文章编号：1672-8785(2012)03-0039-05

近红外光谱技术与偏最小二乘法及模糊聚类法相结合的糖品种分类方法

魏俞涌[1]　陈永明[2]　林　萍[2]　何　勇[2]

(1. 嘉兴职业技术学院，浙江 嘉兴 314001；

2. 浙江大学生物系统工程与食品科学学院，浙江 杭州 310058)

摘　要：提出了一种近红外光谱技术与偏最小二乘法及模糊聚类法相结合的可用于快速无损鉴别糖品种的新方法。采用近红外光谱仪获取了白砂糖、木糖醇、麦芽糖和葡萄糖等四种糖类别各 30 个样本的光谱漫反射特征曲线。运用偏最小二乘法提取了糖分类与特征值，并将提取到的经过归一化处理的 11 种主成分结果作为模糊聚类模型的建模参数。设定聚类数为 4，建立模糊聚类模型，并对 40 个未知样本进行了预测。预测结果的准确率达到 100 %，说明本文提出的方法对于糖类别具有很好的分类和鉴别能力，同时也为光谱分析技术在对品种的快速、无损分类与识别中的应用提供了新的思路。

关键词：近红外光谱；糖品种；偏最小二乘；模糊聚类

中图分类号：TP391　　**文献标识码**：A　　**DOI**：10.3969/j.issn.1672-8785.2012.03.008

Discrimination of Varieties of Sugar Based on Partial Least Squares and Fuzzy Clustering Methods

WEI Yu-yong [1], CHEN Yong-ming [2], LIN Ping [2], HE Yong [2]

(1. Jiaxing Vocational and Technical College, Jiaxing 314001, China;

2. College of Biosystems Engineering & Food Science, Zhejiang University, Hangzhou 310058, China)

Abstract: A new method which combines Partial Least Squares (PLS) with a near infrared spectroscopy is proposed for the nondestructive discrimination of the varieties of sugar. A near infrared spectrometer is used to obtain the diffusion spectral characteristic curves from the samples of white granulated sugar, xylitol, maltose and glucose. Then, the PLS is used to derive the variety and characteristic values of the sugar. The derived eleven main components which are normalized are used as the parameters for establishing a fuzzy clustering model. By setting four clusters, the fuzzy clustering model is established and is used to predict forty unknown sugar samples. The prediction accuracy is up to 100 %. This shows that the new method has a good ability to fast discriminate the variety of sugar.

Key words: near infrared spectroscopy; varieties of sugar; partial least squares; fuzzy clustering

收稿日期：2012–01–28

基金项目：“十一五”国家科技支撑项目 (2006BAD10A0403)

作者简介：魏俞涌 (1965-)，男，浙江嵊州人，讲师，主要从事汽车电器教学与科研工作。E-mail: weiyuyons66@sina.com

0 引言

长期以来，糖已经大大改变了人的体形和生活习惯。例如，当今越来越多美国人的身体形态已经不适合汽车消费市场上的汽车室内构型了。为此，美国的一些汽车研发部门正在着手解决这个问题，比如美国福特汽车公司正在考虑设计新的汽车构型，以满足美国当今肥胖购车群体的消费需求。

在借鉴糖对美国人生活的影响的发展历史上，我们必须考虑亚洲 (包括当今经济发展较快的中国) 人对糖的摄入量在将来将会改变人们生活习惯与健康状况的情况。因此，人们根据自身的身体状况对糖及含糖食品的归属类别进行鉴别并科学选择食用不同品种的糖显得尤为重要。

在对糖的品种进行分类测试时，传统的化学方法测试时间长，测试过程繁琐，而且非专业人士一般无法掌握。通过采用基于近红外光谱技术的快速无损鉴别技术 [1−3] 获取市场上出售的白砂糖、木糖醇、麦芽糖和葡萄糖等四种糖的光谱漫反射特征曲线，然后采用偏最小二乘法 (Partial Least Squares, PLS) 提取各类糖的主成分 (Principal Component, PC) [4−5] 并将提取的经过归一化处理的主成分作为模糊 c 均值 (Fuzzy c-means, FCM) 聚类预测模型的建模参数 [6−9]，可以准确地对糖的品种进行分类和预测。由此可见，通过将近红外光谱技术的快速检测特性与相应的数学建模方法相结合，我们就能开发出相应的计算机产品，以满足海关、医疗药物机构、超市甚至老百姓对糖品种进行快速鉴别的需求。

