文章编号: 1002-1582(2009)01-0070-04

改进偏最小二乘法在近红外牛奶成分测量中的应用*

李振庆, 黄梅珍, 倪一, 丁海峰, 汤洁蔚, 窦晓鸣
(上海交通大学 物理系光学工程研究所, 上海 200240)

摘 要: 采用 Nicolet Nexus 870 红外-近红外傅里叶变换光谱仪测量了 36 个市售巴氏杀菌纯牛乳样品的透射光谱。在近红外光谱 1254～1875nm 和 2045～2372 nm 波段内,为了选择携带信息量大的波长区域,采用改进偏最小二乘回归法,包括间隔偏最小二乘法、移动窗口偏最小二乘法和可变窗宽移动窗口偏最小二乘法对巴氏杀菌纯牛乳中脂肪、蛋白质及乳糖成分分别建立模型,进行了分析和比较,结果表明,采用改进偏最小二乘法所选出的波长区与目标值的相关程度高,可以较好地建立牛奶的预测模型。

关 键 词: 光谱学; 改进偏最小二乘方法; 牛奶
中图分类号: TN215 **文献标识码**: A

Using improved PLS methods for milk components determination by near infrared spectra

LI Zhen-qing, HUANG Mei-zhen, NI Yi, DING Hai-feng, TANG Jie-wei, DOU Xiao-ming
(Institute of Optical Engineering, Department of Physics, Shanghai Jiao Tong University, Shanghai 200240, China)

Abstract: 36 transmittance spectra of pasteurised milk are tested by a FT-IR-Spectrometer (Nicolet Nexus 870). In order to find the most informative region, three improved partial-least squares regression methods, including interval partial-least square regression method (iPLS), moving widow partial-least square regression method (MWPLSR) and changeable size moving window partial least-square regression method (CSMWPLS) are used for fat, protein and lactose model building between 1254nm to 1875nm and 2045nm to 2372nm. The result shows that the spectra region optimized by CSMWPLSR can relate the targets better and thus build a good milk component predictive model.

Key words: spectroscopy; improving partial least square regression method; milk

0 引 言

近红外光谱法是一种快速简便非破坏性的方法,它已被广泛地应用于农作物或食品的定量分析[1],作为在食品科学与农业科学方面的一种重要检测技术,它已经越来越受到人们的重视。

牛奶是一种重要的农副产品,随着人们健康意识的提高,牛奶由于其中含有较丰富的蛋白质、脂肪、乳糖,作为人们的一种日常饮用品其成分含量越来越倍受人们的关注,同时,牛奶成分的测量对于有效地利用奶牛,奶牛饲养和年奶业至关重要,它已成为奶牛厂管理决策的重要依据。目前,国内主要乳品检测站对牛奶成分的检测除了采用经典的化学方法外,主要采用基于光谱方法的乳品分析仪器,如 Foss 公司的 MilkScan。

采用近红外光谱技术实现牛奶成分的测量,关键在于从光谱数据中获取有用的光谱信息,建立各成分与光谱信息之间的有效关联。本文的目的是探讨使用近红外光谱技术测量牛奶中主要成分如蛋白质,脂肪、乳糖等含量的有效建模方法。比较间隔偏最小二乘法、移动窗口偏最小二乘法和可变窗宽移动窗口偏最小二乘法对巴氏杀菌纯牛乳中脂肪、蛋白质及乳糖成分建模的预测效果。

1 原理及方法

1.1 测量原理

近红外光谱是分子振动光谱的倍频和合频吸收光谱,主要是 X-H 键 (X 为 C,O,N,S 等)的吸收,不同基团产生的光谱在吸收峰位置和强度上有所不同,根据朗伯-比耳吸收定律(Lambert-Beer Law),随着样品成分含量的变化,其光谱特征也将发生变化。这是近红外光谱分析方法的理论基础[2]。

1.2 数据模型

1.2.1 间隔偏最小二乘法

偏最小二乘法 (partial-least squares regression method, PLS) 一般是在全谱范围来建立模型的。由

* 收稿日期: 2008-01-14; 收到修改稿日期: 2008-06-04 E-mail: mzhuang@sjtu.edu.cn
基金项目: 医学光电科学与技术教育部重点实验室(福建师范大学)资助项目
作者简介: 李振庆(1981-),男,上海交通大学博士研究生,从事光学检测技术研究。

