文章编号　1004-924X(2012)10-2170-06

可见-近红外光谱测定血红蛋白的等效波段选择

刘振尧,潘　涛*
(暨南大学 光电信息与传感技术广东普通高校重点实验室,广东 广州 510632)

摘要:将可见-近红外光谱和改进的移动窗口偏最小二乘(MWPLS)方法应用于人类血红蛋白(HGB)无试剂快速检测的高精度波段优选。为了避免模型评价失真,提出了一种新的模型评价体系。首先,从全体 205 个样品中随机抽取 70 个作为检验集,余下的 135 个作为建模集,并划分为具有相似性的定标集(80 个样品)和预测集(55 个样品)共 50 次;其次,对每一次划分都分别建模和优化,使得模型具有稳定性;最后,利用检验集对优选出的模型进行再次检验。实验结果表明:可见-短波近红外波段 400～1100 nm 可以作为人体全血 HGB 的信息波段;进一步采用 MWPLS 方法从 400～1100 nm 中选出全局最优波段为 492～890 nm,并得到包含 77 个等效波段的模型空间。以 492～890 nm 为例,检验效果预测均方根偏差(V-SEP)、预测相关系数(V-RP)和相对预测均方根偏差(V-RSEP)分别为 2.58 g L^{-1}、0.988 和 1.97%,得到的样品的 HGB 预测值与临床实测值吻合精度很高,可望应用于临床。

关　键　词:人类全血;血红蛋白;VIS-NIR 光谱;波段选择;等效模型空间
中图分类号:O657. 33　　**文献标识码**:A　　**doi**:10.3788/OPE.20122010.2170

Equivalent waveband selection of VIS-NIR spectroscopic measurement for hemoglobin

LIU Zhen-yao, PAN Tao*

(*Key Laboratory of Optoelectronic Information and Sensing Technologies of Guangdong Higher Educational Institute, Jinan University, Guangzhou* 510632, *China*)
* *Corresponding author, E-mail: tpan@jnu.edu.cn*

Abstract: The VIS-NIR spectroscopy combined with the improved Moving Window Partial Least-square (MWPLS) method was applied to a high accurate waveband selection for the rapid no-reagent determination of Hemoglobin (HGB) in human whole blood. A new modeling evaluation system was proposed to avoid the evaluation distortion. First, seventy samples were randomly selected from a total of 205 samples as the validation set, the remaining 135 samples were used as the modeling set, and the modeling set was divided into similar calibration (80 samples) and prediction (55 samples) sets for a total of 50 times. Then, modeling and optimization were performed in each division to get stable model. Finally, the optimized model was validated again using the validation set. Experimental results indicate that the VIS-short NIR region 400－1 100 nm can be used as the information waveband of HGB in human whole blood, the global optimal waveband 492－890 nm is further selected from 400－

收稿日期:2012-05-29;**修订日期**:2012-07-05.
基金项目:国家自然科学基金资助项目(No. 61078040);广东省科技计划资助项目(No. 2009B030801239, No. 2009A030301002)

1100 nm with MWPLS method, and a model space including 77 equivalent wavebands is obtained. By taking the 492－890 nm for an example, validation effects V-SEP, V-RP, and V-RSEP are 2.58 g L^{-1}, 0.988, and 1.97%, respectively. It concludes that HGB prediction values of the samples are highly close to the clinic measured values, which may be used in clinical diagnosis.

Key words: human blood; hemoglobin; VIS-NIR spectroscopy; waveband selection; equivalent model space

1　引　言

血红蛋白(HGB)是红细胞中负责运载氧的蛋白质，HGB 含量可用于判断贫血和体内铁的营养状况，是重要的临床生化指标。血液中血红蛋白的常规检测方法需要消耗化学试剂且过程复杂，因此建立无需化学试剂、简便快速的血红蛋白定量检测方法具有重要意义。

近红外(NIR)光谱是一种获得物质定量和定性信息的手段，其测试简单便捷、适合于无试剂或无创检验[1-3]，因此利用 NIR 光谱建立无试剂、快速测定 HGB 的方法成为近年来的重要研究方向[4-6]。血液是一种多组分的复杂体系，利用光谱直接测量血液样品需要克服很多噪音干扰，目前血液 HGB 的光谱预测精度还未达到临床水平。光谱信息波段的优选和模型稳定性是需要改进的两个重要方向[7-9]，本文通过改进移动窗口偏最小二乘(MWPLS)方法[10]，完成了具有稳定性的 HGB 的可见-近红外波段优选。

