

# 基于 NIRS 的食用醋品牌溯源模型的建立与优化

管晓<sup>1</sup>, 古方青<sup>1</sup>, 刘静<sup>2</sup>, 杨永健<sup>3</sup>

(1. 上海理工大学医疗器械与食品学院, 上海 200093) (2. 上海海事大学信息工程学院, 上海 201306)

(3. 上海市食品药品检验所, 上海 201203)

**摘要:** 本文主要探讨了近红外光谱(NIRS)结合模式识别技术应用于食用醋品牌溯源研究。采集了四个品牌(四川保宁香醋、山西东湖老陈醋、镇江恒顺香醋、镇江香醋)共160组食醋样品的近红外漫反射光谱,通过主成分分析(PCA)进行光谱变量压缩及剔除8个异常样本数据后,随机选取其中的114组样品组成训练集用于建立溯源模型,剩余38组样品用作测试集进行模型验证。比较了MSC、SD、SNV等几种不同光谱预处理方法以及它们的不同组合对溯源模型的影响,同时考察了PLS-DA与SIMCA两种建模方法对模型的影响。结果表明:选择MSC与SD相结合的方法对光谱数据进行预处理,并采用SIMCA建模方法所建立的醋品牌溯源模型对四大品牌醋的正确识别率分别可达100%、100%、91.7%、90%。由此说明采用近红外光谱技术结合模式识别技术可有效实现食用醋品牌溯源的目的。

**关键词:** 食醋; 品牌溯源; 模型; 近红外光谱

文章编号: 1673-9078(2014)11-200-203

DOI: 10.13982/j.mfst.1673-9078.2014.11.035

## Development and Optimization of the Brand Traceability Model of Vinegar Based on Near-infrared Spectroscopy

GUAN Xiao<sup>1</sup>, GU Fang-qing<sup>1</sup>, LIU Jing<sup>2</sup>, YANG Yong-jian<sup>3</sup>

(1. School of Medical Instruments and Food Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China) (2. College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China)

(3. Shanghai Institute for Food and Drug Control, Shanghai 201203, China)

**Abstract:** In this paper, near-infrared spectroscopy (NIRS) and pattern recognition technology were used to establish a brand traceability model for vinegar. NIR diffuse reflectance spectra of 160 vinegar samples from four different brands (SICHUAN BAONING vinegar, SHANXI EASTLAKE vinegar, HENGSHUN CHINKIANG vinegar, and CHINKIANG vinegar), were collected. After the spectral variables were compressed and eight abnormal sample data were removed using principal component analysis (PCA), 114 samples were randomly selected to form a training set for the construction of a traceability model, and the remaining 38 samples were used to form a test set for model validation. The effects of different pretreatment methods of the NIR spectra including multiplication signal correction (MSC), second derivative (SD), standard normal variate (SNV) etc., and different combinations of these methods on the traceability model were compared, and the effects of two modeling methods, partial least squares-discriminant analysis (PLS-DA) and soft independent modeling by class analogy (SIMCA) were examined. The results demonstrated that the correct recognition rates of the four different brands of vinegars could reach 100%, 100%, 91.7%, and 90%, respectively, when the spectral data were pretreated by MSC combined with SD and the model was developed using SIMCA. Therefore, the brand traceability of vinegar could be effectively achieved using NIR and a pattern recognition technique.

**Key words:** vinegar; brand traceability; model; near infrared spectrum

食用醋酸爽醇厚,除了赋予食品特有的风味外,自身还含有丰富的营养成分,起到食疗保健和药用功能。固态分层发酵是醋酸发酵过程的关键工艺,不同品牌的食用醋因为使用的微生物发酵菌种、发酵原料

收稿日期: 2014-04-29

基金项目: 国家自然科学基金项目(31101348)

作者简介: 管晓(1979-),男,博士,副教授,主要从事食品质量控制方面的研究

等的差异,赋予了它们不同的风味特征和营养价值,但仍无法直接通过感官加以区别。正因如此,导致部分不法商贩受利益驱使会以次充好、制假勾兑,严重侵害了品牌醋的名誉与利益,也极大损害了消费者的权益,因此有必要开发出一种品牌醋溯源技术来鉴别产品真伪,以维护正常的市场秩序。

