

文章编号:1673-1689(2008)03-0099-05

原料乳中大肠菌群快速预测模型的建立

刘建学¹, 钟莉娟², 钟先锋¹, 李佩艳¹, 徐宝成¹, 易军鹏¹, 张晓慧¹

(1. 河南科技大学 食品与生物工程学院, 河南 洛阳 471003; 2. 信阳师范学院 物理电子工程学院, 河南 信阳 464000)

摘要:采用基于余弦相似度因子分析结合偏最小二乘回归分析的近红外光谱方法,研究了原料乳中大肠菌群的测定原理与测试方法,建立了预测数学模型,并与逐步回归分析、偏最小二乘回归分析等方法进行了测试结果的对比分析。结果表明,该分析方法所得结果与微生物生化测定方法所得结果的相关系数 R^2 为 0.954 3;通过模型显著性 F 检验,在置信度 $\alpha=0.05$ 的情况下,该预测模型显著。

关键词:近红外光谱;大肠菌群;余弦相似度;偏最小二乘分析;原料乳

中图分类号:O 433.4; TS 201.3

文献标识码:A

A Rapid Quantitative Assay Model of Coli Group in Raw Milk

LIU Jian-xue¹, ZHONG Li-juan², ZHONG Xian-feng¹,

LI Pei-yan¹, XU Bao-cheng¹, YI Jun-peng¹, ZHANG Xiao-hui¹

(1. College of Food and Bioengineering, Henan University of Science and Technology, Luoyang 471003, China; 2. Department of Physics and Electronic Engineering, Xinyang Normal University, Xinyang 464000, China)

Abstract: A rapid, non-damaging assay, based on the reflectance of transmittance spectroscopy approaches with near infrared spectroscopy (NIRS), was described for quantifying the number of coli group in raw milk. The quantitative prediction model on dynamic cosine similarity clustering and partial least square was established according to the spectra of raw milk sample. For coli group, the correlation efficient (R^2) between the evaluation of NIRS and biochemical method was 0.9543. The model was also compared with partial least squares and step-by-step regression analysis. The result showed that the approach was effective when the confidence was 0.05.

Key words: near infrared spectroscopy; *Escherichia coli*; cosine similarity clustering; partial least square analysis; raw milk

乳品在加工生产过程中很容易被污染或滋生一些有害微生物,这对于原料乳的质量、成品乳的出厂都会产生很大影响。目前,乳品生产厂家、卫

生防疫部门、质量技术监督部门在测定乳品微生物学指标(如细菌总数、大肠菌群总数等)时是采用先取样,再培养,然后通过显微镜人工计数推算的方

收稿日期:2007-03-18.

基金项目:河南省高校青年骨干教师基金项目(09003016).

作者简介:刘建学(1964-),男,河南杞县人,工学博士,教授.主要从事农产品加工及贮藏工程方面的研究. Email: jx_liu@163.com

法,这样测一个样品需要2 d才能完成,即使采用现有的快速检测方法,也需要4 h或6 h以上,所需检测时间长,费工费力,而且检测结果远远滞后于生产需要。因此,乳品生产厂家即使有较大生产能力,由于成品乳不能及时出厂,也不敢扩大生产规模,这种状况在某种程度上严重制约了乳品生产行业的发展。

傅里叶变换近红外光谱分析技术(FTNIR)是一种可以对物质成分进行定量分析或甄别的快速分析方法。它是利用红外辐射与分子振动或转动的相互作用,通过记录试样的近红外吸收光谱进行定性、定量和结构分析的方法。其最显著特点是不破坏所测对象,检测时间短且无污染^[1]。该技术不仅越来越广泛地应用于蛋白质、核酸等生物大分子结构研究上,而且已开始用来更深入地研究细胞和组织等更加复杂的体系^[2-7]。所以,迫切需要寻找一种对乳品中有害菌进行快速、方便的检验方法,并对其加以控制。作者利用大肠菌群的近红外光谱,采用傅立叶变换近红外光谱分析方法对其光谱特征进行分析,以达到快速检测乳品及确保乳品卫生安全的目的。