客观、合理的模型评价方法对于光谱分析是至关重要的。很多实验结果表明，定标集、预测集的不同划分会引起预测效果和模型参数的波动，使结果不稳定。由于需要大量实验，既往的研究很少涉及到模型稳定性。随机选择样品用于最后的模型检验是合理的，但是在模型优选环节，为了避免模型评价失真，有必要考虑建模定标集和建模预测集的相似性。基于上述的随机性、相似性和稳定性，本文提出了一种新的模型评价体系，使得建立的模型具有客观性和稳定性，从而具有实用性。首先，从样品中随机抽取部分样品作为检验集，余下样品作为建模集；再把建模集划分为具有相似性的定标集和预测集，重复多次划分，对每一次划分都分别建模和优化，使得模型具有稳定性；最后，利用检验集对优选出的模型进行再次检验。

2　实验与方法

2.1　实验材料、仪器和测量方法

收集了 205 份人类全血样品，HGB 含量使用深圳迈瑞公司的 BC-3000Plus 生化分析仪测定，所得实测值用于光谱分析的建模与检验。全体样品的 HGB 实测值为 78～168 g·L^{-1}，其均值、标准偏差分别为 131 和 17.1 g·L^{-1}。

光谱仪器为丹麦 FOSS 公司的 XDS Rapid ContentTM 型近红外光栅光谱分析仪和透射样品附件。光谱扫描 400～2 498 nm(除了全近红外区还包含大部分可见光区)；波长间隔 2 nm；400～1 100 nm、1 100～2 498 nm 波段分别用 Si 和 PbS 探测器；透射附件使用光程为 2 mm 的比色皿。每个样品测量 3 次，3 次光谱的均值作为样品的光谱数据。实验温度、湿度分别为 25±1 ℃、45±1% RH。205 个全血样品的可见-近红外光谱如图 1 所示。

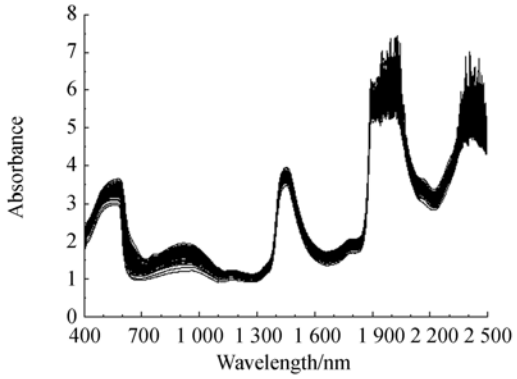

图 1　205 个全血样品的可见-近红外光谱

Fig. 1　VIS-NIR spectra of 205 whole blood samples

2.2 样品划分、模型优选框架

首先，从 205 个样品中随机抽取 70 个样品作为检验集，余下的 135 个样品作为建模集；再把建模集划分为具有相似性的定标集（80 个样品）和预测集（55 个样品），重复划分 50 次。样品划分过程如图 2 所示。对于每一个划分 i 建立定标预测模型，计算均方根偏差和相关系数，分别记为 M-SEC$_i$，M-RC$_i$，M-SEP$_i$ 和 M-RP$_i$，进一步计算它们对于所有划分的平均值和标准偏差，其中预测效果指标分别记为 M-SEP$_{Ave}$，M-SEP$_{Std}$，M-RP$_{Ave}$和 M-RP$_{Std}$，用于评价模型的预测精度和稳定性，并基于 M-SEP$_{Ave}$优选模型参数（波段、PLS 因子数等）。最后，利用检验集对优选出的模型进行再次检验，计算检验的预测均方根偏差、预测相关系数和相对预测均方根偏差（相对于实测值的均值），分别记为 V-SEP、V-RP 和 V-RSEP。

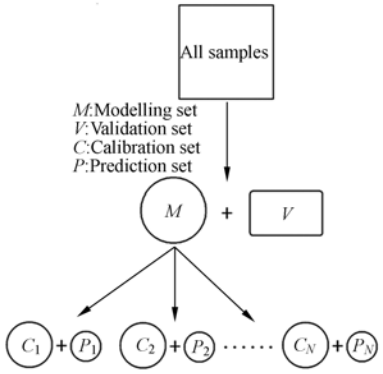

图 2　样品划分

Fig. 2　Division of samples

采用 HGB 实测值来定义定标集和预测集的相似性，当定标集和预测集的 HGB 平均值、标准偏差都接近时，可以定义两个集合相似。采用计算机程序将建模集随机划分为定标集、预测集足够的次数，然后计算每次划分得到的定标集、预测集的 HGB 平均值和标准偏差，并计算整个建模集的 HGB 平均值、标准偏差，分别记为 HGB$_{C,Ave}$，HGB$_{C,Std}$，HGB$_{P,Ave}$，HGB$_{P,Std}$，HGB$_{Ave}$，和 HGB$_{Std}$，相似程度定义为：

$$a_0=\max\left\{\frac{|\mathrm{HGB}_{C,Ave}-\mathrm{HGB}_{P,Ave}|}{\mathrm{HGB}_{Ave}},\frac{|\mathrm{HGB}_{C,Std}-\mathrm{HGB}_{P,Std}|}{\mathrm{HGB}_{Std}}\right\},\tag{1}$$