于信息光谱常常位于某一波段,为了找出目标信息含量最为丰富的波段,可以将所测得的光谱等分成 n 个间隔,把每个间隔内连续的 w 个波长点作为一个窗口(如图 1 所示),设置一最大主成分数,对每一窗口进行偏最小二乘分析,利用交叉验证,根据 RMSECV(root mean square error of cross validation)找出对应每一波段的最佳主成分数。对照整个波谱分成的若干波段建模后的 RMSECV,最终找出目标光谱信息含量最佳的第 i 个波段[3]。

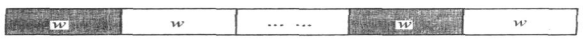

图 1　间隔偏最小二乘法区间选择方式

1.2.2　窗口移动偏最小二乘法

相对于偏最小二乘方法 PLS,在移动窗口偏最小二乘法 MWPLSR[4]中,选取一个宽度为 w 的光谱窗口,从整个光谱(假设有 n 个波长点)的第一个波长点开始依次向右移动一个波长点直至最后,如图 2 所示,落在窗口内的波长点为 i 到$(i + w - 1)$,其中 i 为窗口的起始波长点,w 为窗口宽度,从而可以从整个光谱中依次选择$(n - w + 1)$个包含 w 个波长点的子波长区,设置一最大主成分数,对每个子波长区分别建立偏最小二乘法 PLS 模型,分别得到不同主成分数里对应 PLS 模型的预测误差 RMSEP(root mean square error of prediction),从而找出含有用信息的一个或几个波长区。对于已选出的子波长区建立最终的 PLS 模型一般有两种方法:一种是合并这些子波长区建立 PLS 模型;一种是利用几子波长区建立 PLS 模型。

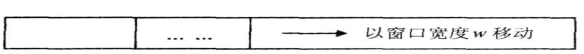

图 2　偏最小二乘窗口移动方式

1.2.3　可变窗宽窗口移动偏最小二乘法

相对于一般偏最小二乘法,CSMWPLS[5]可以优化一段波长区,从而得到信息含量较为丰富的波段,它的工作原理为对于给定的一段含有 p 个波长点的光谱区域,设定窗口宽度 $w = 1$,从整个光谱的第一个波长点开始依次向右移动一个波长点直至最后,然后改变窗口的宽度 $w + 1$,再从起始点移动窗口到最后,循环直至 $w = p$,设定一最大主成分数,对于得到的 $p(p + 1)/2$ 个波段分别进行 PLS 分析,根据建模后的 RMSECV 找出信息含量丰富的最佳窗口以及最佳主成分数,窗口移动过程如下图 3 所示。

2　数据测量与模型建立

2.1　数据测量

使用傅里叶变换红外光谱仪(Nicolet Nexus 870)对来源于不同生产厂商不同批号的 36 份巴氏杀菌均质奶进行透射光谱测量,其中光谱仪分辨率 2cm^{-1},光谱范围为 12000—2000cm^{-1},样品池为 1mm 光程石英样品池,为降低随机噪声,每次测量均扫描 128 次后取平均,为减小样品池位置带来的误差每个样品测量 3 次取平均后共得 36 条光谱,光谱图见图 4。样品中,脂肪,蛋白质和乳糖含量参考值由上海市乳品质量监督检验站使用经典化学方法测量后提供。

