

基于 SOM 和 SVM 的食醋品质近红外定性分析

Research on qualitative analysis of vinegar by using near-infrared spectroscopy combined with SOM and SVM

郝勇^{1,2} 赵翔¹ 温钦华¹ 陈斌²

HAO Yong^{1,2} ZHAO Xiang¹ WEN Qin-hua¹ CHEN Bin²

(1. 华东交通大学机电工程学院, 江西 南昌 330013; 2. 江苏大学食品与生物工程学院, 江苏 镇江 212013)

(1. College of Mechanical and Electronic Engineering, East China Jiaotong University, Nanchang, Jiangxi 330013, China; 2. School of Food and Biological Engineering, Jiangsu University, Zhenjiang, Jiangsu 212013, China)

摘要:研究近红外光谱(near-infrared spectroscopy, NIRS)结合自组织映射(self-organization mapping, SOM)和支持向量机(support vector machine, SVM)用于食醋酿造年份和品牌的判别分析。连续小波变换(continuous wavelet transform, CWT)用于光谱预处理;主成分分析(principal component analysis, PCA)用于食醋光谱降维和样品空间分布分析。结果表明:CWT 预处理可以有效消除食醋光谱的平移误差;PCA 可以极大地减少光谱变量,提高建模效率;对于食醋酿造年份的识别,采用 CWT—PCA—SOM 的正确识别率(correct recognition rate, CRR)为 97.37%,采用 CWT—PCA—SVM 的 CRR 为 100%;对于食醋品牌的鉴别,CWT—PCA—SOM 和 CWT—PCA—SVM 两种方法的 CRR 均为 100%。近红外光谱结合 CWT—PCA—SOM 和 CWT—PCA—SVM 方法在食醋酿造年份及其品牌鉴别中均得到很好的分析结果,该方法具有良好的应用前景。

关键词:自组织映射;支持向量机;食醋;近红外光谱;正确识别率

Abstract: Near-infrared spectroscopy (NIRS) combined with two discriminative analysis methods including self-organization mapping (SOM) and support vector machine (SVM) were used for discriminant analysis of vinegar with different production year and brand. Continuous wavelet transform (CWT) was adopted for spectra pre-processing. Principal component analysis (PCA) was used for spectra dimension reduction and space distribution analysis. The results shown that CWT can effective eliminate spectra translation error. PCA can greatly reduce characteristic spectrum variables and improve modeling efficiency. For identification of vinegar with different pro-

duction year, the CWT-PCA-SOM method can get 97.37% correct recognition rate (CRR), and the CWT-PCA-SVM method can get 100% CRR. For identification of vinegar brand, the CWT-PCA-SOM and CWT-PCA-SVM methods can obtain 100% CRR. Near-infrared spectroscopy combined with CWT-PCA-SOM and CWT-PCA-SVM methods can both obtain better analysis results for identification of vinegar with different production year and brand, and this method has good application prospect.

Keywords: self-organization mapping; support vector machine; vinegar; near-infrared spectroscopy (NIRS); correct recognition rate (CRR)

食醋品质主要取决于原料及发酵过程,其对食醋质量的评价起着至关重要的作用。食醋中的香气成分主要包括醇类、酯类、醛类、酸类等,它们的种类及比例变化与其构成的香气类型与食醋的品质类型有着直接关系,并且随着发酵时间的不同,这些香气成分及含量都会随之发生变化^[1-2]。

食醋的成分较复杂,对其品质的分析主要采用仪器分析方法,研究的重点内容包括食醋种类判别、掺假分析及其香气成分含量测定等^[3-4]。Boffoa 等^[5]采用¹H NMR 结合 KNN、SIMCA 和 PLS—DA 3 种方法对酿造食醋、苹果醋和酒精醋进行判别分析;Zhang Qin-yi 等^[6]采用电子鼻结合神经网络的方法对食醋、冰醋酸和 5% 的稀释冰醋酸混合样品集进行判别分析,分别按照样品类型、原材料、总酸、发酵方法和产地进行分类,分别得到了 72.1%, 76.5%, 77.9%, 94.1%, 82.4% 的正确识别率。Liu Fei 等^[7]采用近红外光谱结合最小二乘支持向量机对 4 种类型的果醋进行分类得到了较好的分析结果,识别率为 100%;Cocchi 等^[8]应用顶空固相微萃取气相色谱(HS—SPME/GC)方法结合小波包变换对香醋的真伪进行鉴别分析,WTPER 算法在色谱信号解析中得到较好的应用,该方法可以实现盲样识别,且判别模型结构简单;Hsieh 等^[9]应用 SNIF—NMR 方法对米醋发酵过