1 材料与方法

1.1 仪器设备

实验使用美国 ASD 公司生产的 Handheld FieldSpec 光谱仪，其光谱采样间隔为 1.5 nm，测定范围在 325 ∼ 1075 nm 之间，扫描次数为 30 次，探头的视场角为 20° 。光源采用与光谱仪相配套的 14.5 V 卤素灯。得到的光谱数据经 ASD View Spec Pro 软件转化为 ASCII 码形式，再由 Unscrambler V9.7 和 MATLAB V7.1 软件进行分析处理。

1.2 样品来源及光谱的获取

我们从超市买来太古纯正白砂糖 (Saccharose)、禾甘木糖醇 (Xylitol)、双歧麦芽糖 (Bifid Sugar) 和红芍口服葡萄糖 (Dextrose) 等四种糖。各取 4 包样本，每包 400 g 。各种糖样本均用直径为 120 mm 、高度为 10 mm 的培养皿盛装。为了减小实验误差并保证被测物体与仪器等距，我们将每个培养皿装满后作为一个实验样本。每个品种各做 40 个样本，共计 160 个样本。将全部实验样本随机分成建模集和预测集。其中，建模集有 120 个样本 (每个品种各 30 个)，预测集有 40 个样本 (每个品种各 10 个) 。光谱仪经白板校准后进行测试。光谱仪置于样本的上方，距糖表面 120 mm 。对每一个样本扫描 30 次。

1.3 光谱数据的预处理

为了消除来自高频随机噪声、基线漂移、样本不均匀以及光散射等的影响，我们需要进行光谱预处理以消除噪声。先对数据进行 Savitzky Golay Derivatives 处理，再采用 Savitzky-Golay 平滑法，选用的平滑点数为 9，此时能够很好地滤除各种因素所产生的高频噪声。最后对数据进行标准正态变量 (Standard Normal Variate, SNV) 处理。由于光谱曲线在首端和末端存在较大噪声，我们只选取 400 ∼ 1000 nm 波段的光谱进行分析。

1.4 偏最小二乘法

偏最小二乘法 (Partial Least Squares, PLS) 是一种很有效的多元统计方法。它能够建立光谱数据与成分之间的相互关系，适合于光谱分析中的线性模型，是一种已被广泛使用的近红外光谱数据处理方法。但是当因变量和自变量不完全呈线性关系时，采用线性处理可能会给分析结果带来一定的偏差，因此需要在 PLS 模型的基础上引入非线性部分，即模糊聚类算法。

1.5 模糊聚类算法

传统的概率密度算法需要使用概率密度函数，为此要假设合适的模型，而且不易处理聚类不是致密而是壳形的情形。模糊聚类算法可以

摆脱这类限制，使得模糊算法中的向量可以属于多个聚类。本文采用 FCM 算法对 PLS 产生的 PC 值进行预测。下面介绍一下 FCM 算法。

FCM 目标函数 $J_m(U,V)$ 为

$$J_m(U,V)=\sum_{i=1}^{n}\sum_{j=1}^{c}u_{ij}^{m}\cdot\|X_i-V_j\|^2 \qquad (1)$$

式中，$\|X_i-V_j\|=[\sum_{k=1}^{s}(x_{ki}-V_{kj})^2]^{1/2}$ 为样本 $X_i=\{x_{kj}\}$ 与聚类中心 $V_j=\{V_{kj}\}$ 之间的欧式距离，其中 $i=1,2,\cdots,n$，n 为样品数；$j=1,2,\cdots,c$，c 为分类数；$k=1,2,\cdots,s$，s 为特征变量数；u_{ij} 为样本 X_i 对第 j 类的隶属度；m 为加权指数，且 $m>1$ (为了加强 X_i 属于各类程度的对比度)；m 的值越大，所得分类矩阵的模糊程度就越大 (一般 m 取 1.1 ~ 2.0) 。目标函数 $J_m(U,V)$ 表示样本 x_i 与各个聚类中心 V_j 的带权距离平方和，其权重为样本 x_i 隶属类 C_j 的隶属度 u_{ij} 的 m 次方，而最佳聚类是使目标函数 $J_m(U,V)$ 最小。因此，若要得到最佳聚类结果，则要求得到适当的隶属度 u_{ij} 和聚类中心 V_j 。当 $m>1$ 且 $X_i\neq V_j$ 时，可以用式 (2) 和式 (3) 迭代计算出隶属度 u_{ij} 和聚类中心 V_j [7−9] 。

