核磁共振氢谱结合 PCA－SVM 算法分类鉴别食用植物油

李　玮,姜　洁,杨红梅,王　浩,贾婧怡
(北京市食品安全监控和风险评估中心,北京 100041)

摘　要:采用核磁共振氢谱(^{1}H-NMR)结合主成分分析-支持向量机法(PCA-SVM)对7种市面上常见的食用植物油进行了分类研究。首先运用PCA法对预处理后的各食用植物油的^{1}H-NMR谱图积分数据进行降维处理,然后选用前2个主成分作为SVM的输入变量,建立预测模型,再对测试集样品进行预测,以实现食用植物油的种类鉴别,并与簇类独立软模式法(SIMCA)的分类结果进行了比较。结果显示:采用网格划分法优化得到模型最优核函数参数值为1.7411,最优惩罚参数值为0.3299,以最优参数建立的PCA-SVM食用植物油分类模型对测试集的分类正确率为100%,高于SIMCA分类法的61.90%。建立的^{1}H-NMR结合PCA-SVM法食用植物油分类模型,可以快速、有效的鉴别食用植物油种类,适合实际食品检测工作中建模样本有限的实际,为食用植物油的品质鉴别和质量控制提供分析方法。

关键词:核磁共振氢谱,食用植物油,主成分分析-支持向量机,分类方法

Classification of edible vegetable oils based on ^{1}H-NMR spectroscopy and PCA-SVM

LI Wei, JIANG Jie, YANG Hong-mei, WANG Hao, JIA Jing-yi
(Beijing Municipal Center for Food Safety Monitoring and Risk Assessment, Beijing 100041, China)

Abstract: To establish a method for the classification of edible oils by ^{1}H-NMR spectroscopy and PCA-SVM and to compare its effectiveness with that of SIMCA. First, the PCA method was used to reduce the dimensionality of independent variables. Then the first two principal components were selected as input variables of the support vector machine (SVM), based on the established PCA-SVM prediction model. The seven kinds of oils could be identified by the proposed technique. The results revealed that the value of g and c were 1.7411 and 0.3299, respectively, which were optimized by grid method. The accuracy of prediction could reach to 100% with the PCA-SVM model, while that was only 61.90% with SIMCA model. It was validated by results that the combination of ^{1}H-NMR spectroscopy with PCA-SVM could achieve the classification of edible oils quickly and effectively.

Key words: ^{1}H-NMR; edible oils; PCA-SVM; classification
中图分类号:TS255.1　　**文献标识码**:A　　**文 章 编 号**:1002-0306(2018)08-0205-05
doi:10.13386/j.issn1002-0306.2018.08.037
引文格式:李玮,姜洁,杨红梅,等.核磁共振氢谱结合PCA-SVM算法分类鉴别食用植物油[J].食品工业科技,2018,39(8):205-209.

食用植物油是人们日常饮食不可缺少的食物之一,是我国居民维生素E的首要来源[1]。目前市场上的食用油种类繁多,食用油因其种类不同、营养价值不同而价格差异很大。一些不法商家为谋取个人利益,常以菜籽油、棕榈油等廉价植物油掺兑优质、高价油品中以降低生产成本,从中谋取暴利。更有生产厂家甚至将非食用油脂(俗称"地沟油")按一定比例勾兑到正规厂家生产的优质油品中,严重地损害消费者的利益和危害消费者的身体健康[2]。为保护合法生产经营者和消费者的利益,有必要进行食用油种类的鉴别。

目前,对食用植物油分类多采用气相色谱法、红外、拉曼光谱结合聚类分析、辨别分析等化学计量学的方法,但这些方法都存在各自的缺点[3-7]。例如,气相色谱法需要对样品进行衍生化,前处理繁琐、检测时间长;红外、拉曼等光谱法由于受其检测原理所限,其图谱对食用植物油混合物体系内各成分解释能力有限。另外,聚类、辨别分析等传统的化学计量