近年来,近红外光谱(Near infrared spectroscopy, NIRS)结合化学计量学技术在食品定性与定量分析中

已逐步得到应用, 不仅因为近红外光谱技术无需预处理、快速灵敏等优点, 还因为该技术成本低、适于现场操作; 同时模式识别作为化学计量学的一个重要分支发展迅速, 可有效应用于食品溯源研究。常用的模式识别主要包括纯模式识别和类建模技术。纯模式识别是通过分割有许多组别的多维空间来进行分类, 采用纯模式识别判别的每个样品总是归属于一个且只有一个群组, 这些方法主要包括线性判别分析 (Linear Discriminant Analysis, LDA)、偏最小二乘判别分析 (Partial Least-Squares Discriminant Analysis, PLS-DA)、支持向量机 (Support Vector Machines, SVM) 等<sup>[1]</sup>, 尤以 PLS-DA 方法应用最为普遍<sup>[2,3]</sup>; 而类建模技术则侧重于类元素的类比建模, 在类建模方法中每个独立的类都分别建模, 模型中满足规则的对象被认为是一个类的成员, 而不满足的对象作为非成员而被拒绝<sup>[4]</sup>。当对一个以上的类进行建模时, 可能会遇到三种不同情况: 每个样品被分配到单一类别中、代表好几种类别以及不属于任何类别。与纯模式识别不同的是, 类建模技术原则上可以找出不代表任何检测类别的样本, 此外类建模技术中由于每个类别单独建模, 添加任何额外的类别可以不重新计算已经存在的类模型, 大大降低了工作量<sup>[5]</sup>, 最常用的类建模技术是簇独立软模式分类法 (Soft Independent Modeling of Class Analogy, SIMCA)。

基于此, 本研究以四种不同品牌食用醋共 160 组样品为研究对象, 利用近红外光谱结合模式识别对醋品牌进行溯源, 着重研究了不同光谱数据变量预处理方法, 以及不同模式识别对建模结果的影响, 以达到优化食用醋品牌溯源模型的目的。

## 1 材料与方法

### 1.1 食醋样品的准备

本研究中搜集的 160 组食醋样品均购自于当地大型超市, 根据产地和品牌差异将食醋分为四大类: 四川保宁香醋 26 组, 山西东湖老陈醋 43 组、镇江恒顺香醋 51 组、镇江香醋 40 组。每类食醋根据陈酿年份的不同进一步细分为 3-6 种。

### 1.2 光谱获取

所有食醋样品近红外光谱均由 MPA 型多通道傅里叶变换光谱仪进行测定, 并通过 OPUS 6.5 软件进行处理。光谱采集模式为漫反射, 测定光谱范围为 4000~10000  $\text{cm}^{-1}$ , 分辨率 4  $\text{cm}^{-1}$ , 扫描次数 32 次。每组样品均分别测定 5 次, 取平均值作为样品典型光谱。

## 2 光谱数据分析处理

### 2.1 光谱预处理

漫反射模式下采集近红外光谱信号时, 光谱信息包含了样品所含的物质结构以及成分含量信息, 但也会因外界噪声干扰、仪器响应、杂散光等因素导致噪音信号, 从而影响后续的光谱分析。因此有必要结合化学计量学方法对原始光谱信息进行预处理。本研究分别采用多元散射校正 (Multiplicative Signal Correction, MSC)、二阶导数 (Second Derivative, SD)、标准归一化 (Standard Normal Variate, SNV) 等方法以及它们的组合对光谱数据进行预处理。MSC 是运用数学函数进行光谱色散与反射校正, 从而实现缩减甚至抵消光谱偏移目的; SD 通常可以消除基线漂移、克服谱带重叠现象, 由此达到强化谱带特征的作用; SNV 则是一个加权归一化方法, 常用于减小近红外光谱基线信号和散射带来的偏差。

