基于傅里叶变换红外光谱技术和软独立模式分类法的牛奶分类识别

穆海波[1]，殷秀秀[2,3]，艾连中[1]，顾小红[2,*]
(1.光明乳业股份有限公司 乳业生物技术国家重点实验室，上海 200436；2.江南大学 食品科学与技术国家重点实验室，江苏 无锡 214122；3.江南大学食品学院，江苏 无锡 214122)

摘 要：利用傅里叶变换红外光谱法(FTIR)结合软独立模式分类法(SIMCA)对不同类别的牛奶进行识别。通过对光谱数据基线校正和 Savitzky-Golay 平滑处理后，在 3100～850cm^{-1} 光谱区域，利用留一交互验证法建立获得主成分分析(PCA)最优模型。在 α＝5% 显著水平下，最优模型对纯牛奶、低乳糖奶、低脂奶和高蛋白奶的识别率分别为 80%、80%、100% 和 80%，拒绝率分别为 93%、100%、100% 和 93%。表明 FTIR 结合 SIMCA 可成为快速识别牛奶类别的有效方法。

关键词：傅里叶变换红外光谱法；软独立模式分类法；牛奶；模式识别

Discrimination of Milk by FTIR and Soft Independent Modeling of Class Analogy

MU Hai-bo[1]，YIN Xiu-xiu[2,3]，AI Lian-zhong[1]，GU Xiao-hong[2,*]
(1. State Key Laboratory of Dairy Biotechnology, Bright Dairy & Food Co. Ltd., Shanghai 200436, China；
2. State Key Laoratory of Food Science and Technology, Jiangnan University, Wuxi 214122, China；
3. School of Food Science and Technology, Jiangnan University, Wuxi 214122, China)

Abstract: Fourier transform infrared spectroscopy (FTIR) combined with soft independent modeling of class analogy (SIMCA) method was employed to the identification of different varieties of milk. The optimized PCA model was built by leave-one-out cross-validation (LOOCV) method after series of pre-treatments such as baseline correction and Savitzky-Golay smoothing in the region of 3100 — 850 cm^{-1}. Under the α =5% significance level, the identification rates of this model for pure milk, low lactose milk, low fat milk and high protein milk were 80%, 80%, 100% and 80%, respectively, and the rejection rates were 93%, 100%, 100% and 93%, respectively. This indicates that FTIR combined with SIMCA is a valid method for rapid identification of different varieties of milk.

Key words: FTIR；SIMCA；milk；pattern recognition
中图分类号：TS252.7 文献标识码：A 文章编号：1671-5187(2012)02-0034-04

近年来，乳制品的营养价值逐渐被广大消费者认可，消费量逐年上涨，市售牛奶的种类也是越来越多，除了最常见的纯牛奶以外，还有高蛋白奶、低脂奶、低乳糖奶等，从而满足不同消费者的营养需求。牛奶中蛋白质、脂肪和乳糖含量是决定牛奶品质的核心指标，因此对不同种类牛奶进行分类鉴别具有重要意义[1]。

傅里叶红外光谱(fourier translation infrared spectroscopy，FTIR)是一种快速、无损的检测技术，该方法具有操作简便、无需样品前处理、不消耗化学试剂等优点[2]。目前，已有将 FTIR 技术结合化学计量学应用于乳制品、蜂蜜、茶叶等的模式识别研究的相关报道。Karoui 等[3]利用衰减全反射傅里叶变换红外光谱(ATR-FTIR)技术结合主成分分析(PCA)建立了奶酪产地的分类模型，达到了快速识别的效果；冯宇等[4]采用漫反射傅里叶变换红外(DR-FTIR)光谱技术结合 PCA 对 4 种

收稿日期：2011-12-25
作者简介：穆海波(1977 —)，女，工程师，硕士，主要从事乳制品的研究开发。E-mail：muhaibo@brightdairy.com
* 通信作者：顾小红(1971 —)，女，高级工程师，硕士，主要从事光谱研究及食品快速无损检测。
E-mail：xiao_gu@yahoo.com.cn.com

茶叶进行了聚类分析，结果表明PCA结合马氏距离判据的方法对4种茶叶可进行鉴别；胡乐乾等[5]应用中红外分析技术结合模式识别技术，对掺入糖浆的洋槐和紫云英蜂蜜与真蜂蜜的品质差异进行了模式识别分类研究。结果显示红外光谱结合主成分分析能达到区分真假蜂蜜的效果。