收稿日期:2017-08-18
作者简介:李玮(1984-),女,博士,高级工程师,研究方向:食品营养与安全,E-mail:liwei@bjmu.edu.cn。
基金项目:北京市科技计划重大项目(D16110500210000)。

学方法在处理分类问题时,一般需要事先知道样本的先验分布,并要求有足够多的样本数据,而这些要求在实际应用中往往难以达到。所以,在实际工作中应选用适合小样本集的数据处理分析方法。支持向量机(support vector machine,SVM)是统计学习理论中最实用的部分,其在分类问题中既考虑分类误差最小,又考虑分类线的结构,提高了机器学习的泛化能力。此外,SVM 还通过引入核函数的方法使计算的复杂度不再取决于空间维数,而是取决于样本数量,尤其是样本中的支持向量数,特别适合小样本集数据的分类[8-11]。

核磁共振(nuclear magnetic resonance,NMR)是鉴定有机化合物结构和研究化学动力学等的极为重要的手段,具有前处理简单,不损伤样品,结构信息丰富等优点。NMR 技术在食品领域的应用初期主要采用低场技术研究水在食品中的状态[12],但随着超导技术、计算机技术和脉冲傅立叶变换波谱仪的迅速发展,其在食品领域的研究及应用领域逐渐扩大[13-15]。近年来,国内外研究者采用高场 NMR 技术结合化学计量学方法,在食品的真伪鉴别、产地溯源等领域展开了大量的研究与应用[16-19],但这些研究多采用聚类、辨别分析等传统的化学计量学方法,少见采用 SVM 结合 NMR 的分类方法报道。

本研究对市售大豆油、花生油、玉米油、葵花籽油、橄榄油、芝麻油、菜籽油这 7 种市面上常见食用植物油的^{1}H-NMR 图谱进行测定,结合 PCA-SVM 算法建立分类模型。随机选取已知食用油种类的预测样本对模型进行检验,根据预测结果分析模型的可靠性,并与传统的簇类独立软模式法(soft independent modeling of class analogies,SIMCA)算法建立的分类模型进行对比。

1 材料与方法

1.1 材料与仪器

7 种食用植物油　共 119 个样品,超市,其中大豆油 19 个,花生油 18 个,玉米油 19 个,葵花籽油 17 个,橄榄油 17 个,芝麻油 17 个,菜籽油 10 个;氘代氯仿($CDCl_3$)　氘代度:99.8%,美国 CIL 公司;Norell 5 mm 核磁管　美国 Norell 公司。

AVANCE 600 MHZ 超导傅里叶变换 NMR 仪　配有 CPBBO 探头,Topspin 3.2 处理软件及 60 位自动进样器,瑞士 Bruker 公司。

1.2 实验方法

1.2.1　样品溶液的制备　吸取植物油样品 200 μL 于 2 mL 的 EP 管中,加入 800 μL $CDCl_3$,涡旋 30 s,使样品与试剂混合均匀。取 600 μL 混合溶液于 5 mm 核磁管中,待测。

1.2.2　训练集与测试集样本的确定　在 Matlab 中用 randperm 函数随机将每种食用植物油样品分为两组,一组为训练集(training set)样本,一组为测试集(testing set)样本,保证训练集与测试集样本数量比约为 3∶2,于是得到 75 个样本的训练集和 42 个样本的测试集。

1.2.3　仪器条件　NMR 仪^{1}H 载波频率为 600.13 MHz,采用 Bruker 标准脉冲 zg30,检测温度为 298 K,^{1}H 的 90°脉冲宽度为 11.90 μs,谱宽为 12019.23 Hz,中心频率为 3600.78 Hz,脉冲延迟时间为 4 s,扫描次数为 32,空扫次数为 2。

