

基于可见近红外光谱的糖类别快速鉴别研究

林 萍, 陈永明, 何 勇*

浙江大学生物工程与食品科学学院, 浙江 杭州 310029

摘 要 提出了一种用近红外光谱技术快速鉴别糖类别的新方法。采用近红外光谱获取白砂糖、木糖醇、双歧糖和葡萄糖等四种糖类别的光谱反射特征曲线, 采用偏最小二乘法进行模式特征分析, 经过交互验证法判别, 确定最佳主成分数为 11。完成特征提取后, 将 11 种主成分作为神经网络的输入变量, 建立了 3 层 BP 神经网络。四个类别的糖样本数均为 40, 共计 160 个样本, 将其分成训练集样本 120 个和预测集样本 40 个, 对 40 个未知样本进行预测, 准确率为 100 %。说明所提出的方法对于糖类别具有很好的分类和鉴别能力。

关键词 偏最小二乘法; BP 神经网络; 糖; 光谱技术

中图分类号: S566 **文献标识码:** A **DOI:** 10. 3964/j. issn. 1000-0593(2009)02-0382-04

引 言

现在市场上供应糖的品种越来越多, 功能也各异, 消费者应该根据自身的需要选择适合自己的糖。市场上的糖大致有四种: 白砂糖、木糖醇、双歧糖和葡萄糖。

白砂糖即白糖, 蔗糖含量一般在 95 % 以上, 是饮料、食品和制药工业不可少的重要原料。适当食用白糖有助于提高机体对钙的吸收, 但过多就会妨碍钙的吸收。糖尿病患者不宜直接食用食糖, 最好是以甜味剂替代。

葡萄糖是生物体内新陈代谢不可缺少的营养物质, 它的氧化反应放出的热量是人类生命活动所需能量的重要来源。在消化道中, 葡萄糖比任何其他单糖都容易被吸收, 而且被吸收后能直接为人体组织利用。

木糖醇为白色晶体, 外表和蔗糖相似, 是多元醇中最甜的甜味剂, 甜度相当于蔗糖, 热量相当于葡萄糖。木糖醇仅仅能被缓慢吸收或部分被利用, 热量低是它的一大特点。木糖醇是人体糖类代谢的中间体, 无需胰岛素促进, 木糖醇也能透过细胞膜, 被组织吸收利用, 供细胞以营养和能量, 且不会引起血糖值升高, 是适合糖尿病患者食用的营养性的食糖代替品。

双歧糖是以异麦芽低聚糖和果糖低聚糖为基料研制生产的一种甜度相当于蔗糖 3 ~ 4 倍, 热量只有蔗糖 1/6 的新型糖。因其甜度高、热量低, 不易为肠胃吸收, 所以没有蔗糖容易引起肥胖和龋齿的副作用, 是喜欢甜食又关心健康者、以及忌食蔗糖者的理想糖源。由于其热量低, 对血糖影响很

小, 基本不依赖胰岛素, 糖尿病人使用更安全。

本文提出了一种快速简便无损鉴别糖类的可见-近红外光谱法。国内外很多学者利用近红外光谱技术进行物质品种鉴别, 如苹果品种^[1]、中药材^[2]、酸奶品种^[3]、奶粉品种^[4]、杨梅品种^[5]、脸谱识别^[6]等。但在糖的品种鉴别方面, 还鲜有报道。

偏最小二乘法^[7]是一种目前应用较广泛的光谱分析的数学方法。可以实现全谱或部分谱数据的分析; 数据矩阵分解和回归交互结合为一步, 得到的特征值向量直接与被测组分或性质相关; 对于样本数量少, 变量多的情况下更具有优越性。人工神经网络是模仿人脑神经细胞的结构和功能的系统^[8]。目前应用较广的是一种基于误差反向传播算法的 BP 神经网络。由于具有高度非线性映射的能力, 现已在各领域得到了广泛的应用, 取得了令人满意的结果。本文用基于偏最小二乘法与 BP 神经网络相结合的识别方法, 研究用近红外光谱技术对糖进行种类鉴别, 并对 40 个未知样本进行预测, 准确率为 100 %, 40 个样本的预测相对偏差均在 5 % 以下, 说明本文提出的方法对于糖具有很好的分类和鉴别能力。

1 材料与方法

1.1 仪器设备

美国 ASD (analytical spectral device) 公司的 Handheld FieldSpec 光谱仪, 光谱采样间隔为 1.5 nm, 测定范围在 325 ~ 1075 nm 之间, 扫描次数 30 次, 探头视场角为 20°。光源

