拟物中迁移规律的研究[J]. 中国卫生检验杂志,2012(6):1 221-1 224,1 228.

[9] 汪辉,曹小彦,彭新凯,等. 高效液相色谱—二极管阵列法测定高蛋白食品中的三聚氰胺[J]. 食品与机械,2007,25(5):114-115,124.

[10] 吴富忠. 高效液相色谱法测定密胺餐具中三聚氰胺单体迁移量[J]. 浙江预防医学,2015(2):211-213.

[11] 林晓珊,吴惠勤,黄晓兰,等. 气相色谱—串联质谱法快速测定乳制品中三聚氰胺及其 3 种类似物[J]. 质谱学报,2014(6):537-543.

[12] Wei-Chih Cheng, Shu-Kong Chen, Tien-Jen Lin, et al. Determination of urine melamine by validated isotopic ultra-performance liquid chromatography/tandem mass spectrometry[J]. Rapid Communications in Mass Spectrometry, 2009, 23(12): 1 776-1 782.

[13] 勇艳华,顾鑫荣,袁斌,等. 液相色谱串联质谱法测定蛋及蛋制品中三聚氰胺残留[J]. 粮油加工,2010(2):86-88.

[14] 皮林格,巴纳. 食品用塑料包装材料—阻隔功能、传质、品质保证和立法[M]. 范家起,张玉霞,译. 北京:化学工业出版社,2004.

[15] Pennarun P Y, Dole P, Feigenbaun A. Overestimated diffusion coefficients for the prediction of worst case migration from PET: Application to recycied PET and to functional barriers assessment[J]. Packaging Technoiogy and Science, 2004(17): 307-320.

[16] 李丹,李忠海,袁列江,等. 纸塑包装中 PCBs 的迁移模型及效果评价[J]. 食品与机械,2012,28(3):162-166.

(上接第 40 页)

样品的收集和数据积累是提高模型稳健性与预测能力的基础,在今后的研究中应该不断增加品牌醋的种类、样品组的数量从而使模型数据库更加充实;另外,本课题中所用样品是液体,具有挥发性,贮藏环境、测量条件以及机器差异都会影响光谱的采集,一台机器上构建的模型可能无法适用于其他机器,因此数据的传递性、模型的普适性将是近红外光谱技术发展中亟需解决的问题。

参考文献

[1] 王莉,刘飞,何勇. 应用可见近红外光谱技术进行食用醋品牌和 PH 值的快速检测[J]. 光谱学与光谱分析,2008,28(4):813-816.

[2] 徐广通,袁洪福,陆婉珍. 现代近红外光谱技术及应用进展[J]. 光谱学与光谱分析,2000,20(2):134-142.

[3] Ken Watanabe, Shawn D Mansfield, Stavros Avramidis. Application of near-infrared spectroscopy for moisture-based sorting of green hem-fir timber[J]. Journal of Wood Science, 2011, 57(4): 288-294.

[4] Justin J Skowno, Jonathan Saul Karpelowsky. Near-infrared spectroscopy for monitoring renal transplant perfusion[J]. Pediatric Nephrology, 2014, 29(11): 2 241-2 242.

[5] Ana Henriques, Paulo Cruz, Jorge Martins, et al. Determination of melamine content in amino resins by near-infrared spectroscopy[J]. Wood Science and Technology, 2013, 47(5): 939-940.

[6] Gastaldi D, Canonico F, Irico S, et al. Near-infrared spectroscopy investigation on the hydration degree of a cement paste[J]. Journal of Materials Science, 2010, 45(12): 3 169-3 174

[7] Liu Fei, He Yong, Wang Li. Detection of organic acids and pH of fruit vinegars using near-infrared spectroscopy and multivariate calibration[J]. Food and Bioprocess Technology, 2011, 4(8): 1 311-1 340.

[8] 宋海燕,秦刚,刘海芹. 基于近红外光谱技术的瓶装醋定性检测[J]. 光谱学与光谱分析,2012,32(6):1 547-1 549.

[9] 赵强,张工力,陈星旦. 多元散射校正对近红外光谱分析定标模型的影响[J]. 光学精密工程,2005,13(1):53-58.

[10] 李武,胡冰,王明伟. 基于主成分分析和支持向量机的太赫兹光谱冰片鉴别[J]. 光谱学与光谱分析,2014,34(12):3 235-3 240.

[11] Huang Hong-zhong, Wang Hai-kun, Li Yan-feng. Support vector machine based estimation of remaining useful life: current research status and future trends[J]. Journal of Mechanical Science and Technology, 2015, 29(1): 151-153.

[12] 李玉军,汤晓君,刘君华. 粒子群优化算法在混合气体红外光谱定量分析中的应用[J]. 光谱学与光谱分析,2009,29(5):1 276-1 277.