1.2.4　样品测定及谱图处理　在 1.2.2 项实验条件下,调整仪器参数、调谐、控温、匀场、采样及傅里叶变换,得到^{1}H-NMR 图谱。测得谱使用 Bruker Topspin 3.2 软件处理,变换点数为 65536,线宽因子为 1.00 Hz,用指数窗函数处理,基线和相位校正均采用手动方式进行,四甲基硅烷(TMS)为内标信号(δ 0.00)。

1.2.5　分类模型的建立　处理后的图谱用 MestReNova(version 6.0.1,Spain)软件,以 δ 0.005 积分段对化学位移区间 δ 0.15~10.00 进行分段积分,去除谱中 δ 7.21~7.30 区域的信号后进行面积归一化处理,得到样品核磁图谱转换形成的典型二维矩阵,其中每行代表一个样本,每列代表样本在同一化学位移内的强度积分相对值。分别采用 PCA-SVM 和 SIMCA 法,对训练集数据进行分类模型的建立,用得到的模型预测测试集。程序采用 Matlab(V7.0.4,Mathworks Inc,USA)软件编写、运行。

2 结果与分析

2.1 7 种食用植物油的典型^{1}H-NMR 图谱分析

食用植物油中主要成分为甘油三酯,图 1 所示为 7 种食用植物油典型的^{1}H-NMR($CDCl_3$)谱图,图谱显示主要存在 9 组信号峰,参考文献[20-21],对这 9 组主要信号峰进行了信号归属(表 1)。图谱中所示,δ 0.84~0.92 信号为脂肪链末端甲基质子信号,δ 1.22~1.39 信号为长链脂肪酸一般性亚甲基质子信号,δ 1.56~1.66 信号为脂肪链上与羰基相隔一个亚甲基的亚甲基上的质子信号,δ 1.96~2.09 信号为与脂肪链上双键相连的亚甲基质子信号,δ 2.27~2.36 信号为脂肪链上与羰基直接相连的亚甲基质子信号,δ 2.74~2.80 信号为脂肪链上两个双键之间亚甲基质子信号,δ 4.10~4.32 和 δ 5.24~5.29 信号分别为甘油三酯中丙三醇的亚甲基和次甲基质子信号,δ 5.30~5.41 信号为非共轭脂肪酸不饱和质子信号。

表 1　食用植物油中脂类成分^{1}H-NMR 主要化学位移归属

Table 1　^{1}H-NMR major chemical shift assignments of fat in edible oils

序号	化学位移(ppm)	归属基团
1	0.84~0.92	$-CH_3$
2	1.22~1.39	$-(CH_2)_n-$
3	1.56~1.66	$-OCO-CH_2-CH_2-$
4	1.96~2.09	$-CH_2-CH_2=CH_2-$
5	2.27~2.36	$-OCO-CH_2-$
6	2.74~2.80	$=CH-CH_2-CH=$
7	4.10~4.32	$-CH_2OCOR$
8	5.24~5.29	$>CHOCOR$
9	5.30~5.41	$-CH=CH-$