基金项目:国家自然科学基金项目(编号:21265006,31171697)

作者简介:郝勇(1978—),男,华东交通大学机电工程学院,副教授,博士后。E-mail:haonm@163.com

收稿日期:2016—02—05

程中掺入酒醪中的人造冰醋酸的含量进行定量分析,研究表明:米醋中的(D/H)_{CH₃}含量与其掺假量呈较好的线性关系($R^2 > 0.97$),据此可较好实现发酵米醋中掺假量的准确预测。在这些分析方法中,色谱法需要复杂的样品前处理,分析步骤繁琐;电子鼻分析法需要设计特异的气体传感器;多维荧光分析方法数据处理较复杂;而近红外光谱和¹H NMR分析方法由于其分析速度快、分析成本低和易于在线等分析特点具有较好的应用前景。

近红外光谱分析方法在食品和农产品的定性定量分析^[10-12]中广泛应用。然而,由于其光谱包含待测组分基因的倍频和合频吸收,需要借助化学计量学方法对光谱信息进行提取和分析。近红外光谱在食醋品质中的应用主要包括食醋品质指标的定量分析和种类的判别,Casale等^[13]采用近红外光谱结合波段筛选和线性判别方法对苹果醋、酒精醋、大麦醋等7种食醋共计95个样品的陈置效应和氧化效应进行定性判别,识别率均为100%;Liu Fei等^[14]采用近红外光谱对米醋的可溶性固形物和pH值进行定量分析,模型的相关系数均大于0.99;Saiz-Abajo等^[15]采用近红外光谱对酒精醋中的有机酸、灰分和挥发酸等质量参数进行定量分析;Chen等^[16]采用近红外光谱结合变量筛选和非线性回归方法对食醋中的总酸进行定量分析。综上可知,近红外光谱分析方法主要用于食醋品质指标的定量分析,对于食醋品牌和酿造年份的判别分析相关文献报道较少,作者在前期的研究中,采用变量筛选方法结合偏最小二乘判别分析(PLSDA)对食醋的品牌和酿造年份进行了定性分析^[17],研究中采用部分变量结合线性判别分析方法用于消除光谱信息与分析目标间的非线性关系,因此,本试验拟对非线性判别方法在食醋品质分析中的应用开展研究。

本研究拟采用连续小波变换(CWT)^[18-19]对光谱进行预处理,消除背景和基线漂移等因素对光谱的干扰,提高光谱分辨率。主成分分析(PCA)方法用于光谱降维和样品空间分布分析;自组织映射(SOM)^[20-22]和支持向量机(SVM)^[23-25]两种非线性判别分析方法用于食醋酿造年份和品牌的定性判别,旨在为食醋的品质分析提供一种快速分析方法。

1 材料与方 法

1.1 材 料

食醋样品共计160个,其中125个食醋为“恒顺”牌,其余为其它品牌的食醋。对于“恒顺”品牌食醋,生产日期的跨度为2012~2014年,其中30个样品为2012年4月生产,40个样品为2013年6月生产,其余55个样品为2014年9月生产。样品采用Kennard-Stone方法按照近似2:1的比例进行建模集和测试集的划分,因此对于125个恒顺食醋酿造年份的判别,建模样本数为83,测试集样本数为42;对于食醋品牌的判别,建模样本数为105个,测试集样本数为55。

1.2 仪器及光谱采集

试验仪器为Tensor37傅里叶变换近红外光谱仪(德国Bruker公司)。以空气为参比,将食醋样品注入2 mm比色

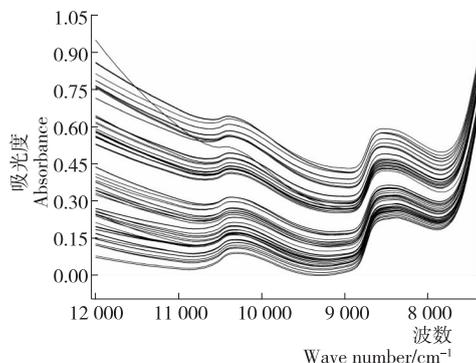


图1 食醋近红外透射光谱

Figure 1 Near-infrared transmission spectra of vinegar

皿,光谱扫描次数64,分辨率 4 cm^{-1} ,光谱采集范围 $7\ 350\sim 12\ 000\text{ cm}^{-1}$ 。每个样品平行采集3条光谱,平均光谱作为最终分析光谱。食醋的近红外透射光谱见图1。