收稿日期: 2007-10-06, 修订日期: 2008-01-08

基金项目: 国家科技支撑计划项目(2006BAD10A0403)和宁波市农业攻关国际合作项目(2008C10037)资助

作者简介: 林 萍, 女, 1982 年生, 浙江大学生物系统工程与食品科学学院博士研究生 *通讯联系人 e-mail: yhe@zju.edu.cn

采用与光谱仪配套的 14.5 V 卤素灯。得到的光谱数据经 ASD View Spec Pro 软件转化为 ASCLL 码形式,再由分析软件 Unscrambler V9.7 和 DPS^[9]对数据进行分析处理。

1.2 样品来源及光谱的获取

从超市买来太古纯正白砂糖、禾甘木糖醇、双歧糖(无糖型)、红芍口服葡萄糖。各取 4 包样本,每包 400 g。各种糖样本均用直径为 120 mm,高度 10 mm 的培养皿装满作为一个实验样本。每个种类各做 40 个样本,共计 160 个样本。全部实验样本随机分成建模集和预测集,建模集有 120 个样本(每个种类 30 个),预测集有 40 个样本(每个种类 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 光谱数据预处理

为了去除来自高频随机噪声、基线漂移、样本不均匀、光散射等影响,需要进行光谱预处理来消除噪声。先对数据进行 Savitzky Golay Derivatives 处理,再采用 Savitzky Golay 平滑法,选用平滑点数为 9,此时能很好滤除各种因素产生的高频噪音。最后对数据进行 SNV 处理。由于光谱曲线在首端和末端有较大噪音,如图 1 所示,所以只取 400~1 000 nm 波段的光谱用于分析。

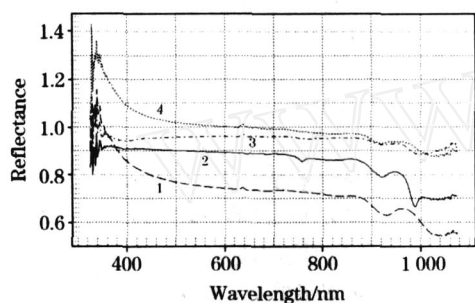


Fig 1 Near infrared reflectance spectra of four different varieties sugar

1: 木糖醇; 2: 白砂糖; 3: 双歧糖; 4: 葡萄糖

1.4 偏最小二乘法(PLS)

偏最小二乘法是一种很有效的多元统计方法,能够建立光谱数据和成分之间的相互关系,适合于光谱分析中的线性模型,是一种广泛使用的近红外光谱数据处理方法。但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的部分,人工神经网络是目前常用的非线性模型。

1.5 BP 神经网络模型

本研究建立了一个 3 层 BP 神经网络。所有的样本随机地分成学习集和预测集数据,其中 120 个样本为学习集样本,40 个样本为预测集样本。网络输入层、隐含层、输出层节点数分别为 11, 8, 1,其中输入层的 11 个节点来自 PLS 分析得到的主成分。最小训练速率为 0.1,设定训练迭代次数为 2 000 次,对输入样本进行标准化处理。

2 试验结果与分析

2.1 不同种类糖的近红外光谱图分析

四种糖的典型近红外光谱如图 1 所示。从图中可以看出,不同种类糖的光谱曲线有明显区别,并具有一定的特征性和指纹性,这一差异为糖的不同种类鉴别奠定了数学基础。选择波长范围在 400~1 000 nm 的光谱,应用 ASD View Spec Pro 软件,把糖的光谱曲线做平均处理,并转换成 ASCII 码,形成反射率矩阵,用偏最小二乘法对其聚类。

2.2 不同品牌糖的聚类分析

对白砂糖、木糖醇、双歧糖和葡萄糖各 30 个,共 120 个样本进行主成分分析。前 3 个主成分的特征值及累计可信度如表 1 所示。由于前 2 个主成分的可信度已达 97.7%,故仅用前 2 个主成分就可表示近红外光谱的主要信息。

Table 1 Prins and reliabilities

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

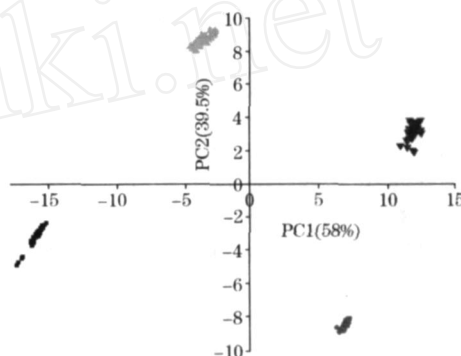


Fig 2 2D scores plot for the prin1 and prin2 of sugar samples

图 2 表示 120 个建模样本的主成分 1 和 2 得分图,横坐标表示每个样本的第一主成分得分值,纵坐标表示每个样本的第二主成分得分值。从图中可以看出,白砂糖(Saccharose)、木糖醇(Xylitol)、双歧糖(Bifid sugar)、葡萄糖(Glucose)四种糖明显分成 4 类,葡萄糖的 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,木糖醇的 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,其余三种糖的聚类性都很好。说明主成分 1 和 2 对四种糖有较好的聚类作用。