### 2.2 建模方法

化学计量学建模在 UNSCRAMBLER 9.7 (CAMO 公司) 软件环境下进行。该软件不仅能实现光谱数据的预处理, 还能提供多种建模方法, 如主成分分析 (Principal Component Analysis, PCA)、SIMCA、PLS-DA。

#### 2.2.1 无监督模式识别

PCA 是一种无监督的模式识别, 作为数据压缩和信息提取工具应用十分广泛, 常用于确定描述数据主要变化趋势的综合变量。PCA 通过投影得到一群分布在空间中的点来模拟数据集中最大变异方向, 这一系列数据点称为主成分 (Principal Components, PCs)。PCs 以递减的顺序描述对象中较高变异, 又因其在计算过程中彼此正交, 每一个 PC 都可以独立解释变异, 这有助于通过揭示对象之间的关系去描述数据结构以及检测对象的偏差情况。为确定变异来源, 原始数据矩阵可分解成对象空间、变量空间及误差矩阵, 误差矩阵代表不能被所提取的 PCs 解释的变量。

#### 2.2.2 有监督模式识别

SIMCA 是类建模技术中最常用的方法之一, 它在 PCA 基础上实现了对训练集中每个类分别建模的目的。建模时每个类单独建立 PCA 模型, 若未知样品与训练集中某一类足够相似, 即可将该样品归到该类。若各类之间的距离足够大, SIMCA 分类效果越好。

PLS-DA 作为另一种模式识别方法, 建模时首先赋值一组虚拟变量作为参考值以描述样品单元中组类

别, 然后计算相关独立变量和一类中指定样品虚拟变量之间的关系。训练时根据 PLS 计算指定样品的虚拟值, 若值大于一个特定预测阈值时, 样品属于该组, 常用阈值定义为 0.5<sup>[10]</sup>。

### 3 结果分析

#### 3.1 食用醋样品近红外光谱

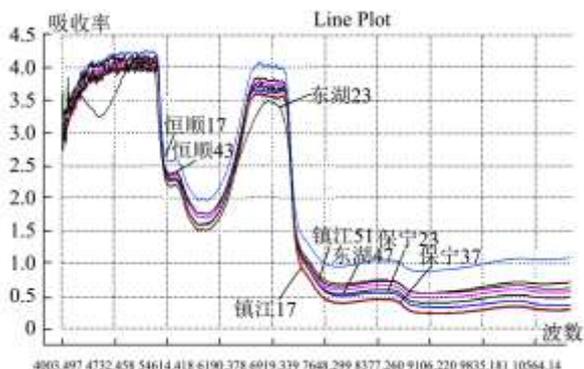


图 1 不同品牌食用醋代表性近红外光谱图

Fig.1 Representative NIR spectra of different brands of vinegar

图 1 所示为四种品牌食用醋代表性样品近红外光谱图。近红外光谱反应的是测试样品中含氢基团的振动倍频与组合频信息, 样品中所含成分、浓度等都会影响光谱信号。本研究每隔 4 cm<sup>-1</sup> 波数进行光谱测量, 在 4000~10000 cm<sup>-1</sup> 波数范围内共获得 2074 个波数-吸收率数据点, 组成光谱数据变量集。各品牌食用醋光谱图均存在一些共性特征: 在 6900 cm<sup>-1</sup> 附近出现的大的吸收峰, 为食醋中醇类和酚类的 O-H 键伸缩振动所引起; 在 5500 cm<sup>-1</sup>、8600 cm<sup>-1</sup> 附近出现的小的吸收, 代表了 C-H 键振动的倍频区域, 这与食用醋中含有的

乙酸等有机酸有关。由于各品牌食用醋主要成分相似, 峰形相近, 并存在严重重叠现象, 无法通过直接观察方法将不同品牌醋区分开, 因此需要结合化学计量学手段将光谱数据集进行处理并建立分类模型, 才能达到品牌分类的目的。

#### 3.2 近红外光谱主成分分析

图 2 所示为部分代表性食用醋近红外漫反射全谱数据主成分得分图。由图可知, PC1、PC2 及 PC3 方差贡献率分别为 93%、4%和 1% (PC2 在纵轴, 未标出), 此时前三个主成分累积方差贡献率已达到 98%, 但仍然有部分样品存在异常, 所以建模前需剔除异常样本以减小模型误差。本实验中提出的异常样本共 8 组, 为四川保宁香醋 2 组, 山西东湖老陈醋 3 组, 镇江恒顺香醋 3 组。