其中：a_0 越小相似度越高，本文选择满足 $a_0<0.1$ 的 50 个划分用于建模。

2.3 MWPLS 方法

MWPLS 方法是把所有位置、所有长度的波段进行比较，其参数归结如下：（1）起点波长（BW）及其编号；（2）波长个数（NW）；（3）PLS 因子数（F）[11-12]。本文在 400～1 100 nm 中进一步优选波段，波长间隔为 2 nm，共 351 个波长点，因此，MWPLS 方法的起点波长编号和 NW 都可设置为 1～351，但是为了减少计算量而不失代表性，NW 的设置改进如下：从 1 连续到 100，从 102 到 200 间隔 2，从 205 到 350 间隔 5，以上参数设置对应 42 436 个波段；另外，PLS 因子数 F 设置为 1～30。对于不同波段和 F 分别建立 PLS 模型，按照 M-SEP$_{Ave}$最小选出最优 PLS 因子数，得到每个波段的 PLS 模型。本文采用 Matlab7.6 软件构建上述参数循环的 MWPLS 算法平台。在此平台上，可以得到每个波段的 PLS 模型、进一步遴选出全局最优波段。

在实际仪器设计中，往往对波段位置和波长个数有一定的约束条件（比如成本、材料性能等），全局最优波段有时不能满足实际要求，因此选择相对于波段位置和波长个数的局部最优模型具有重要的实际意义，这也是 MWPLS 方法需要改进之处。在本文中，为了考察波长个数与预测效果的关系，固定波长个数 NW，按照 M-SEP$_{Ave}$ 最小选出该 NW 对应的局部最优模型，同时得到对应的其余参数 BW、F；而为了考察波段位置与预测效果的关系，固定起点波长 BW，按照 M-SEP$_{Ave}$ 最小选出该 BW 对应的局部最优模型，同时得到对应的其余参数 NW、F。

3 结果与讨论

3.1 可见光、短波近红外、长波近红外等 6 个波段的比较

把全谱 400～2 498 nm 按照可见光、短波近红外、长波近红外等划分如下：（1）可见光区 400～780 nm、（2）短波近红外区 780～1 100 nm、（3）长波近红外区 1 100～2 498 nm、（4）全近红外区 780～2 498 nm、（5）可见-短波近红外区 400～1 100 nm，和全谱 400～2 498 nm 共有 6 种情形，分别建立 PLS 模型，预测效果如表 1 所示。

表 1　可见光、短波近红外、长波近红外等 6 个波段的模型预测效果

Tab. 1　Prediction effects of 6 models corresponding to visible region, short-NIR region, long-NIR region, and so on

Waveband /nm	F	M-SEP$_{Ave}$	M-SEP$_{Std}$	M-RP$_{Ave}$	M-RP$_{Std}$
400～780	12	4.08	0.43	0.974	0.006
780～1 100	8	3.13	0.37	0.984	0.004
1 100～2 498	15	8.28	1.01	0.882	0.031
780～2 498	10	5.35	0.60	0.953	0.011
400～1 100	8	2.82	0.37	0.987	0.004
400～2 498	6	4.17	0.45	0.972	0.006

从表 1 可以看出，可见区和短波近红外区的模型效果明显优于长波近红外区，这是由于全血样品是深红色，靠近红光的区域有着更为显著的光谱信息。可见-短波近红外区 400～1 100 nm 的预测效果和稳定性优于其它所有模型。因此，可见-短波近红外区 400～1 100 nm 可以作为人体全血 HGB 的信息波段。

3.2　用 MWPLS 方法进一步优选波段

采用 MWPLS 方法在 400～1 100 nm 中进一步筛选波段。按照 2.3 节的方法，得到波长个数 NW、起点波长 BW 对应的局部最优模型的 M-SEP$_{Ave}$，分别如图 3、图 4 所示，可以看出，全局最优模型的 NW、BW 分别为 492、200 nm，相应波段为 492～890 nm，预测精度（M-SEP$_{Ave}$，M-R$_{P,Ave}$）和模型稳定性（M-SEP$_{Std}$，M-R$_{P,Std}$）都比 400～1 100 nm 有改善，结果如表 2 所示，为了便于比较，波段 400～1 100 nm 的 PLS 模型效果也列在表 2 中。