$$u_{ij}=\left[\sum_{l=1}^{c}\left(\frac{\|x_i-v_j\|}{\|x_i-v_l\|}\right)^{\frac{2}{m-1}}\right]^{-1} \qquad (2)$$

$$v_j=\sum_{p=1}^{n}[(u_{pj})^m\cdot x_p]\Big/\sum_{p=1}^{n}(u_{pi})^m \qquad (3)$$

具体算法如下：

(1) 固定分数 c，加权指数为 m；收敛门限为 ε；选取初始隶属度矩阵 $U^{(0)}$，其元素 u_{ij} 满足：

$$0\le u_{ij}\le 1,\quad \forall i,j;\quad \sum_{j=1}^{c}u_{ij}=1,\quad \forall i$$

(2) 根据式 (3) 和 $U^{(q)}$ 求出聚类中心 $V_j^{(q)}$，其中 q 为迭代次数。

(3) 根据式 (2) 和求得的 $V_j^{(q)}$，求出 $U^{(q+1)}$ 。

(4) 若 $\max\{|U^{(q)}-U^{(q+1)}|\}\le\varepsilon$，则停止迭代，$U^{(q+1)}$ 及相应的 $V_j^{(q)}$ 为所求结果；否则返回步骤 (2)，继续迭代。

(5) 在得到的隶属度矩阵 U 中，令每列中的最大元素为 1，其余为 0，从而得到一个普通分类矩阵 (即分类结果) 。

2 试验结果与分析

2.1 不同品牌的糖的聚类分析

对四种糖 (白砂糖、木糖醇、麦芽糖和葡萄糖) 的共 120 个样本进行主成分分析。表 1 列出了前三个主成分的特征值及累计可信度 。由于前两个主成分的可信度已达 97.7 %，我们仅用前两个主成分就可以表示原近红外光谱的主要信息。

表 1 前三个主成分的累计可信度

主成分	PC1	PC2	PC3
累计可信度	58.5 %	97.7 %	99.5 %

图 1 为 120 个建模样本的主成分 1 、 2 的得分图。其中，横坐标表示每个样本的第一主成分的得分值，纵坐标表示每个样本的第二主成分的得分值。从图 1 中可以看出，白砂糖、木糖醇、麦芽糖和葡萄糖已经明显分成四类：木糖醇的 30 个样本聚合在第一象限，双歧糖的 30 个样本聚合在第二象限，白砂糖的 30 个样本聚合在第三象限，葡萄糖的 30 个样本聚合在第四象限。这四种糖中除了白砂糖有个别样本稍微有所偏离之外，其余三种糖的聚类性都很好，说明主成分 1 、 2 对四种糖有较好的聚类作用。

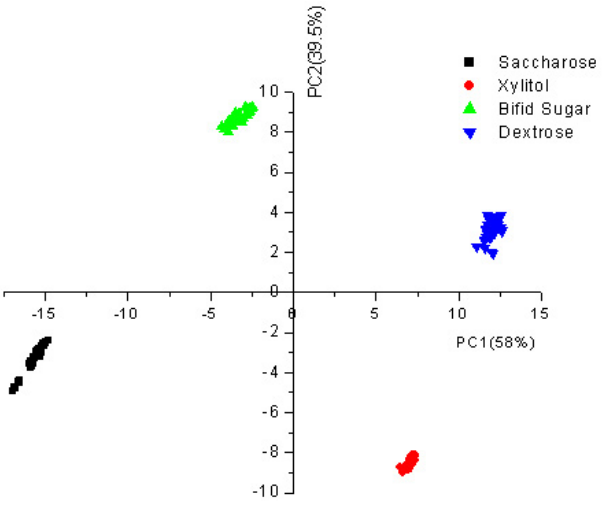


图 1　120 个样本的主成分 1 和主成分 2 的得分图

2.2 PLS 成分的提取和模型参数的确定

采用 PLS 分析方法并经交互验证法判断得到，最佳主成分数为 11 。表 2 列出了主要的模型参数。选用 11 个主成分模型的残差 (RV) 为 0.000985，相关系数 (R) 为 0.999916；同时，其校正集标准偏差 (SEC) 为 0.014538 。由于是对糖的种类进行鉴别，用 1 、 2 、 3 、 4 分别表示四种不同品种的糖，并将其作为 Y 变量。