1 材料与方法

1.1 试验材料与仪器

试验材料采用洛阳巨尔乳业有限公司提供的原料乳样品,大肠菌群生化分析所用培养基及相关试剂全部使用分析纯级。主要试验仪器为德国BRUKER公司生产VECTOR-33型傅里叶变换近红外光谱分析仪及广州光学仪器厂生产显微镜。

1.2 试验方法

1.2.1 生化试验 原料乳中大肠菌群数每100 mL检样内大肠菌群最可能数(MPN)表示,其测定方法参照文献^[8]的方法进行。

1.2.2 光谱扫描 测试样品时,先打开傅里叶变换近红外光谱仪,预热20 min。待仪器工作正常后,设定仪器各参数。光谱仪的扫描精度为 4 cm^{-1} ,扫描次数32次,扫描光谱范围为 $3\,000\sim 9\,000\text{ cm}^{-1}$ 。在参数设置菜单中选近红外光谱模式,积分球检测光路,漫反射方式。按照上述设定的参数作背景扫描,取得背景扫描光谱,存入计算机。取待测试样30 mL,放入已经灭菌、干燥过的50 mL的小烧杯中,置于积分球窗口,进行近红外光谱扫描。将扫描结果即光谱数据存入计算机,每个样品测3次取其平均。而后将光谱数据输入已编制好的计算机应用程序,经运算分析后得出样品中大肠菌群的检

测报告。

2 试验中的数学方法

2.1 余弦相似度聚类

试验中的光谱数据异常庞大,为了获得光谱数据中与大肠菌群有关的信息,必须采用现代数学方法进行降维处理。常用的聚类方法是使用欧氏距离

离相似度 $\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}$,其中 x_i 为试验值, $\bar{x}$ 为试验值的平均值;或者是马氏距离相似度 $\sqrt{(x - \bar{x}) / \sum (x_i - \bar{x})}$,其中 x_i 为向量, $\bar{x}$ 为样本向量的代数平均值向量。以相似度作为聚类的标准,两种方法所得到的都是点与点之间的距离,反映的是样本之间的差异。利用向量的角距离余弦相似度 $(x_i \cdot \bar{x}) / (|x_i| \cdot |\bar{x}|)$,其中 x_i 为向量, $\bar{x}$ 为向量组的中心向量。虽然因聚类中心不易确定,造成分组不易而较少使用,但角距离体现了向量和中心向量的贴合程度,利于确定状态的接近情况,便于确定 n 维空间包含信息特性的一组基,在这些方面角距离法具有一定的优势。

如图1所示,在平面I中有3个向量,当选定 a , b 两个为中心向量的时候,由于 c 与 a 的角距离比 c 与 b 的角距离小,因而可把 c 与 a 聚为一类。在空间中亦如此(见图2),利用向量的角距离,可以把向量按照它们的角距离进行分组。对于空间中任何一个向量 a ,都可以按角距离把它分离它最近的一组中去。计算各个组的平均角距离,并对平均角距离为最大的那一组进行拆分,达到每组的平均角距离的范围为最小(见图3)。由图3可分辨出平均角距离范围最大的一组,对该组进行拆分。在拆分的时候,先任意选定平均角距离边界上的一点(即图3中椭圆上的一点),以它为拆分的一个中心;再在平均角距离边界上,找到与该选定点关于中心向量对称的一点作为另一个中心,并且计算出每组的平均角距离作为分组后合并的标准;对于拆分后的各个中心向量,如果它们的角距离小于拆分后重新成的组 a 或组 b 的平均角距离,则将这俩个中心合并。这是因为,对于重新分组后产生的组的分布不一定是各向同性,必须将两组合并,以保证每个组的分布都是各向同性的(这样标准差就得到了控制)。继续拆分,就可以对一系列数据从一组逐步分解到多个组,进而得到若干个具有代表性的向量。余弦相似度动态聚类方法的计算框图见图4^[9-10]。