2.2 PCA-SVM 分类模型的建立

2.2.1　自变量的降维　经数据处理后的训练集数据

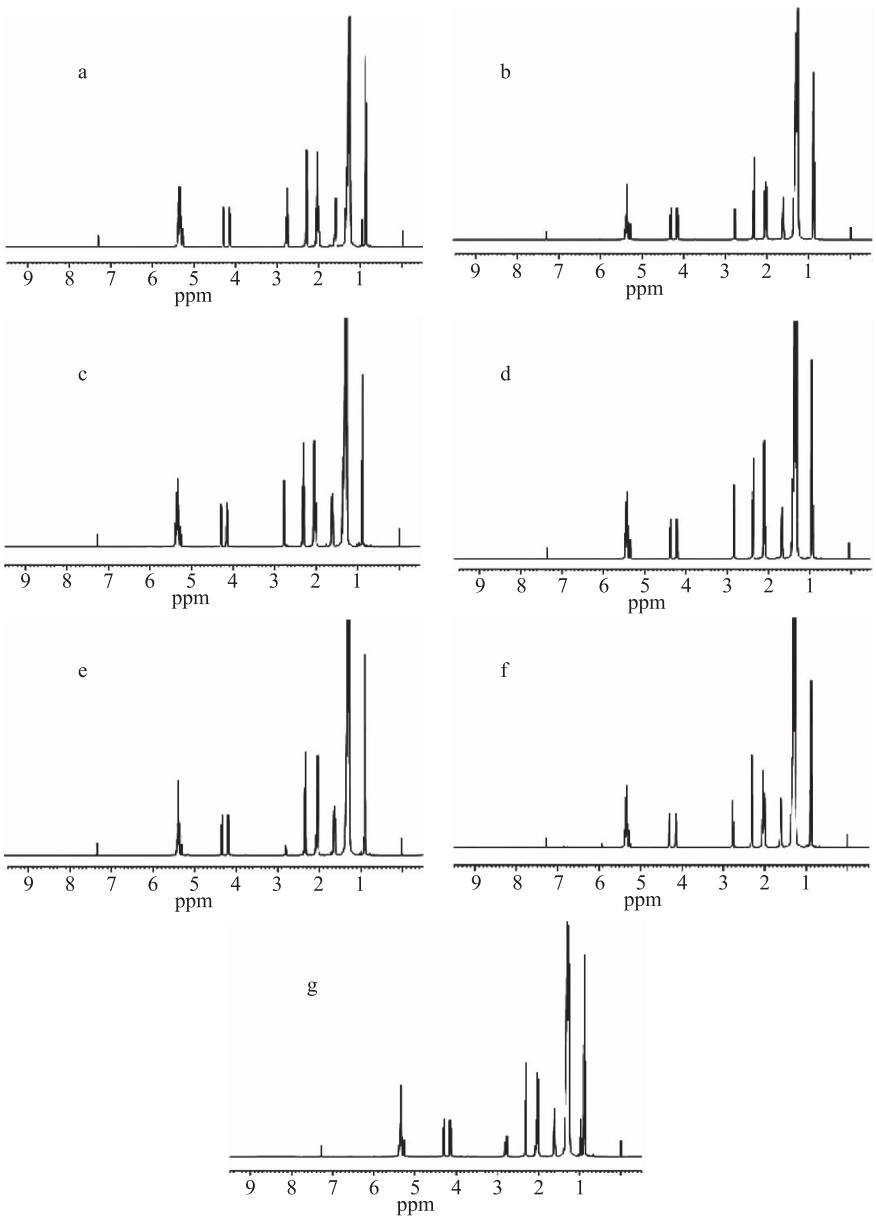

图1　7种常见植物油的典型^{1}H-NMR谱图($CDCl_3$)

Fig.1　^{1}H-NMR spectrum of edible oils($CDCl_3$)

注:a为大豆油;b为花生油;c为玉米油;d为葵花籽油;e为橄榄油;f为芝麻油;g为菜籽油。

是由75个样本,1950个变量(图谱分段积分获得)构成的75×1950的一个矩阵,其中每行代表一个样本,每列代表样本在同一化学位移内的强度积分相对值。当过多的自变量因子输入分类模型时,不仅会影响到模型的运算速度,还会引入噪音,影响了模型的预测精度。为解决训练集自变量过多问题,在建立模型前,首先采用PCA分析方法,将训练集数据自变量降维,得到新的特征变量的潜在变量数为2,累计贡献率在95%以上。因此,在建立分类模型时,以PCA分析得到的新特征变量代替原有的自变量作为模型建立的输入自变量。