1.3 模式识别方法及其评价指标

SOM网络是一种由全部互相连接的神经元排列形成的无导师自组织学习网络,该方法是通过在学习过程中逐步缩小神经元间的作用邻域,并依据相关的学习规则增强中心神经元的激活程度,从而去掉各神经元之间的侧向连接。SOM网络由输入层和自组织特征映射层组成。输入层只有一个节点,对应于输入矢量;自组织特征映射层由一系列组织在低维网格上的有序节点组成。每个节点对应一个权矢量。SOM网络训练步骤:①网络初始化。给输出层每个节点赋予初始权值 W ,设定初始邻域和学习速率 η ,定义训练结束条件;②计算获胜的神经元。随机选取一个样本 $X_i = (x_1, x_2, \dots, x_3)$ 。计算 X 到每个输出节点之间的距离,选出与样本 X_i 距离最近的神经元 g ;③权值更新。根据式(1)调整神经元 g 和其邻域内包含的神经元的权值 W_{ij} ;④判断算法是否结束,若未结束,返回步骤②。

$$W_{ij} = W_{ij} + \eta(X_i - W_{ij}), \quad (1)$$

式中:

W_{ij} ——输入层节点 i 到竞争层节点 j 的权值。

SVM是基于统计学理论的机器学习方法,适用于解决小样本非线性和高维模式识别,具有很好的泛化能力。SVM是采用分离超平面作为分离训练数据的线性函数来解决非线性问题,运用核函数技术将输入空间中的非线性问题,通过函数映射到高维特征空间,在高维空间中构造线性判别函数。常用的核函数主要包括polynomial核函数、RBF核函数和sigmoid核函数。

上述两种模式识别方法,采用测试集样品的正确识别率(CRR)评价判别模型的优劣,表达式见式(2)。

$$\text{正确识别率} = \frac{\text{正确识别的样品数}}{\text{样品总数}} \times 100\%。 \quad (2)$$

采用Matlab R2010a对数据进行分析。CRR越接近于100%,判别模型的精度越高。

2 结果与分析

2.1 主成分分析用于食醋酿造年份及其品牌鉴别

主成分分析(PCA)方法是经典的特征提取和降维技术,

它可以在不具备任何相关知识背景的情况下对未知样品进行分类判别。将食醋的光谱经主成分变换后,前 3 个主成分的累计贡献率均为 99.997%,因此在后续的计算中仅采用前 3 个主成分代替食醋光谱进行分析。图 2(a)为不同生产年份的 125 个食醋样品光谱数据经主成分分解后保留前 3 个主成分分布图,图 2(b)为混合品牌的 160 个食醋样品的前 3 个主成分分布图。由图 2 可知,对于不同生产年份的食醋样品,2012 年、2013 年以及部分 2014 年的食醋很难区分;而

同品牌的食醋间也存在主成分空间的重叠。尽管从主成分分布图中可以观察到食醋酿造年份及其品牌的聚类趋势,但是对于重叠部分仍然无法进行明确区分,需采用光谱预处理方法对光谱包含的有效信息进行提取以改变样品光谱在主成分空间的分布。研究采用连续小波变换(CWT)^[23-24]对光谱进行预处理。图 3 为食醋光谱经“Haar”小波基函数在分解尺度为 60 条件下进行的导数运算光谱。由图 3 可知,CWT 预处理后的光谱平滑度和聚集度得到明显改善。