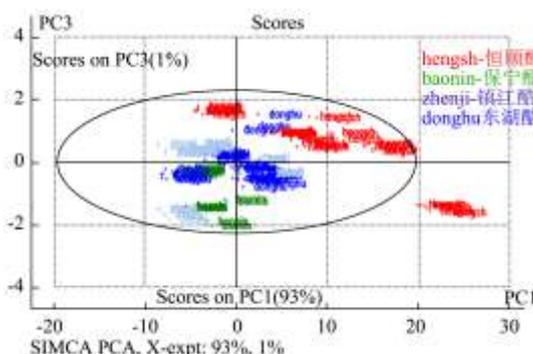


图 2 代表性食用醋近红外全谱数据变量主成分得分图

Fig.2 PCA scores of NIR spectra for the representative vinegar samples

#### 3.3 质量参数预测

表 1 不同预处理方法对模型参数的影响

Table 1 Effects of different pretreatment methods on model parameters

模型参数	预处理方法							
	-	MSC	SD	SNV	MSC+SD	SD+SNV	MSC+SNV	MSC+SD+SNV
R <sup>2</sup>	0.71	0.87	0.83	0.81	0.92	0.86	0.89	0.76
RMSEC	0.137	0.013	0.019	0.021	0.012	0.017	0.014	0.034
RMSEP	0.119	0.011	0.023	0.015	0.014	0.015	0.016	0.021

注: - : 未进行数据预处理。

160 个食用醋样品剔除 8 个异常值之后, 选取 18 组保宁醋、30 组东湖醋、36 组恒顺醋、30 组镇江醋组成训练集用于建模, 剩余 38 组样品组成预测集用于评价模型效果。建模前对光谱数据集采用不同的预处理方法进行预处理 (如表 1 所示), 并考察预处理方法的选择对模型预测精度的影响。为评估预测模型的可靠性, 选择了三个统计学参数, 即模型相关系数 (Determination Coefficient, R<sup>2</sup>)、校准均方根误差

(Root-Mean-Square Error of Calibration, RMSEC)、预测均方根误差 (Root-Mean-Square Error of Prediction, RMSEP)。表 1 所示为选择不同预处理方法进行数据处理并回归分析建模后模型参数的变化情况。

由表 1 结果可知, 原始光谱数据未经预处理直接回归建模后模型的 R<sup>2</sup>、RMSEC、RMSEP 分别为 0.71、0.137、0.119, 而经过预处理之后, 模型效果均有不同

程度的改善,由此可见预处理过程可提高数据变量相关性,进而提高预测模型的可靠性。同时也发现,先后采用 MSC 和 SD 两种预处理方法时,模型相关系数最高,达到 0.92,但先后采用 MSC、SD 和 SNV 三种预处理方法组合使用时,模型相关系数反而降低(0.76),这可能是由于预处理造成的数据过拟合现象,反而使得光谱变量的线性组合相关性降低。因此本实验选择 MSC+SD 的数据预处理方法,以便于下一步建模。

### 3.4 不同建模方法的效果比较

表 2 两种建模方法对溯源结果的影响比较

Table 2 Effects of two modeling methods on the traceability results

	SIMCA		PLS-DA	
	正确分	正确分	正确分	正确分
	类组数	类率/%	类组数	类率/%
保宁醋(6组)	6	100	5	83.3
东湖醋(10组)	10	100	8	80
恒顺醋(12组)	11	91.7	10	83.3
镇江醋(10组)	9	90	9	90
总计(38组)	36	94.7	32	84.2