2.3 模型参数的确定和 PLS 成分的提取

采用 PLS 分析方法,经交互验证法判断,最佳主成分数为 11,主要的模型参数见表 2。选用 11 个主成分模型其残差(R_v)0.000 985,相关系数(r)0.999 916,同时,它的校正集标准偏差(SEC)为 0.014 538。由于是对糖的种类进行鉴别,用 1, 2, 3, 4 分别代表四种不同品种的糖,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构成新变量。提取的新变量保留了尽可能多的原有变量的有用信息,并且变量数尽可能地少,变量之间相互独立,这就

Table 2 Main parameters in PLS model

PLS	R_V	r	SEC
1	0.531 000	0.764 739	0.722 656
2	0.030 750	0.988 309	0.170 998
3	0.006 359	0.997 587	0.077 864
4	0.002 901	0.998 960	0.051 141
5	0.002 034	0.999 320	0.041 357
6	0.001 659	0.999 575	0.032 702
7	0.001 357	0.999 755	0.024 828
8	0.001 339	0.999 840	0.020 029
9	0.001 115	0.999 867	0.018 314
10	0.001 015	0.999 894	0.016 315
11	0.000 985	0.999 916	0.014 538

Table 3 Prediction results for unknown samples by BP model

样本号	预测值	真实值	RSD/ %	样本号	预测值	真实值	RSD/ %
(1)	1.013 7	1	1.37	(21)	2.972 2	3	0.92
(2)	1.010 3	1	1.03	(22)	2.955 0	3	1.50
(3)	1.007 7	1	0.77	(23)	2.958 5	3	1.38
(4)	1.020 8	1	2.08	(24)	2.969 3	3	1.02
(5)	1.007 2	1	0.72	(25)	2.943 8	3	1.87
(6)	1.007 4	1	0.74	(26)	2.969	3	1.03
(7)	1.018 1	1	1.81	(27)	2.910 4	3	2.98
(8)	1.028 6	1	2.86	(28)	2.950 2	3	1.66
(9)	1.046 3	1	4.63	(29)	2.954 2	3	1.52
(10)	1.012 2	1	1.22	(30)	2.916 7	3	2.77
(11)	1.978 8	2	1.06	(31)	3.888 9	4	2.77
(12)	1.993 5	2	0.32	(32)	3.990 7	4	0.23
(13)	1.983 5	2	0.82	(33)	3.971	4	0.72
(14)	1.986 2	2	0.69	(34)	3.989 2	4	0.27
(15)	1.972 4	2	1.38	(35)	3.990 9	4	0.22
(16)	2.010 8	2	0.54	(36)	3.989	4	0.27
(17)	1.997 4	2	0.13	(37)	3.987	4	0.32
(18)	2.002 7	2	0.13	(38)	3.984 6	4	0.38
(19)	1.986 8	2	0.66	(39)	3.987 6	4	0.31
(20)	1.998 3	2	0.08	(40)	3.990 5	4	0.23

1: Saccharose; 2: Xylitol; 3: Bifid sugar; 4: Glucose

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,他们既保留尽量多与因变量的相关性,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。从表 2 中可看出, Y 变量与 PC11 的相关性最大。

2.4 基于特征变量建立 BP 神经网络模型

光谱波段是 400 ~ 1 000 nm, 如果全部将其作为神经网络的输入,大大加大了神经网络的计算量,影响学习速率。而且有些区域样品的光谱信息很弱,与样品的组成或性质间缺乏一定的相关性。而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,将这 11 个特征变量作为 BP 神经网络的输入变量,即网络输入层、隐含层、输出层节点数分别为 11, 8, 1。训练集和预测集样本数为 120 和 40。对 40 个未知样本的预测结果见表 3, 预测样本 1 至 10 号真实值为 1, 预测样本 11 至 20 号真实值为 2, 预测样本 21 至 30 号真实值为 3, 预测样本 31 至 40 号真实值为 4, 依次代表白砂糖、木糖醇、双歧糖、葡萄糖这 4 个种类。拟合残差为 0.000 023, 40 个样本的预测相对偏差均在 5 % 以下,即预测的正确率达到了 100 %。