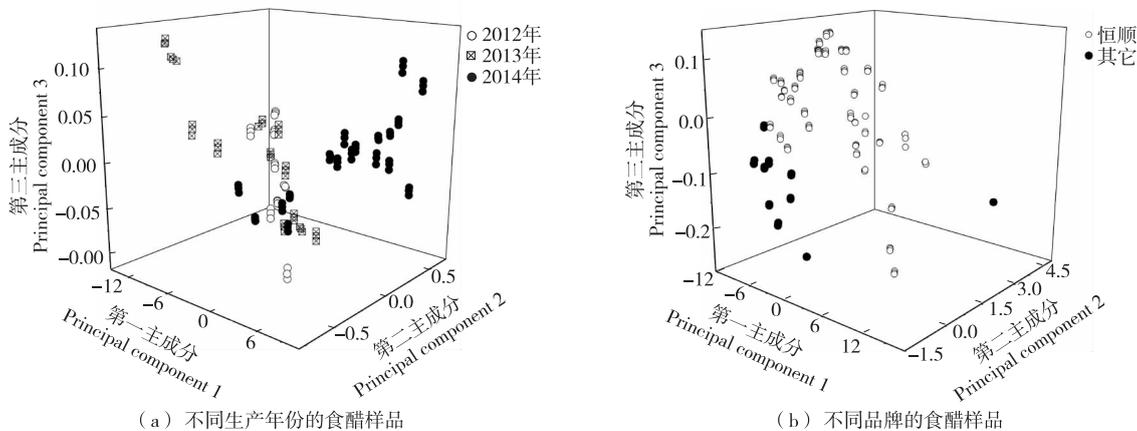


图 2 食醋样品前 3 个主成分分布图

Figure 2 Distribution of the three principal components for vinegar samples

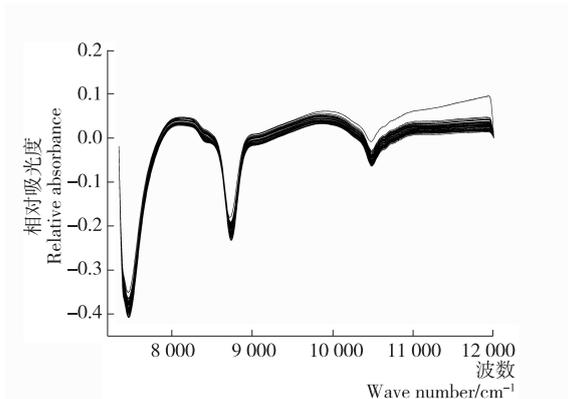


图 3 食醋样品连续小波变换光谱图

Figure 3 Continuous wavelet transform spectra of vinegar samples

为了进一步考察 CWT 方法对样品分类的影响,分别对经 CWT 处理后的食醋光谱进行主成分分析,保留前 3 个主成分进行分析。图 4 为经 CWT 处理后不同生产年份和品牌的食醋的主成分分布图。由图 4 可知,经 CWT 处理后,样本的主成分空间发生改变(图 2),聚类结果得到了部分改善。然而采用该方法无法准确的计算不同生产年份和各个品牌间的明确界限和 CRR 值,因此需要进一步采用其它有监督的模式识别方法对样品进行识别。

2.2 SOM 用于食醋酿造年份及其品牌分析

SOM 是由输入层和自组织特征映射层组成的两层网络。创建 SOM 神经网络前需要确定网络的结构,即确定竞

争层神经元的数目及拓扑结构。在本研究中,采用 4×4 的拓扑结构,因此 16 个神经元用于对应食醋的相应特征,迭代次数设为 200。分别将食醋原始光谱和 CWT 光谱的前 3 个主成分作为 SOM 网络训练的输入特征。SOM 训练和预测结果见表 1。由表 1 可知,对于食醋生产年份识别的 SOM 模型,PCA—SOM 的正确识别率为 63.16%,而 CWT—PCA—SOM 得到较好的分类结果,正确识别率提高为 97.37%,表明 CWT 预处理方法具有提取有效信息和抑制噪声的作用。对于 CWT—PCA—SOM 模型的校正集样本:2012 年食醋样本对应的获胜神经元编号为 3、4、7、8;2013 年食醋样本对应的获胜神经元编号为:7、11、12、15、16;2014 年食醋样本对应的获胜神经元编号为:1、2、3、5、6、9、10、13;而对于测试集样本,如果对应的神经元编号为 3 时,无法正确判断是属于 2012 年还是 2014 年生产,因为校正样本中 2012 年和 2014 年的获胜神经元都包含了 3 号;同理,如果测试集样本对应 7 号神经元,也无法准确判断该样品是属于 2012 年还是 2013 年生产;另外编号 14 的神经元为死神经元(校正集样本中各年份对应的获胜神经元中均没有包含 14 号)。对于食醋品牌识别的 SOM 模型(表 2),CWT—PCA—SOM 得到较好的分类结果,正确识别率为 100%,对于食醋品牌的 SOM 分类,校正集中“恒顺”牌对应的神经元编号为 1~12 号,其它品牌对应的编号为 13~16,而测试集中,“恒顺”品牌的神经元编号均介于 1~12,其它品牌的编号介于 13~16,因此对于食醋品牌的判别,CWT—PCA—SOM 的识别率为 100%。