本实验选取脂肪、蛋白质和乳糖含量各不同的商品奶为研究对象，通过ATR-FTIR法采集样品红外谱图，在PCA分析的基础上建立不同类别牛奶的识别模型，以期通过ATR-FTIR结合软独立模式分类法(SIMCA)实现不同类别牛奶的识别。

1 材料与方法

1.1 材料与仪器

实验选用光明乳业股份有限公司、内蒙古蒙牛乳业(集团)股份有限公司、内蒙古伊利实业集团股份有限公司及南京卫岗乳业有限公司生产的市售液态奶为研究对象，根据蛋白质、脂肪和乳糖含量的差异将总样品分为4大类，每类20个共80个样品，分别为纯牛奶(蛋白质2.9～3.1g/100mL、脂肪3.3～3.8g/100mL、乳糖4.8～5.0g/100mL)、低乳糖奶(乳糖0.5g/100mL)、低脂奶(脂肪1.3g/100mL)和高蛋白奶(蛋白质3.2～3.4g/100mL)，并将总样本分成60个训练集样品和20个预测集样品。

Nicolet Nexus 470红外光谱仪(装配水平衰减全反射附件和DTGS检测器) 美国Thermo Electron公司。

1.2 方法

1.2.1 红外光谱采集

采样参数：波数范围4000～675cm^{-1}，扫描次数32次，分辨率4cm^{-1}，样品温度20℃。

采样方法：轻微摇匀样品后取1.0mL均匀分布于样品槽内，以空气为背景，采集红外谱图，每个样品平行采集10次，取平均值作为最终的红外谱图，同理采集去离子水的红外谱图。

1.2.2 分析方法

模式识别是对表征事物或现象各种形式的信息进行处理和分析，是对事物或现象进行描述、辨认、分类和解释的过程，是一种将样本进行聚类的过程[6]。从处理问题的性质和解决问题的方法等角度，模式识别分为有监督的分类(supervised classification)和无监督的分类(unsupervised classification)两种。软独立模式分类法(SIMCA)是一种有监督的模式识别方法，其识别思想是对训练集中每一类已知样本分别进行主成分分析(PCA)并建立数学模型，然后将未知样品与已建立的模型进行拟合，确定未知样品属于哪一类或不属于任何一类[7]。

PCA是SIMCA分析的核心，也是一种数据压缩的常用方法，把原有的各个特征利用线性变化得到一批新的特征，每个特征都是原有特征的函数，但新特征总数少于原有特征数，这样新特征既保留了原有特征的主要信息，又减少了特征个数。在PCA分析基础上建立SIMCA模型后，对未知样品进行预测，利用识别率和拒绝率考察模型的预测效果，其中识别率即是指被考察未知样品落在正确类模型区域内的比率见式(1)，而拒绝率是指被考察类模型对其他不属于该类的未知样品的拒绝程度，即不属于该类的样品落在该类模型区域外的几率，见式(2)[8]。

$$\text{识别率}/\% = \frac{\text{识别同类样本数}}{\text{该类未知样本总数}} \times 100 \quad (1)$$

$$\text{拒绝率}/\% = \frac{\text{识别非同类样本数}}{\text{非该类未知样本总数}} \times 100 \quad (2)$$

1.2.3 数据处理

本实验数据分析基于Unscrambler 9.7和Omnic 8.0软件平台。

2 结果与分析

2.1 红外光谱图的建立及分析

以空气为背景分别采集各类液态奶和水的红外谱图，再利用差减技术得到扣除水后牛奶的红外光谱图[9]。图1为扣除水背景后纯牛奶的红外光谱图，可以看出牛奶的吸收主要集中在3000～2800cm^{-1}和1800～900cm^{-1}两个区域。3000～2800cm^{-1}为CH_3和CH_2的对称和反对称伸缩振动，1745cm^{-1}为脂肪中C＝O键的伸缩振动，1650、1560～1530cm^{-1}和1240cm^{-1}为蛋白质Ⅰ、Ⅱ和Ⅲ类酰胺键的特征吸收区域，1100～900cm^{-1}糖分子中的C－OH和C－O－C的伸缩振动[10]。

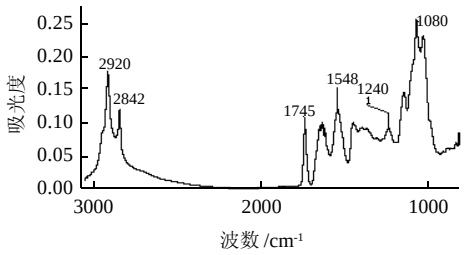

图1 牛奶的红外光谱图

Fig.1 Infrared spectra of milk

2.2 红外光谱分析方法

2.2.1 红外光谱数据的预处理

采集到的原始光谱数据不仅包括样品的信息，还包

括如仪器高频噪音、基线漂移等各种噪音干扰。为了尽量消除干扰信息的影响，在分析之前都要对光谱数据进行预处理[11]。分别采用4种预处理方法：A：原始数据；B：基线校正＋导数处理；C：基线校正＋标准矢量归一化；D：基线校正＋Savitzky-Goaly多项式法平滑。