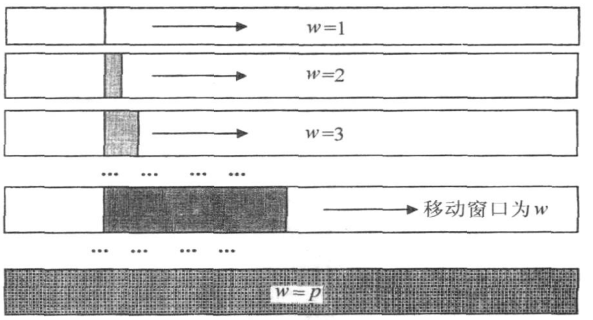

图 3　可变窗宽窗口移动偏最小二乘法窗口选择方式

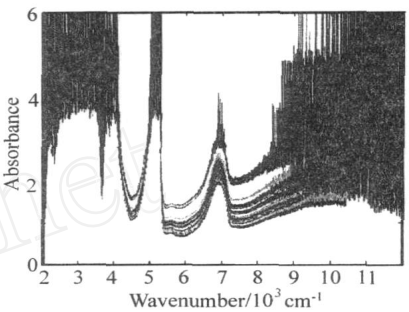

图 4　原始光谱图

2.2　数据预处理

由图 4 可知,检测器所检测到的光谱信号除含有主要成分的信息外,由于牛奶溶液为乳浊液,在低波长区域由于瑞利散射较为严重以及高波长区由于所选光程较大导致光谱信息噪声较大。为减弱噪声对目标信息的影响,对测得的 36 组光谱数据进行预处理,首先选取光谱中噪声较小的波段 1254—1875nm 和 2045—2372nm,之后对光谱进行分段平滑处理。

在实验测量过程中,由于实验条件及实验仪器等多方面的影响,实验数据可能会出现偏离总体趋势的个别点,这些偏离点的存在将影响到数据处理及实验结果,因此,在数据处理过程中采用马氏距离法[6],来判断奇异样品。

2.3　数据建模

为提高模型质量,依次使用 PLS[7]、iPLS、MWPLS 和变窗宽移动窗口偏最小二乘 CSMWPLS 结合多元散射校正(MSC)建立预测模型。

由于对各个目标参数的建模方法和过程是相同的,下文以脂肪为例,介绍建立脂肪模型的步骤。

步骤 1:粗选波段建模

采用 PLS 对脂肪建模,光谱波段 1254—1875nm 及 2045—2372nm。模型及预测结果见表 1。步骤 2:使用 iPLS 方法,分别将光谱等分为 40 和 80 个窗口建立脂肪模型,检验选择波段建模是否优于全谱建模,建模结果如图 5 和图 6 所示。

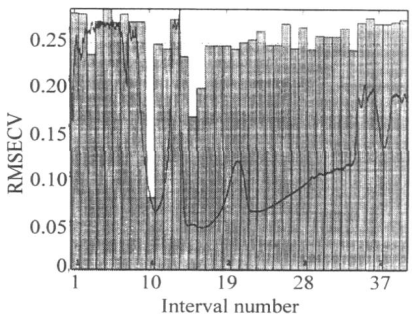

图 5 对脂肪的 iPLS 建模,间隔 40

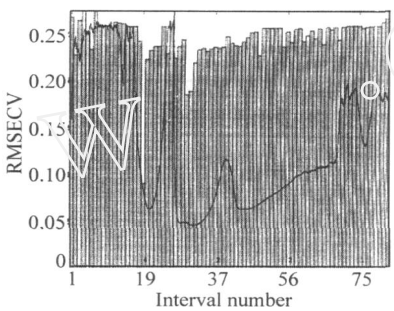

图 6 对脂肪的 iPLS 建模,间隔 80

在图 5,图 6 中,水平实线为利用全谱建立 PLS 模型对应 RMSECV,可以看出,RMSECV 在某些特定波段变小,且即使建模所取间隔不同,RMSECV 较小的波段位置仍大致相同。由此可见,选择此特定波段建模,效果将优于全谱建模。

步骤 3:使用 MWPLS 方法建模

虽然 iPLS 方法指出选择波段建模的效果优于全谱建模,并给出全谱中残差变小波段的大致位置,但仍无法精确波段起止点。因此采用 MWPLS 方法进一步决定所需波段位置并建立模型。设定窗口宽度,并对全谱移动窗口。在每一个窗口位置,计算每个成分数在此窗口位置时的 RMSECV,选取以某一给定的成分数达到较小的误差水平的波段,以此区间及其对应窗宽、窗口位置建模。设窗宽 149,成分数范围(1—20),全谱进行 MWPLS,脂肪全谱范围的结果如图 7 所示,横坐标为波数,纵坐标为 Ln(SSR)(SSR:sums of squared residues)。图 8 为局部波段放大图,具体所选波段及建模结果见表 1。

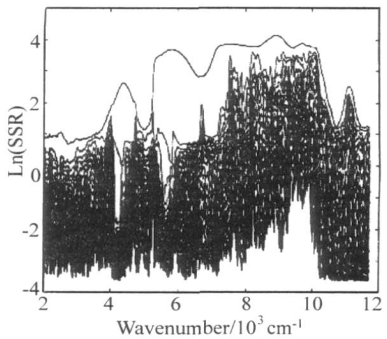

图 7 脂肪 MWPLS 误差曲线(全谱)