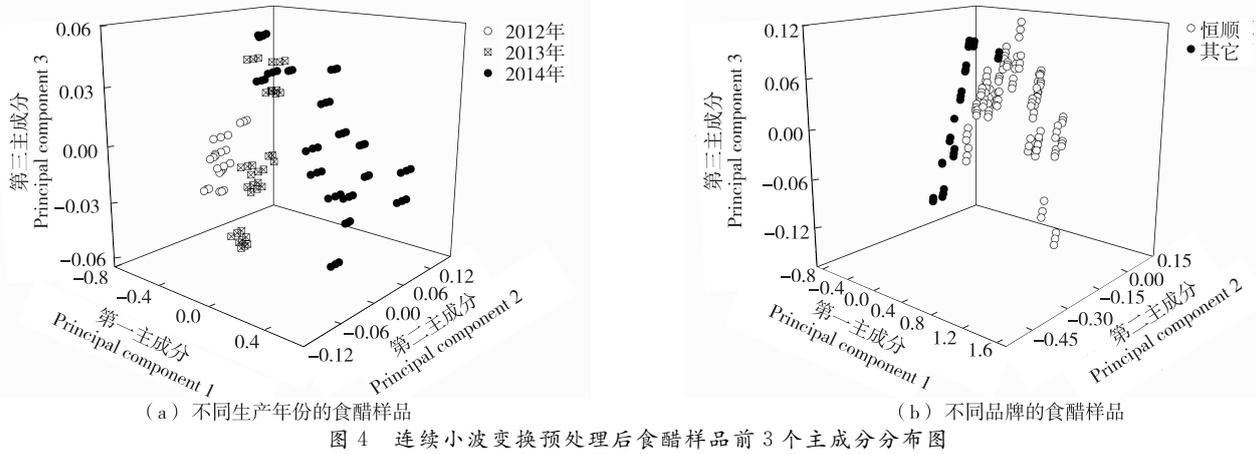


Figure 4 Distribution of the three principal components for CWT spectra of vinegar samples

2.3 SVM 用于食醋酿造年份及其品牌分析

在创建 SVM 分类模型时,需要对核函数及相关参数进行优化,得到最佳的分类效果。RBF 核函数由于简单多变的函数形式被广泛应用^[26-27]。利用留一法交叉验证寻找最佳的参数 c (惩罚因子) 和参数 g (RBF 核函数的方差)。在 SVM 模型优化时,优化时间与待优化的光谱特征变量数有关,变量数目越多优化时间越长。当模型的性能相同时,为了减少计算时间,优先选择惩罚因子 c 比较小的参数组合,主要是因为 c 越大,最终得到的支持向量数越多,模型就越复杂,计算量越大。

为了探讨光谱预处理方法和特征变量数目对模型精度和优化速度的影响,分别对原始光谱、PCA 处理后的光谱和 CWT-PCA 处理后的光谱进行 SVM 模型的建立及评价,评价指标包括识别率和模型优化时间。结果见表 3。由表 3 可

知,光谱经 PCA 降维后,优化时间明显减小,食醋生产年份模型的优化时间从 1 599.64 s 减小为 87.83 s,食醋品牌模型的优化时间从 2 026.40 s 减小为 70.55 s; CWT-PCA-SVM 可实现食醋生产年份和品牌的 100% 识别。

3 结论

本研究采用 SOM 和 SVM 两种模式识别方法结合近红外光谱对食醋酿造年份及其品牌进行识别。CWT 和 PCA 方法的引入进一步提高了 SOM 和 SVM 方法的建模效率和判别精度。近红外光谱结合 SOM 和 SVM 方法可实现食醋酿造年份及其品牌的正确判别,与 SOM 方法相比,SVM 在食醋酿造年份的识别中可以得到更好的识别结果。近红外光谱结合 CWT-PCA-SOM 和 CWT-PCA-SVM 为食醋品质鉴别提供一种快速的分析方法。与前期的研究结果