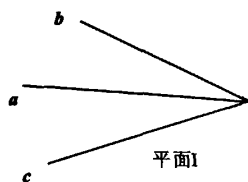


图1 平面向量聚类原理图

Fig.1 Planar vectors clustering

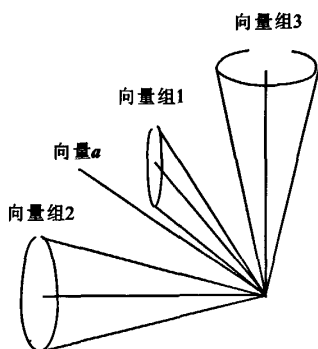


图2 空间向量聚类原理图

Fig.2 Dimensional vectors clustering

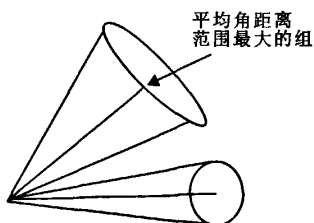


图3 拆分示意图

Fig.3 Cluster splitting

2.2 偏最小二乘回归分析

利用余弦相似度动态聚类法找到原料乳的特征波数,通过偏最小二乘法建立这些特征波数点的吸光度和原料乳中大肠菌群的数量关系,从而建立原料乳中大肠菌群数的余弦相似度偏最小二乘法预测模型。

2.3 逐步回归分析

利用逐步回归法搜索原料乳的特征波数,再建立原料乳的近红外光谱特征波数点的吸光度和大肠菌群数的逐步回归法预测模型。

采用 C++ Builder 作为开发平台,分别编制偏最小二乘法、逐步回归分析法和基于余弦相似度聚类的偏最小二乘法的光谱分析程序。

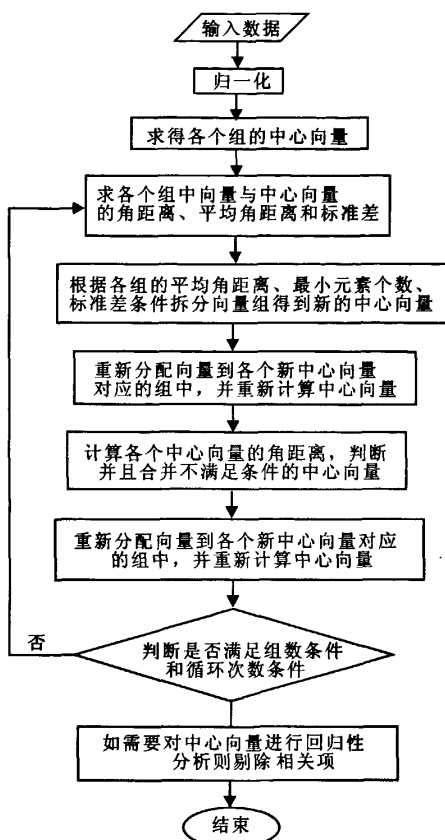


图4 余弦相似度聚类流程

Fig.4 The flow chart of cosine similarity clustering

3 结果与分析

试验中共检测原料乳样品 45 个。将各样品依据生化检测方法所测大肠菌群数从少到多顺次编号,按 4 取 1 的方法组成定标集和预测集(即定标集为 3,预测集为 1。)。余下的 1 个样品放入定标集,这样定标集总样品数为 34,预测集总样品数为 11。

利用上述试验方法分别对 34 个原料乳试样的近红外光谱进行分析,通过与生化检测方法所得结果进行定标计算,建立原料乳中大肠菌群的数学预测模型。运算过程中剔除 3 个样品,实际参与建模运算的样品数为 31 个。

采用偏最小二乘回归分析方法、逐步回归分析方法和基于余弦相似度聚类分析结合偏最小二乘回归分析的近红外光谱方法,依次建立原料乳中大肠菌群数的近红外光谱分析模型,各预测模型对 11 个原料乳样品中大肠菌群数的预测结果与生化分析结果的相关性分别见图 5~7。各预测模型的统计量见表 1。

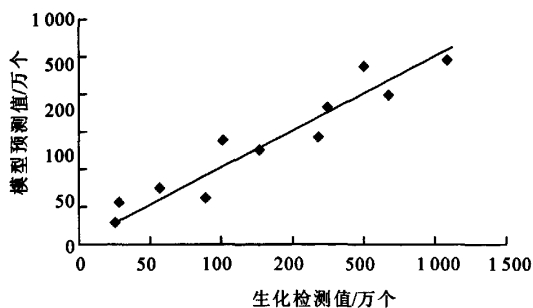


图 5 偏最小二乘法大肠菌群预测值与生化检验值的相关图

Fig. 5 Comparison of coli group values by partial least square method and biochemical method