2.2.2 特征波长的选取

在实际分析过程中，虽然红外光谱数据量越多包含的信息量也越多，但在一些区域光谱信息弱且与成分相关性差，这些信息在建模时反而干扰有效信息的提取[12]。因此建模时有必要对特征波长进行选取，从而提高模型效果。

水在中红外区域有非常强的吸收谱带，在3400cm^{-1}附近有强且宽的O－H的伸缩振动，在600～400cm^{-1}有水分子的摇摆振动，2390～2280cm^{-1}为CO_2反对称伸缩振动区间，这些谱带会干扰和掩盖其他成分的吸收[13]。结合光谱分析结果分别选用全谱Ⅰ(3100～850cm^{-1}，其中2390～2280cm^{-1}区域用直线代替)和自选Ⅱ(3000～2800cm^{-1}＋1800～900cm^{-1})区域建立PCA模型。

经A、B、C、D这4种光谱数据预处理方法，在Ⅰ和Ⅱ光谱区域进行PCA分析并建立SIMCA模型，得到在5%显著性水平下的不同预处理结果见表1。

表1 训练集样品的识别率和拒绝率

Table 1 Recognition rates and rejection rates for training sets with different spectral pretreatments

预处理	牛奶种类	全谱Ⅰ (3100～850cm^{-1})		自选Ⅱ(3000～2800cm^{-1}＋1800～900cm^{-1})	
		识别率/%	拒绝率/%	识别率/%	拒绝率/%
原始数据	纯牛奶	47	87	53	80
	低乳糖奶	100	100	100	100
	低脂奶	100	100	100	100
	高蛋白奶	60	82	40	84
基线校正＋导数处理	纯牛奶	0	71	40	82
	低乳糖奶	100	100	100	100
	低脂奶	100	100	100	100
	高蛋白奶	13	67	53	80
基线校正＋标准矢量归一化	纯牛奶	33	71	53	53
	低乳糖奶	93	89	33	80
	低脂奶	100	78	100	100
	高蛋白奶	27	73	0	73
基线校正＋Savitzky-Goaly多项式法平滑	纯牛奶	53	89	67	82
	低乳糖奶	100	100	100	100
	低脂奶	100	100	100	100
	高蛋白奶	67	84	47	89

由于低乳糖奶和低脂奶两类液态奶在乳糖和脂肪含量上与其他类牛奶存在较大差异，因此更有利于识别，从表1可以看出，通过原始数据建立的模型对于这两类牛奶已有很好的识别效果，识别率达到100%。而市售液态奶在蛋白质含量上变化范围较小，根据分类得到的高蛋白奶和纯牛奶本身差异不明显，因此需要在建模时通过光谱数据预处理、光谱区域的选择等方法来优化模型，从而提高模型的识别效果。以识别效果和拒绝效果为参考依据，纵向比较分析表1中数据发现，基线校正＋Savitzky-Goaly多项式法平滑的光谱数据域处理方法得到的模型在识别率和拒绝率上均高于其他方法，效果最优；横向比较表1中数据，在D光谱数据预处理条件下，Ⅰ和Ⅱ光谱区域对SIMCA模型的识别效果影响不明显，综合考虑模型中高蛋白奶和纯牛奶的识别率和拒绝率，本研究最终选用D(基线校正＋Savitzky-Goaly多项式法平滑)＋全谱Ⅰ(3100～850cm^{-1})对光谱数据进行预处理。

2.3 模型的建立及样品预测

选取3100～850cm^{-1}为特征波长区域，进行基线校正和多项式平滑处理，采用交互留一验证法对样品进行PCA模型并建立SIMCA模型用于预测未知样品，在α＝5%显著性水平下，得到未知样品与模型间距(S_i/S_0 vs H_i)的结果见图2。未知样品与模型间距表示用于预测的未知样品与不同类别牛奶模型的模型中心之间的距离关系，图2中低脂奶模型的中心即图中横纵坐标轴所构成的封闭区域，落在该区域的未知样品即被判定为同一类型，而落在封闭区域以外的未知样品则被判定为不是同类样品。图中4种形状的图案分别代表4类液态奶，可知，低脂奶(三角形图案)全部分布在模型的中心，另外3类牛奶均分布在模型中心以外，说明该模型对低脂奶具有很好的识别效果。