表 1 SOM 方法对食醋样品不同酿造年份的识别结果

Table 1 Recognition results of different production year of vinegar based on SOM method

年份	PCA			WT-PCA		
	校正神经元	测试神经元	测试识别率/%	校正神经元	测试神经元	测试识别率/%
2012 年	{4 8 12 14 15}	{4 7}		{3 4 7 8}	{4 8}	
2013 年	{3 6 7 11 14 15}	{6 12}	63.1%	{7 11 12 15 16}	{4 12 15}	97.37
2014 年	{1 2 5 9 10 13 16}	{10 13 14 16}		{1 2 3 5 6 9 10 13}	{2 3 5}	

表 2 SOM 方法对食醋样品不同品牌的识别结果

Table 2 Recognition results of different brand of vinegar based on SOM method

品牌	PCA			WT-PCA		
	校正集神经元	测试集神经元	识别率/%	校正集神经元	测试集神经元	识别率/%
恒顺牌	{1 2 3 5 6 7 8 10 11 12 13 14 15}	{2 3 6 7 8 10 11 12 13 14}	89.58	{1 2 3 4 5 6 7 8 9 10 11 12}	{13 14 15 16}	100
其它	{4 9 13 15 16}	{4 9 13 15 16}		{2 3 4 6 7 8 9 10 11 12}	{13 14 15 16}	

表 3 SVM 方法对食醋样品不同分类指标的识别结果

Table 3 Recognition results of different indexes of vinegar based on SVM method

识别指标	生产年份			品牌		
	优化时间/s	校正识别率/%	测试识别率/%	优化时间/s	校正识别率/%	测试识别率/%
Full spectrum	1 599.64	100	68.42	2 026.40	100	100
PCA	87.83	100	68.42	70.55	100	100
CWT-PCA	85.67	100	100.00	72.95	100	100

比较,本研究采用两种非线性定性分析方法得到的分析结果与部分变量结合线性判别分析方法建模结果相当,结果表明:在复杂的数据建模时,可以通过非线性建模方法替代优选变量结合线性建模的分析方法。