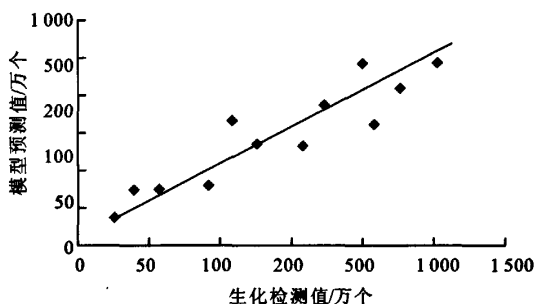


图 6 逐步回归法大肠菌群预测值与生化检验值的相关图

Fig. 6 Comparison of coli group values by step-by-step regression method and biochemical method

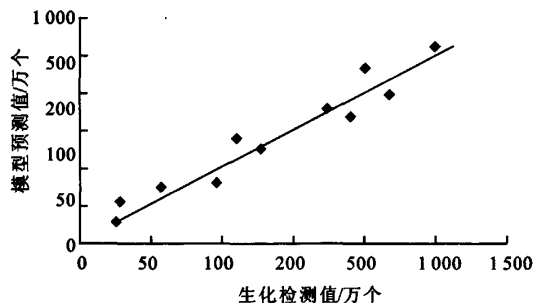


图 7 余弦相似度聚类+偏最小二乘法大肠菌群预测值与生化检验值的相关图

Fig. 7 Comparison of coli group values by cosine similarity clustering method and partial least square method

为了检验各方法对原料乳中大肠菌群数的预测性能,对各预测模型进行了相关系数检验和显著性 F 检验。3 种数学模型的预测值与生化检验值之间的相关系数 R^2 分别为 0.912 6、0.892 1 和 0.954 3。经 F 检验计算, F 值分别为 8.494、3.816

和 9.046,在 $\alpha=0.05$,即置信度为 95%的情况下,其 F 检验临界值 $F_{\alpha=0.05}(7,23)=2.44$ 。由此可以说明,3 种方法所建数学模型在进行原料乳中大肠菌群数的预测中理论上都是可行的。但从 F 值计算结果也可以看出,利用逐步回归分析方法所建模型的显著性不强,刚过临界值;另一方面,从相关系数分析来看,其相关系数只有 0.892 1,这在实际生产中是达不到检测精度要求的,因而该模型不能用于生产中检测原料乳中的大肠菌群数。其他两种模型可用于生产实际,且基于余弦相似度聚类分析结合偏最小二乘回归分析的近红外光谱方法优于其他两种方法,是较理想的代替人工生化法来检测大肠菌群数的方法,所用检测时间大大少于生化检验时间。

表 1 原料乳中大肠菌群预测模型的统计量

Tab. 1 Statistics of prediction models for *E. coli* group in raw milk

建模方法	预测模型相关方程	R^2	F 值
偏最小二乘法	$y=0.990\ 2x+0.001\ 7$	0.912 6	8.494
逐步回归法	$y=0.964\ 1x+0.120\ 4$	0.892 1	3.816
余弦聚类加偏最小二乘法	$y=1.026\ 1x-0.122\ 2$	0.954 3	9.046

虽然本试验借助近红外光谱分析技术找到了能够替代人工生化检测大肠菌群的快速分析方法,但近红外光对于大肠菌群的作用机理尚不清楚,到底是何种物质(或生物大分子)对近红外光的吸收也不明白,这都有待于进行近红外光谱分析技术的进一步深入研究来解决。

4 结 语

该试验利用近红外光谱分析技术,结合现代数学处理方法,研究了原料乳中大肠菌群的光谱特征,得到如下结论:

1) 采用 C++ Builder 作为开发平台,分别编制了偏最小二乘法、基于余弦相似度因子分析的偏最小二乘法及逐步回归分析法的光谱分析程序。

2) 通过样品的近红外光谱分析,找到了原料乳中大肠菌群的特征吸收光谱,运用所编程序建立了用于原料乳中大肠菌群数预测的数学模型,建立了原料乳中大肠菌群数的快速分析测定方法。