从原有变量中提取新变量，即对原有变量进行线性组合来构成新变量。提取的新变量尽可能保留原有变量的有用信息，尽可能减少变量数，变量之间相互独立，这就是特征变量。PLS 成分是从原有自变量的样本数据矩阵 X 中提取的相互正交的成分。它们既保留了与因变量的相关性，又保留了较多的方差，从而在消除原有自变量共线性的同时，使建立的回归模型仍能充分地反映出自变量与因变量之间的相互关系。

表 2 PLS 模型中的主要参数

主成分	残差 (RV)	相关系数 (R)	校正集标准偏差 (SEC)	主成分	残差 (RV)	相关系数 (R)	校正集标准偏差 (SEC)
1	0.531000	0.764739	0.722656	7	0.001357	0.999755	0.024828
2	0.030750	0.988309	0.170998	8	0.001339	0.999840	0.020029
3	0.006359	0.997587	0.077864	9	0.001115	0.999867	0.018314
4	0.002901	0.998960	0.051141	10	0.001015	0.999894	0.016315
5	0.002034	0.999320	0.041357	11	0.000985	0.999916	0.014538
6	0.001659	0.999575	0.032702				

2.3 模糊聚类

光谱波段为 400 ~ 1000 nm 。如果将其全部作为模糊输入 FCM 算法的输入端，那么无疑将会大大增加其计算量。而且样品在有些区域中的光谱信息很弱，与样品的组成或性质之间缺乏一定的相关性。而用 PLS 主成分分析得出的前面 11 个主成分已经包含了大部分光谱信息。因此，我们对这 11 个特征变量进行了归一化处理，并将处理结果作为 FCM 算法的建模参数。我们分别将建模和预测样本数选择为 120 个和 40 个 (见表 3) 。设初始化聚类数为 4 ，当迭代次数达到 12 次时，目标函数值为 315.448349 ，完成预测。图 2 所示为聚类过程；表 4 列出了预测结果。由此可见，预测结果的正确率达到 100 % 。

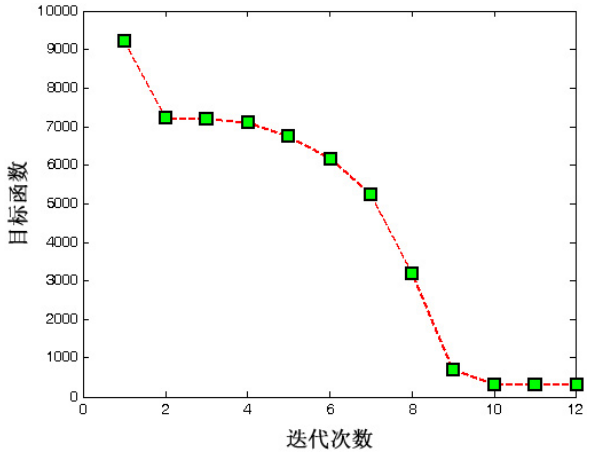


图 2 模糊聚类的过程图

表 3 模糊聚类建模与预测样本

编号	建模、预测样本	预测因子
1	白砂糖 (编号 1-30)	PC1-PC11
2	木糖醇 (编号 31-60)	PC1-PC11
3	麦芽糖 (编号 61-90)	PC1-PC11
4	葡萄糖 (编号 91-120)	PC1-PC11
5	未知样本 (编号 121-160)	PC1-PC11

表 4 模糊聚类结果

编号	建模样本	预测样本编号
1	白砂糖 (编号 1-30)	121-130
2	木糖醇 (编号 31-60)	131-140
3	麦芽糖 (编号 61-90)	141-150
4	葡萄糖 (编号 91-120)	151-160