参考文献

- [1] Pizarro C, Esteban-Diez I, Saenz-Gonzalez C, et al. Vinegar classification based on feature extraction and selection from headspace solid-phase microextraction/gas chromatography volatile analyses: A feasibility study[J]. *Analytica Chimica Acta*, 2008, 608(1): 38-47.
- [2] 刘杨岷, 张家骊, 王利平, 等. 食醋风味成分比较研究[J]. *食品与机械*, 2005, 21(5): 40-43.
- [3] Ubeda C, Callejon R M, Hidalgo C, et al. Determination of major volatile compounds during the production of fruit vinegars by static headspace gas chromatography-mass spectrometry method [J]. *Food Research International*, 2011, 44(1): 259-268.
- [4] Callejon R M, Amigo J M, Pairo E, et al. Classification of Sherry vinegars by combining multidimensional fluorescence, parafac and different classification approaches [J]. *Talanta*, 2012, 88(15): 456-462.
- [5] Boffoa E F, Tavaresa L A, Ferreira M M C, et al. Classification of Brazilian vinegars according to their ¹H NMR spectra by pattern recognition analysis[J]. *LWT - Food Science and Technology*, 2009, 42(9): 1 455-1 460.
- [6] Zhang Qin-yi, Zhang Shun-ping, Xie Chang-sheng, et al. Characterization of Chinese vinegars by electronic nose[J]. *Sensors and Actuators B: Chemical*, 2006, 119(2): 538-546.
- [7] Liu Fei, He Yong, Wang Li. Determination of effective wavelengths for discrimination of fruit vinegars using near infrared spectroscopy and multivariate analysis[J]. *Analytica Chimica Acta*, 2008, 615(1): 10-17.
- [8] Cocchi M, Durante C, Foca G, et al. Application of a wavelet-based algorithm on HS—SPME/GC signals for the classification of balsamic vinegars [J]. *Chemometrics and Intelligent Laboratory Systems*, 2004, 71(2): 129-140.
- [9] Hsieh Chang-wei, Li Po-hsien, Cheng Ju-yun, et al. Using SNIF—NMR method to identify the adulteration of molasses spirit vinegar by synthetic acetic acid in rice vinegar [J]. *Industrial Crops and Products*, 2013, 50: 904-908.
- [10] Yip W L, Soosainather T C, Dyrstad K, et al. Classification of structurally related commercial contrast media by near infrared spectroscopy[J]. *Journal of Pharmaceutical and Biomedical Analysis*, 2014, 90(5): 148-160.
- [11] Coppa M, Revello-Chion A, Giaccone D, et al. Comparison of near and medium infrared spectroscopy to predict fatty acid composition on fresh and thawed milk[J]. *Food Chemistry*, 2014, 150(1): 49-57.
- [12] 杨代明, 方宣启. 非线性化学指纹图谱技术在食醋鉴别中的应用研究[J]. *食品与机械*, 2014, 30(5): 68-71.
- [13] Casale M, Abajo M J S, Saiz J M G, et al. Study of the aging and oxidation processes of vinegar samples from different origins during storage by near-infrared spectroscopy[J]. *Analytica Chimica Acta*, 2006, 557(2): 360-366.
- [14] Liu Fei, He Yong, Wang Li. Comparison of calibrations for the determination of soluble solids content and pH of rice vinegars using visible and short-wave near infrared spectroscopy[J]. *Analytica Chimica Acta*, 2008, 610(2): 196-204.
- [15] Saiz-Abajo M J, Gonzalez-Saiz J M, Pizarro C. Prediction of organic acids and other quality parameters of wine vinegar by near-infrared spectroscopy: A feasibility study[J]. *Food Chemistry*, 2006, 99(3): 615-621.
- [16] Chen Quan-sheng, Ding Jiao, Cai Jian-rong, et al. Rapid measurement of total acid content (TAC) in vinegar using near infrared spectroscopy based on efficient variables selection algorithm and nonlinear regression tools [J]. *Food Chemistry*, 2012, 135(2): 590-595.
- [17] 夏蓉, 郝勇. 近红外光谱在食醋品牌和贮藏年份鉴别中的应用研究[J]. *中国酿造*, 2012, 31(11): 27-29.
- [18] Shao Xue-guang, Ma Chao-xiong. A general approach to derivative calculation using wavelet transform[J]. *Chemometrics and Intelligent Laboratory Systems*, 2003, 69(2): 157-165.
- [19] Leung A K M, Chau F T, Gao J B. Wavelet transform: A novel method for derivative calculation in analytical chemistry [J]. *Analytical Chemistry*, 1998, 70(2): 5 222-5 229.
- [20] Yan Ai-xia, Nie Xiang-lei, Wang Kai, et al. Classification of Aurora kinase inhibitors by self-organizing map (SOM) and support vector machine (SVM) [J]. *European Journal of Medicinal Chemistry*, 2013, 61: 73-83.
- [21] Merlin P, Sorjamaa A, Mailliet B, et al. X-SOM and L-SOM: A double classification approach for missing value imputation [J]. *Neurocomputing*, 2010, 73(9): 1 103-1 108.
- [22] Nakayama N, Oketani M, Kawamura Y, et al. Classification of acute liver failure of indeterminate etiology: usefulness of clustering analysis using a self-organizing map (SOM)[J]. *Journal of Hepatology*, 2011, 54(1): S369-S370.
- [23] Devos O, Downey G, Duponchel L. Simultaneous data pre-processing and SVM classification model selection based on a parallel genetic algorithm applied to spectroscopic data of olive oils [J]. *Food Chemistry*, 2014, 148(1): 124-130.
- [24] Wang Ya-sheng, Yang Meng, Wei Gao, et al. Improved PLS regression based on SVM classification for rapid analysis of coal properties by near-infrared reflectance spectroscopy[J]. *Sensors and Actuators B: Chemical*, 2014, 193(2): 723-729.
- [25] Sugumaran V, Ramachandran K. Effect of number of features on classification of roller bearing faults using SVM and PSVM[J]. *Expert Systems with Applications*, 2011, 38(4): 4 088-4 096.
- [26] Mu T T, Nandi A K. Breast cancer detection from FNA using SVM with different parameter tuning systems and SOM-RBF classifier[J]. *Journal of the Franklin Institute*, 2007, 344(3/4): 285-311.
- [27] Norinder U, Ek M E. QSAR investigation of NaV1.7 active compounds using the SVM/Signature approach and the Bioclipse Modeling platform[J]. *Bioorganic & Medicinal Chemistry Letters*, 2013, 23(1): 261-263.