3 结 论^[10]

应用近红外光谱技术建立了糖类别鉴别的模型,该模型的预测效果很好,对未知样品的预测相对误差均在 5 % 以下,识别率达到 100 %。说明运用近红外光谱可以快速、准确对糖类别进行鉴别。采用偏最小二乘法和 BP 神经网络相结合的方法,大大提高了识别的精确度。将从 PLS 分析中得到的特征变量作为 BP 神经网络的输入,不但减少了神经网络的计算量,加快了训练速率,而且因为去除了光谱干扰信息,也提高了预测的正确率。因此,用偏最小二乘法结合 BP 神经网络的模式识别和近红外光谱技术研究糖类别鉴别是可行的。

参 考 文 献

- [1] HE Yong, LI Xiao-li, SHAO Yong-ni. International Journal of Food Properties, 2007, 10(1): 9.
- [2] LIU Shu-hua, ZHANG Xue-gong, ZHOU Qun, et al (刘沐华, 张学工, 周 群, 等). Spectroscopy and Spectral Analysis (光谱学与光谱分析), 2005, 25(6): 878.
- [3] HE Yong, FENG Shui-juan, LI Xiao-li (何 勇, 冯水娟, 李晓丽). Spectroscopy and Spectral Analysis (光谱学与光谱分析), 2006, 26(11): 2021.
- [4] HUANG Min, HE Yong, CEN Hai-yan, et al (黄 敏, 何 勇, 岑海燕, 等). Spectroscopy and Spectral Analysis (光谱学与光谱分析), 2007, 27(5): 916.
- [5] CEN Hai-yan, BAO Yi-dan, HE Yong (岑海燕, 鲍一丹, 何 勇). Spectroscopy and Spectral Analysis (光谱学与光谱分析), 2007, 27(3): 503.
- [6] Cheng, J, Liu Q S, Lu H Q, et al. Pattern Recognition, 2006, 39(1): 81.
- [7] He Yong, Li Xiaoli, Deng Xunfei. Journal of Food Engineering, 2007, 79(4): 1238.

- [8] BAO Yi-dan , WU Yan-ping , HE Yong (鲍一丹 , 吴燕萍 , 何 勇) . Journal of Agricultural Mechanization Research (农机化研究) , 2004 , (3) : 162 .
- [9] TANG Qi-yi , FENG Ming-guang (唐启义 , 冯明光) . DPS Data Processing System for Practical Statistics (实用统计分析及其 DPS 数据处理系统) . Beijing : Science Press (北京 : 科学出版社) , 2002 .
- [10] LU Wan-zhen , YUAN Hong-fu , XU Guang-tong , et al (陆婉珍 , 袁洪福 , 许广通 , 等) . Modern Near-Infrared Spectroscopy Analysis Technology (现代近红外光谱分析技术) , Beijing : Chinese Petroleum Chemical Industry Press (北京 : 中国石化出版社) , 2000 . 167 .

Fast Discrimination of Varieties of Sugar Based on Spectroscopy Technology

LIN Ping , CHEN Yong-ming , HE Yong *

College of Biosystems Engineering and Food Science , Zhejiang University , Hangzhou 310029 , China

Abstract Visible and near-infrared reflectance spectroscopy (NIRS) was applied in the discrimination of sugar varieties. NIRS is a pollution-free , rapid , quantitative and qualitative analysis method , with the characteristics of high speed , non-destructiveness , high precision and reliable detection data , etc. Four kinds of sugar were gained from the local market and each species was divided into 40 samples. One hundred twenty samples were used as the training set and the remainders (total 40 samples) formed the prediction set. Samples were scanned by a spectroradiometer within a wavelength region of 325-1 075 nm. Three pre-processing methods were applied on the spectra prior to building the PLS regression model. The multivariable analysis using partial least square (PLS) was applied to abstract characteristics of the pattern. Through full cross validation , 11 principal components presenting important information of spectra were confirmed. The correlation coefficient (R) , residual variance (R_v) and standard error of calibration (SEC) were 0.999 916 , 0.000 985 and 0.014 538 respectively. Then , these 11 principal components were taken as the input of BP neural network. This model was used to predict the varieties of 40 unknown samples. Through training and prediction , the recognition rate of 100 % was achieved by BP neural network. This model has come to be reliable and practicable. Thus , it is concluded that PLS analysis combined with BP neural network is an available alternative for pattern recognition based on the spectroscopy technology.

Keywords Partial least square ; BP neural network ; Sugar ; Spectroscopy technology

(Received Oct. 6 , 2007 ; accepted Jan. 8 , 2008)

* Corresponding author