3) 对各模型的预测结果进行分析比较,认为逐步回归方法不适于生产中对原料乳中大肠菌群数的检测,而偏最小二乘法和基于余弦相似度聚类及偏最小二乘法的近红外光谱分析技术可以作为生

产中进行原料乳中大肠菌群数的快速检测方法,其预测模型显著,预测值与生化检测值之间的相关系数分别达到0.912 6和0.954 3。

参考文献(References):

- [1] 严衍祿. 近红外光谱分析基础与应用[M]. 北京:中国轻工业出版社,2005.
- [2] 慈云祥,臧凯赛,高体玉. 几种微生物的红外光谱研究[J]. 高等学校化学学报, 2002,23(6):1047—1049.
CI Yun-xiang,ZHANG Kai-sai,GAO Ti-yu. FTIR study of microbes[J]. *Chemical Journal of Chinese Universities*, 2002, 23(6):1047—1049. (in Chinese)
- [3] Dieter Helm, Dieter Naumann. Identification of some bacterial cell components by FTIR spectroscopy[J]. *FEMS Microbiology Letters*, 2000,126:75—80.
- [4] Naumann D, Labischinski H, Giesbrecht P. Modern Techniques for Rapid Microbiological Analysis[M]. New York: VCH Verlag Chemie, 2001.
- [5] Naumann D. The ultra rapid differentiation and identification of pathogenic bacteria using FTIR techniques[J]. *Fourier and Computerized Infrared Spectroscopy*, 1998, 553: 268—269.
- [6] Naumann D, Fifala V, Labischinski H, et al. The differentiation and identification of pathogenic bacteria using FTIR and multivariate statistical analysis[J]. *Mikrochimical Acta*, 1999,(1): 373—377.
- [7] Naumann D, Fifala V, Labischinski H, et al. The rapid differentiation and identification of pathogenic bacteria using FT-IR and multivariate statistical analysis[J]. *Journal of Molecular Structure*, 2001,174: 165—170.
- [8] 姚占芳,吴云汉. 微生物学实验技术[M]. 北京:气象出版社,1998.
- [9] 刘建学,李守军. 基于 Matlab 与 C++Builder 偏最小二乘分析的编程实现[J]. 江南大学学报:自然科学版,2006,5(1):34—37.
LIU Jian-xue,LI Shou-jun. Partial least square and its realization by using the interface of Matlab and C++ builder[J]. *Journal of Southern Yangtze University: Natural Science Edition*,2006,5(1):34—37. (in Chinese)
- [10] 刘建学,李守军. 基于余弦相似度的因子分析在食品成分检测中的应用[J]. 食品科学,2005,26(6):109—112.
LIU Jian-xue,LI Shou-jun. Dynamical clustering by cosine similarity algorithm[J]. *Food Science*,2005,26(6):109—112. (in Chinese)

(责任编辑:李春丽)

《江南大学学报(自然科学版)》

征稿、征订启事

《江南大学学报(自然科学版)》(双月刊)是由教育部主管、江南大学(国家“211工程”重点建设高校)主办的自然科学类学术期刊。本刊主要刊载通信与控制工程、信息工程、机械工程、产品系统设计理论、纺织工程、应用化学、材料工程、土木工程、数理科学等学科的学术论文、研究报告,以及反映学科前沿研究动态的高质量综述。

本刊优先刊登国家自然科学基金和省部级及其以上科研项目析出论文,同时发表与企业及生产实际密切相关的应用性研究成果。热忱欢迎广大高校教学、科研人员及相关领域的专家、学者,在读硕士、博士研究生赐稿。

本刊为A4开本,128页,每册订价8.00元;全年共6期,48.00元;本刊邮发代号:28—189,全国各地邮局均可订阅,亦可向本刊编辑部直接订购。本刊可破季订阅。热忱欢迎广大读者订阅本刊。

邮编:214122

电子邮箱:xbzrkx@jiangnan.edu.cn

地址:江苏省无锡市蠡湖大道1800号

电话:0510—85913519

《江南大学学报(自然科学版)》编辑部