不同建模方法的选择同样对溯源模型的结果有重要影响。实验中选取了 PLS-DA 和 SIMCA 方法分别建立食用醋品牌溯源模型,并利用模型对测试集中的 38 组样本进行分类预测,结果如表 2 所示。由此可知,PLS-DA 和 SIMCA 方法建立的模型对测试集均有较好的预测效果,总分类正确率均在 84% 以上。特别是 SIMCA 模型分类正确率达到了 94.7%,尤其对 6 组保宁醋和 10 组东湖醋都达到了 100% 的正确分类率,对恒顺醋和镇江醋的分类正确率也超过了 90%,均优于 PLS-DA 模型。这可能是由于 SIMCA 建模时需分别对四种品牌醋建立主成分模型,每个主成分模型都是独立的类模块,模块间数据变量之间无关联;而 PLS-DA 是以整个训练集为建模对象,在数据变量间建立预测模型,虽然每个品牌醋都设定了不同虚拟变量值,但由于模型中数据变量的关联性使模型对未知试样的预测精度较低。因此在利用 MSC+SD 进行数据预处理基础上,选择 SIMCA 方法建模是一种更为准确有效的食用醋品牌溯源的方法。

## 4 结论

本文主要讨论了不同光谱数据预处理及模式识别方法对建立食用醋近红外品牌溯源模型的影响,并确定了溯源模型的建立方法。结果表明采用 MSC 与 SD

组合的光谱预处理方法,并结合 SIMCA 建模技术,所得溯源模型对保宁醋、东湖醋、恒顺醋和镇江醋四种品牌醋的预测结果最优,分别达到 100%、100%、91.7%、90%,达到了理想的品牌溯源目的。下一步的研究仍有必要进一步加大建模样品数量,使建立的模型更具有代表意义与普适性。

## 参考文献

- [1] Cozzolino D, Smyth H E, Gishen M. Feasibility study on the use of visible and near-infrared spectroscopy together with chemometrics to discriminate between commercial white wines of different varietal origins [J]. *Journal of Agricultural and Food Chemistry*, 2003, 51(26): 7703-7708
- [2] Lai Y, Ni Y, Kokot S. Discrimination of rhizoma corydalis from two sources by near-infrared spectroscopy supported by the wavelet transform and least-squares support vector machine methods [J]. *Vibrational Spectroscopy*, 2011, 56(2): 154-160
- [3] Dian P H, Andueza D, JESTIN M, et al. Comparison of visible and near infrared reflectance spectroscopy to discriminate between pasture-fed and concentrate-fed lamb carcasses [J]. *Meat Science*, 2008, 80(4): 1157-1164
- [4] Pontes M J C, Santos S R B, Ara Jo M C U, et al. Classification of distilled alcoholic beverages and verification of adulteration by near infrared spectrometry [J]. *Food Research International*, 2006, 39(2): 182-189
- [5] Di Egidio V, Oliveri P, Woodcock T, et al. Confirmation of brand identity in foods by near infrared transreflectance spectroscopy using classification and class-modelling chemometric techniques-the example of a belgian beer [J]. *Food Research International*, 2011, 44(2): 544-549
- [6] Vitale R, Bevilacqua M, Bucci R, et al. A rapid and non-invasive method for authenticating the origin of pistachio samples by nir spectroscopy and chemometrics [J]. *Chemometrics and Intelligent Laboratory Systems*, 2012, 138(3): 334-336
- [7] Cozzolino D, Cynkar W U, Shah N, et al. Can spectroscopy geographically classify Sauvignon Blanc wines from Australia and New Zealand? [J]. *Food Chemistry*, 2011, 126(2): 673-678
- [8] Woodcock T, Downey G, Odonnell C. Near infrared spectral fingerprinting for confirmation of claimed PDO provenance of honey [J]. *Food Chemistry*, 2009, 114(2): 742-746
- [9] Casale M, Oliveri P, Armanino C, et al. NIR and UV-vis spectroscopy, artificial nose and tongue: comparison of four

fingerprinting techniques for the characterisation of Italian red wines [J]. *Analytica Chimica Acta*, 2010, 668(2): 143-148  
[10] Galtier O, Dupuy N, Le Dreau Y, et al. Geographic origins

and compositions of virgin olive oils determined by chemometric analysis of NIR spectra [J]. *Analytica Chimica Acta*, 2007, 595(1-2): 136-144