2.2.2　惩罚参数c和核函数参数g的优化　SVM对模型参数的选择很敏感,为了得到比较理想的分类准确率,需要调节相关的惩罚参数c和核函数参数g。惩罚参数c代表模型对误差的宽容度,过高的c会导致过学习状态的发生,即训练集准确率很高而测试集准确率很低,而c过小将导致训练不完全。核函数参数g是径向基函数自带的一个参数,它决定了数据映射到新的特征空间后的分布。模型中这两个参数对支持向量个数、模型计算的复杂度和精确度都有很大的影响[22]。因此,为了提高模型的准确度和降低模型的复杂度,本研究分别采用网格划分法(Grid)和粒子群优化算法(PSO)对模型的惩罚参数c和核函数参数g进行了优化。表2结果显示,在建立SVM分类模型时,采用Grid法优化得到的c和g值,对测试集的预测的正确率更高,因此选用Grid法优化的c和g的值作为SVM分类模型的建模参数。

2.2.3　PCA-SVM分类模型的建立　采用2.2.1中得到的训练集数据新特征变量作为输入自变量,以食

表2 两种优化方法所建校正模型对训练集和测试集样本进行分类判别的结果比较

Table 2 Comparison of the correct recognition and prediction ability of the calibration models developed by Grid and POS

优化方式	优化参数		训练集			测试集		
	c	g	样本数	误判数	正确率(%)	样本数	误判数	正确率(%)
PSO	0.5409	12.1079	75	0	100	42	2	95.23
Grid	0.3299	1.7411	75	0	100	42	0	100

表3 PCA-SVM 和 SIMCA 分类模型的结果对比

Table 3 Comparison of the correct recognition and prediction ability of the calibration models constructed by PCA-SVM and SIMCA

建模方式	训练集			测试集		
	样本数	误判数	正确率(%)	样本数	误判数	正确率(%)
PCA-SVM	75	0	100	42	0	100
SIMCA	75	0	100	42	16	61.90

用植物油种类作为因变量,以 2.2.2 中得到的最优 c,g 值作为模型参数,建立 PCA-SVM 分类模型。采用测试集样品对模型的分类准确率进行评价,结果见表2中 Grid 项结果。结果显示,建立的 PCA-SVM 分类模型对训练集和测试集的分类正确率均为100%。SVM 是一种有监督的机器学习方法,该方法在小样本、非线性和高维数据空间的模式识别问题上拥有传统模式识别方法所不具备的独特优势,特别适用于小样本量的复杂体系分析及数据挖掘。因此,在原理上,在样本量增加情况下,分类的准确率应不会有显著降低,但实际结果,还应通过增加样品量来验证。

2.3 PCA-SVM 分类模型与 SIMCA 分类模型的分类结果对比

2.3.1 SIMCA 分类模型的建立 采用 1.2.2 中得到的训练集数据建立 SIMCA 分类模型,选用中心化法(Center,Ctr)进行数据标度换算。优选的主成分数分别是4、4、4、3、4、5、2 时,7 种食用植物油被 100% 聚类识别。在利用训练集样本建立的 SIMCA 辨别模型对测试集样本进行验证,结果见表3。

2.3.2 两种分类模型效果对比 分别采用 PCA-SVM 和 SIMCA 两种方法建立 7 种食用植物油的分类模型,分别对测试集数据进行预测,结果见表3。结果显示,PCA-SVM 法和 SIMCA 法模型的训练集的分类正确率相同,均为 100%,但对测试集的进行分类时,PCA-SVM 模型要的正确率要远远高于 SIMCA 模型。

SIMCA 是一种基于主成分分析的有监督模式识别方法,其核心思想是对训练集中的每个样本分类分别建立一个主成分分析模型以对其进行描述。该方法以经典的统计学数学理论为依据的,着眼于最大似然的基点,要求"残差平方和"最小,因而通常需要训练样本数目接近无限大时其有效性才能被真确的显露出来。SVM 是一种有监督的机器学习方法,该方法在小样本、非线性和高维数据空间的模式识别问题上拥有传统模式识别方法所不具备的独特优势,因此目前在涉及统计分类以及回归分析的诸多相关领域中得到了广泛的应用,特别适用于复杂体系分析及数据挖掘。但因在实际的食品检测工作中所能获得的样本数量往往非常有限,因此 SVM 法更适合实际监测工作的要求。