3 结论

采用近红外光谱技术获得了四种糖的光谱特征曲线，然后利用取得的光谱信息结合 PLS 分析法获得了识别模型的输入特征置信区间，并将得到的特征变量作为模糊聚类建模参数，成功地对糖品种进行了分类和识别。本文建立的 PLS/ 模糊聚类方法为光谱分析技术在对品种的快速、无损分类与识别中的应用提供了一种新的思路。若将本方法结合计算机硬件技术开发出相应的产品，则可为海关、医药部门、超市甚至老百姓提供行之有效的糖食品快速分类与识别方法，因而具有重要的实际意义。

参考文献

[1] Cen Haiyan, He Yong. Theory and Application of Near Infrared Reflectance Spectroscopy in Determination of Food Quality [J]. *Trends in Food Science & Technology*, 2007, **18**(2): 72–83.

[2] He Yong, Li Xiaoli, Deng Xunfei. Discrimination of Varieties of Tea Using Near Infrared Spectroscopy by Principal Component Analysis and BP Model [J]. 2007, **79**(4): 1238–1242.

[3] 黄敏，何勇，黄凌霞，等. 基于可见－近红外光谱技术的家蚕蚕种鉴别方法的研究 [J]. **红外与毫米波学报**, 2006, **25**(5): 1421–1423.

[4] 李剑，陈德钊，成忠，等. 构建支持向量机－偏最小二乘法为药物构效关系建模 [J]. **分析化学**, 2006, **34**(2): 263–266.

[5] 吴晓华，陈德钊. 化学计量学非线性偏最小二乘算法进展评述 [J]. **分析化学**, 2004, **34**(2): 534–540.

[6] 楼世博，孙章. **模糊数学** [M]. 北京：科技出版社，1987.

[7] 李晶皎，王爱侠，张广渊，等译. **模式识别** [M]. 北京：电子工业出版社， 2006.

[8] 褚小立，袁洪福，陆婉珍. 光谱结合主成分分析和模糊聚类方法的样品聚类与识别 [J]. **分析化学**, 2000, **28**(4): 421–427.

[9] 邓勇，刘琪，李亦学. 基于氨基酸模糊聚类分析的跨膜区域预测 [J]. **化学学报**, 2004, **62**(19): 1968–1972.

新闻动态 News

美国汽车称重站利用红外成像技术捕捉有危险卡车

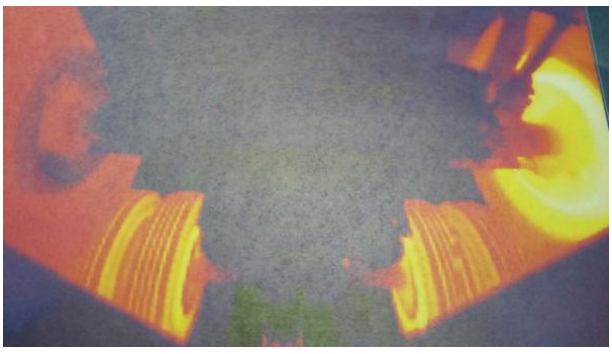
据 www.msnbc.msn.com 网站报道，每天有数万辆商业汽车沿美国华盛顿的道路而过。但是经常从你身边通过的这些大卡车是否安全呢？现在，美国华盛顿州的巡逻队有了一种有助于回答这个问题的新手段。

该州的警察打算在地面上放一台热成像摄像机，以帮助他们精确查找经过汽车称重站的卡车所存在的问题。迄今为止，该州的巡逻队和运输部门已对这项技术进行了试验，而且该州政府机构对试验结果留下了深刻印象。

现在，该州的巡逻队是在汽车称重站对车辆进行检测的，但检测有时候是随机进行的。安全焦点包括轮胎看上去是否好，刹车是否可以工作。检测可能要花费一个小时的时间。

如果安装了热成像摄像机，半挂车底部的图片便可以在计算机上显示出来。巡逻队员看一下该图像就可以精确地查出刹车可能已坏的卡车。通过采用这项技术，可以使检测工作变得更加有效。

该州的巡逻队员说，他们的时间可以更好地花费在找出具有缺陷的卡车方面，所以这是一个非常激动人心的手段。他们计划在 2012 年 3 月选择几个汽车称重站使用这种红外成像摄像机。



□ 高国龙