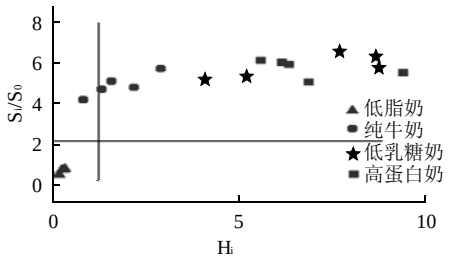

图2 牛奶预测集样品的S_i/S_0 vs H_i(Leverage)

Fig.2 S_i/S_0 vs H_i plots for prediction sets

表2列出了在α＝5%显著性水平下，模型对于各类牛奶样品的识别效果。可以看出模型对于低脂奶的识别效果最好，识别率和拒绝率均达到100%，对于纯牛奶、高蛋白奶和低乳糖奶的识别率和拒绝率均大于80%，说明该模型对于不同品质的牛奶具有较好的识别效果。

表 2 预测集样品的识别率和拒绝率

Table2 Recognition rates and rejection rates for prediction sets

预测集样品	纯牛奶	低乳糖奶	低脂奶	高蛋白奶
识别率 /%	80	80	100	80
拒绝率 /%	93	100	100	93

3 结 论

本实验利用衰减全反射傅里叶红外光谱法(ATR-FTIR)采集牛奶样品的光谱信号，通过光谱数据的预处理，特征波长区域的选择优化建模数据，结合基于 PCA 分析的 SIMCA 模式识别方法对牛奶样品建立分类识别模型，利用模型对未知样品进行识别，模型对于低脂奶的识别率和拒绝率均为 100%，效果良好，对于纯牛奶、高蛋白奶和低乳糖奶的识别率和拒绝率均大于 80%。说明利用红外光谱技术结合软独立模式分类法是进行牛奶分类识别的有效手段。

参考文献：

[1] KAROUI R, DEBAERDEMAEKER J D. A review of the analytical methods coupled with chemometric tools for the determination of the quality and identity of dairy products[J]. Food Chemistry, 2007, 102(3): 621-640.

[2] 刘倩, 孙培艳, 高振会, 等. 衰减全反射傅里叶变换红外光谱技术结合模式识别进行油品鉴别[J]. 光谱学与光谱分析, 2010, 30(3): 663-666.

[3] KAROUI R, BOSSETB J O, MAZEROLLESC G, et al. Monitoring the geographic origin of both experimental French Jura hard cheeses and Swiss Gruyère and L'Etivaz PDO cheeses using mid-infrared and fluorescence spectroscopies: a preliminary investigation[J]. International Dairy Journal, 2005, 15(3): 275-286.

[4] 冯宇, 顾小红, 汤坚, 等. 中红外光谱技术与模式识别相结合鉴别茶叶种类[J]. 食品与生物技术学报, 2007, 26(2): 7-11.

[5] 胡乐乾, 尹春玲, 马渭奎, 等. 红外光谱法对蜂蜜掺伪的模式识别[C]// 中国化学会第十二届全国应用化学年会. 郑州, 2011.

[6] 杜一平, 潘铁英, 张玉兰. 化学计量学应用[M]. 北京: 化学工业出版社, 2008.

[7] 刘树深, 易忠胜. 基础化学计量学[M]. 北京: 科学出版社, 1999.

[8] 张宁, 张德权, 李淑荣, 等. 近红外光谱结合 SIMCA 法溯源羊肉产地的初步研究[J]. 农业工程学报, 2008, 24(12): 309-312.

[9] 翁诗甫. 傅里叶变换红外光谱仪[M]. 北京: 化学工业出版社, 2005.

[10] 常敏. 应用红外光谱技术进行牛奶成分检测的研究[D]. 天津: 天津大学, 2004.

[11] 康继, 顾小红, 汤坚, 等. 中红外反射光谱结合偏最小二乘法快速定量分析葡萄酒[J]. 光谱实验室, 2010, 27(3): 789-796.

[12] 王云. 温度对近红外光谱技术检测牛奶成分影响的研究[D]. 天津: 天津大学, 2006.

[13] IÑÓN F A, GARRIGUE S, de la GUARDIA M, et al. Nutritional parameters of commercially available milk samples by FTIR and chemometric techniques[J]. Analytica Chemica Acta, 2004, 513(2): 401-412.