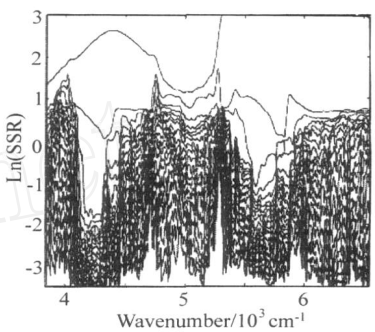

图 8 脂肪 MWPLS 误差曲线,部分波段放大

在图 7 和图 8 中,可以看出,误差随主成份数增大逐渐减小,当主成分数达到一定值时,误差变化幅度越来越小。

步骤 4:使用 CSMWPLS 方法建模

对于依据 MWPLS 的结果选出的波段进一步使用 CSMWPLS,优化建模范围,计算 R 及 RMSEP。在 CSMWPLS 过程中,给定成分数范围(1—20)及窗宽范围(1—200),对步骤 3 中找到的信息含量较丰富的波段进一步优化建模范围。在每一个窗口的建模过程时,随机抽取 16 个样本作为预测样本集,其余样本作为校正集建模,重复该过程 216 次(样本总数的 6 倍),通过 RMSECV 确定该窗口的最佳主成分数。对所有可能的窗口都完成建模后,比较 RMSECV,确定最优建模范围。

2.4 模型及预测结果

根据以上步骤,对脂肪建立不同模型后的结果如表 1 所示,同理,采用 MWPLS 以及 CSMWPLS 方法对蛋白质和乳糖建立数学模型后,结果如表 2、表 3 所示。

表 1 脂肪部分波段建模及其 MSC 方式、成分数、R 值及 RMSEP 值

波段选择方式	波段/nm	波段/nm	MSC 方式	PC	R	RMSEP
粗选	[1254,1875]	[2045,2372]	分段 MSC	5	0.95328	0.0755
		[2045,2372]	MSC	5	0.97531	0.0552
	[1254,1875]		MSC	2	无法建模	
MWPLS	[1667,1790]	[2192,2401]	分段 MSC	5	0.97505	0.0555
		[2192,2401]	MSC	5	0.97755	0.0527
	[1667,1790]		non-MSC	4	0.93403	0.0895
CSMWPLS	[1688,1790]	[2262,2394]	分段 MSC	5	0.97782	0.0524
		[2262,2394]	non-MSC	5	0.97696	0.0534
	[1688,1790]		non-MSC	4	0.93708	0.0874

表 2　蛋白质部分波段建模及其 MSC 方式、成分数、R 值及 RMSEP 值

波段选择方式	波段/ nm	波段/ nm	MSC 方式	PC	R	RMSEP
粗选	[1254,1875]	[2045,2372]	non-MSC	5	0.79611	0.0617
		[2045,2372]	MSC	6	0.93524	0.0361
	[1254,1875]		non-MSC	4	无法建模	
MWPLS	[1629,1797]	[2116,2284]	分段 MSC	4	0.89955	0.0445
		[2116,2284]	MSC	3	0.90326	0.0438
	[1629,1797]		MSC	3	0.82689	0.0574
CSMWPLS	[1634,1797]	[2118,2284]	分段 MSC	4	0.89945	0.0446
		[2118,2284]	MSC	3	0.90713	0.0429
	[1634,1797]		non-MSC	4	0.85813	0.0526

表 3　乳糖部分波段建模及其 MSC 方式、成分数、R 值及 RMSEP 值

波段选择方式	波段/ nm	波段/ nm	MSC 方式	PC	R	RMSEP
粗选	[1254,1875]	[2045,2372]	non-MSC	5	0.72071	0.117
		[2045,2372]	MSC	3	0.71623	0.1178
	[1254,1875]		non-MSC	2	无法建模	
MWPLS	[1615,1705]	[2129,2287]	non-MSC	5	0.84095	0.0913
		[2129,2287]	non-MSC	4	0.8047	0.1002
	[1615,1705]		non-MSC	3	0.64117	0.1296
CSMWPLS		[2214,2261]	non-MSC	4	0.85462	0.0883
	[1623,1694]		non-MSC	4	0.7732	0.1071
	[1623,1694]	[2129,2287]	分段 MSC	4	0.84561	0.0901