3 结论

本研究采用^{1}H-NMR 结合 PCA-SVM 对 7 种市面上常见的食用植物油进行了分类研究。采用网格划分法优化得到模型最优核函数参数为 1.7411,最优惩罚参数为 0.3299。本研究结果显示,采用 PCA-SVM 算法建立的分类模型对测试集样品分类的正确率要远高于 SIMCA 分类模型,因此 SVM 法更适合实际监测工作中建模样本量小的要求,适于食用植物油的快速分类鉴别,可以快速、有效的鉴别食用植物油种类,适合实际食品检测工作中建模样本有限的实际,为食用植物油的品质鉴别和质量控制提供分析方法。

参考文献

[1]马冠生,郝利楠,李艳平,等.中国成年居民食用油消费现状[J].中国食物与营养,2008(9):29-32.

[2]王乐,刘尧刚,陈凤飞,等.地沟油的污染及变质情况研究[J].武汉工业学院学报,2007,26(4):1-4,12.

[3]林晨,张方圆,吴凌涛,等.气相色谱结合化学计量学分析4种食用植物油的指纹图谱[J].分析测试学报,2016,35(4):454-459.

[4]王同珍,陈孝建,安爱,等.气相色谱-质谱技术结合化学计量学对5种动物油进行判别分析[J].分析测试学报,2016,35(5):557-562.

[5]李娟,范璐,毕艳兰,等.红外、近红外光谱-簇类的独立软模式方法识别植物调和油脂[J].分析化学,2010,38(4):475-482.

[6]黄晓东,谢飞飞,尤勇,等.食用植物油傅里叶变换红外光谱鉴别的研究[J].安徽工程大学学报,2014,29(4):4-7.

[7]赵薇,刘翠玲,孙晓荣,等.应用拉曼光谱技术识别食用油的种类[J].食品科技,2015,40(3):274-277.

[8]Blanco M,Coello J,Iturriaga H,et al.Calibration in non-linear near infrared reflectance spectroscopy:A comparison of several methods[J].Analytica Chimica Acta,1999,384(2):207-213.

[9]Despagne F,Massart D L,Chabot P.Development of a Robust Calibration Model for Nonlinear In-Line Process Data[J].Analytical Chemistry,2000,72(7):1657-1665.

[10]Blanco M,Pages J.Classification and quantitation of finishing

oils by near infrared spectroscopy[J]. Analytica Chimica Acta, 2002,463(2):295-303.
[11]王春艳,史晓凤,李文东,等.基于主成分和支持向量机浓度参量同步荧光光谱油种鉴别[J].分析测试学报,2014,33(3):289-294.
[12]R H inrichs,J Gotz,H Weisser.Water-Holding Capacity and Structure of Hydrocolloid - gels WPC - gels and Yogurts Characterised by means NMR[J].Food Chemistry,2003,82(1):155-160.
[13]李玮,姜洁,路勇,等.NMR 氢谱定量测定奶酪中总共轭亚油酸的含量[J].食品科学,2015,36(10):134-138.
[14]姜洁,李玮,路勇,等.核磁共振脉冲宽度法测定婴幼儿乳粉中乳糖、蔗糖含量[J].食品工业科技,2015,36(8):68-71,77.
[15]李玮,贾婧怡,姜洁,等.NMR 氢谱法分析市售奶油中的脂肪酸[J].食品工业科技,2016,37(23):319-323.
[16]李玮,贾婧怡,李龙,等.核磁共振代谢组学技术鉴别天然奶油与人造奶油[J].食品科学,2017,38(12):278-285.
[17]Andreotti G, Trivellone E, Lamanna R, et al. Milk identification of different species: ^{13}C - NMR spectroscopy of triacylglycerols from cows and buffaloes' milks[J]. Journal of Dairy Science,2000,83(11):2432-2437.
[18]G Vigli,Angelos P,Apostolos S,et al.Classification of edible oils by employing ^{31}P and ^{1}H NMR spectroscopy in combination with multivariate statistical analysis.A proposal for the detection of seed oil adulteration in virgin olive oils[J].Journal of Agricultural Food Chemistry,2003,51:5715-5722.
[19]Zou X Q, Huang J H, Jin Q Z, et al. Lipid composition analysis of milk fats from different mammalian species: potential for use as human milk fat substitutes[J].Journal of Agricultural Food Chemistry,2013,61:7070-7080.
[20]Raffaele S,Francesco A,Livio P.^{1}H and of Virgin Olive Oil. ^{13}C NMR An Overview[J]. Magnetic Resonance in Chemistry, 1997,35:S133-S145.
[21]Maria L I, Sopelana P, Maria D G.^{1}H Nuclear Magnetic Resonance monitoring of the degradation of margarines of varied compositions when heated to high temperature[J]. Food Chemistry,2014,165:119-128.
[22]李盼池,许少华.支持向量机在模式识别中的核函数特性分析[J].计算机工程与设计,2005,26(2):2302-2304.