3　实验结果

本文采用改进偏最小二乘法对牛奶中的蛋白质、脂肪和乳糖进行建模分析。结果表明,改进的偏最小二乘法可以提高模型的预测能力:采用 CSMWPLS 对脂肪预测,主成分数为 5 时,相关系数 R 可达 0.97782,RMSEP 为 0.0429;采用 CSMWPLS 对蛋白质预测,主成分数为 3 时,相关系数 R 为 0.90713,RMSEP 为 0.0361;采用 CSMWPLS 对乳糖预测,主成分数为 4 时,相关系数 R 为 0.85462,RMSEP0.08831。本实验所采用的方法,若能进一步对牛奶中的其它微量成分如非脂乳固体、农药残留量以及添加剂等进行分析,将具有更重要的意义。

特别感谢上海市乳品质量监督检验站提供的大力帮助。

参考文献:

[1] Osborne B G, Fearn T; Hindle P H. Practical NIR spectroscopy with applications in food and beverage analysis[M]. Longman Scientific & Technical, UK: Essex, 1993. 227.
[2] 李庆波,汪皑燕,等. 牛奶主要成分含量近红外光谱快速测量法[J]. 食品科学,2002,23(6): 121.
[3] Dayal S, MacGregor J F. Improved PLS algorithms[J]. Bhupinder. Chemometrics, 1997, 11: 73—85.
[4] Jiang J H, Berry R J, Siesler H W, et al. Wavelength interval selection in multicomponent spectral analysis by moving window partial least-squares regression with applications to mid-infrared and near-infrared spectroscopic data[J]. Anal. Chem. 2002, 74: 3555—3565.
[5] Du Y P, Liang Y Z, Jiang J H, et al. Spectral regions selection to improve prediction ability of PLS models by changeable size moving window partial least squares and searching combination moving window partial least squares[J]. Analyica Chimica Acta, 2004, 501: 183—191.
[6] 陆婉珍,袁洪福,徐广通. 现代近红外光谱分析技术[M]. 北京: 中国石化出版社,2000. 16—17.
[7] 梁逸曾,俞汝勤. 分析化学手册(第十分册),化学计量学[M]. 2000, 12. 161-169, 211—212.

(上接第 69 页)

LEO 通信的 OPTEL-02 水平指向范围为 - 180°—+180°,俯仰指向范围为 - 25°—+ 85°,比较我们从仿真得到的结论可以发现,本文设计的星座星间链路参数完全可以满足实际卫星激光通信终端。

下一步的研究重点是分析该网络的覆盖特性,研究具体的路由算法和接入策略,评估网络性能,对其结构进行进一步的优化。

参考文献:

[1] Pratt T, Bostian C, Allnutt J. 卫星通信(第二版)[M]. 北京:电子工业出版社,2005. 341—342.
[2] Chan V W S. Optical space communications[J]. IEEE, 2000, 6(6): 959—975.
[3] Chan V W S. Optical satellite networks[J]. Journal of Lightwave Technology, 2003, 21(11): 2811—2855.
[4] Chan V W S. Free-space optical communications[J]. Journal of Lightwave Technology, 2006, 24(12): 4750—4762.
[5] Suzuki R, Morikawa E, Yasuda Y. A study of constellation for LEO satellite communication network[C]. 21st International Communications Satellite Systems Conference and Exhibit, USA: Washington, 2003 AIAA, 2003. 2324.
[6] Suzuki R, Motoyosh S. A study of constellation for LEO satellite network[C]. 22nd AIAA International Communications Satellite Systems Conference & Exhibit, USA: Monterey, California, AIAA, 2004. 3236.
[7] Gerber A J, Tralli D M, Bajpai S N. Medium earth orbit (MEO) as an operational observation for NOAA's post GOES-R environmental satellites[J]. SPIE, 2005, 5659: 261—271.
[8] A. H. BALLARD, "Rosette Constellations of Earth Satellites[J], IEEE Transactions on Aerospace and Electronic Systems, 1980, AES-16(5): 656—673.
[9] 王振永,王平,顾学迈. 卫星网络中永久星间链路的设计方法研究[J]. 通信学报,2006, 27(8): 129—133.
[10] Baister G, Dreischer T, Fischer E. OPTEL family of optical terminals for space based and airborne platform communications links[C]. SPIE, 2005, 5986: 1—9.