(上接第 177 页)

[15]李尽哲,王德芝,黄雅琴.响应面法研制灵芝火棘复合保健饮料[J].食品工业科技,2016,37(21):238-242.
[16]赵镭,刘文,汪厚银.食品感官评价指标体系建立的一般原则与方法[J].中国食品学报,2008(3):121-124.
[17]宋慧.红枣玫瑰花复合保健饮料的生产工艺研究[J].中国食品添加剂,2013,(5):63-69.
[18]高辉,贾长英,吕忠政,等.雪莲果功能性饮料的制备研究[J].食品安全质量检测学报,2016,7(7):2835-2839.
[19]陈晓芳,张倩,吴文倩,等.枇杷花醇提物止咳化痰作用实验研究[J].中成药,2013,35(1):167-169.
[20]李尽哲,叶兆伟,黄雅琴.蛹虫草桑叶复合保健饮料的研制[J].食品研究与开发,2016,37(3):77-79.

(上接第 204 页)

AOAC International,2003,86(2),412-431.
[14]李丽娟,吴青,翁柔丹,等.气相色谱方法同时测定果蔬及粮谷中 6 种除草剂残留[J].食品工业科技,2014,35(18):84-88.
[15]夏虹,彭茂民,王小飞,等.QuEChERS 法和 UPLC-MS-MS 法快速测定稻谷中 4 种磺酰脲类除草剂的残留量[J].食品科技,2015(8):325-328.
[16]彭俏容,于淑新,赵连海,等.QuEChERS-HPLC 快速测定白酒中 13 种邻苯二甲酸酯[J].酿酒科技,2014(1):89-92.
[17]Ferreira I, Fernandes J O, Cunha S C. Optimization and validation of a method based in a QuEChERS procedure and gas chromatography-mass spectrometry for the determination of multi-mycotoxins in popcorn[J]. Food Control, 2012, 27(1):188-193.
[18]Gonzálezcurbelo M á, Hernándezborges J, Borgesmiquel T M, et al. Determination of pesticides and their metabolites in processed cereal samples[J].Food Additives & Contaminants Part A,2012,29(1):104-16.
[19]GB/T 21928-2008.食品塑料包装材料中邻苯二甲酸酯的测定[S].
[20]Gillespie M.,Walters S.Rapid Clean-Up of Fat Extracts for Organophosphorus Pesticide Residue Determination Using C18 Solid-Phase Extraction Cartridges[J].Analytica Chimica Acta, 1991,245(Supplement C):259-265.
[21]李淑娟,于杰,高玉生,等.HPLC-MSMS 测定果蔬中有机磷类农药的基质效应[J].食品工业科技,2017,38(6):49-53.
[22]GB 9685-2008.食品容器、包装材料用添加剂使用卫生标准[S].