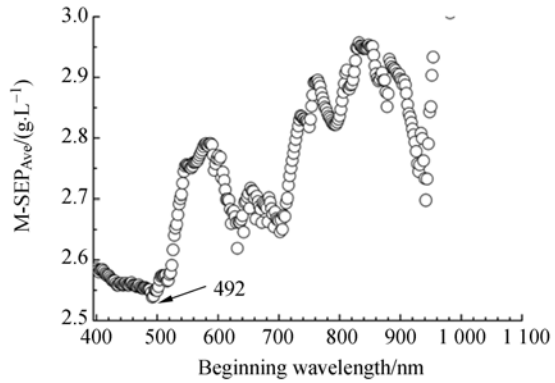

图 3　每个起点波长对应的局部最优模型的 M-SEP$_{Ave}$

Fig. 3　M-SEP$_{Ave}$ of local optimal model corresponding to each beginning wavelength

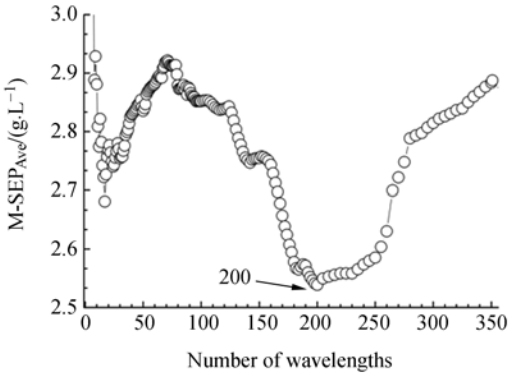

图 4　每个波长个数对应的局部最优模型的 M-SEP$_{Ave}$

Fig. 4　M-SEP$_{Ave}$ of local optimal model corresponding to each number of wavelengths

表 2　最优波段与可见-短波近红外波段的预测效果

Tab. 2　Prediction effect corresponding to optimal waveband and VIS-short NIR region

Waveband /nm	F	M-SEP$_{Ave}$	M-SEP$_{Std}$	M-R$_{P,Ave}$	M-R$_{P,Std}$
492～890	9	2.54	0.29	0.989	0.003
400～1 100	8	2.82	0.37	0.987	0.004

3.3　具有等效性的模型空间

在上一节中，根据 M-SEP$_{Ave}$ 最小（2.54 g·L^{-1}）筛选出全局最优波段为 492～890 nm，但是从统计学角度看，预测精度有轻微波动的模型是等效的，这是因为建模样品存在随机性和局限性。为此，对最优精度允许适当幅度的浮动（本文以 1% 为例，设置上浮的幅度），即将 M-SEP$_{Ave}$ 从 2.54 g·L^{-1} 上浮 1% 到 2.56 g·L^{-1}，对应的波段都是允许的等效最优波段，由此得到包含 77 个波段的等效模型空间。这 77 个等效最优波段的起点波长（BW）从 436～502 nm，终点波长从 858 到 898 nm，波长个数（NW）从 382～422 nm，它们的跨度是 436～898 nm，公共部分是 502～858 nm。77 个等效最优波段的位置如图 5 所示。因此，波段 436～898 nm 包含了 HGB 的充分信息，而 HGB 的定量信息主要集中在波段 502～858 nm。

考虑到分析样品是复杂体系，样品光谱中存在大量的干扰，在实际分析过程中，分析波段往往要宽于被测成分的吸收波段，除了吸收波段外还应该包括一些补偿波段用于克服噪音干扰。本文基于预测误差等效，得到了补偿波段的多种选择，从而得到最优波段的多种选择，可望解决由于仪

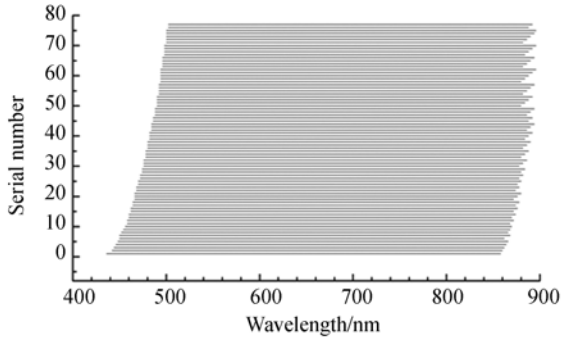

图 5　77 个等效最优波段的位置

Fig. 5　Positions of 77 equivalent optimal wavebands

器设计的实际条件约束(如成本、材料性能等)对波段位置和波长个数的限制。根据实际情况,上浮的幅度还可以做出相应的调整。

3.4　模型检验

以 492～890 nm 为例进行模型检验。即,基于建模样品的数据和已经确定的模型参数(波段和 PLS 因子数),计算出每个波长对应的 PLS 回归系数,再利用得到的 PLS 回归系数和检验样品的光谱计算出检验样品的 HGB 预测值。70 个检验样品的预测值与实测值的比较如图 6 所示,检验效果 V-SEP、V-RP 和 V-RSEP 分别为 2.58 g·L^{-1}、0.988 和 1.97%。结果表明,样品的 HGB 预测值与临床实测值有很高的吻合精度,可望应用于临床。

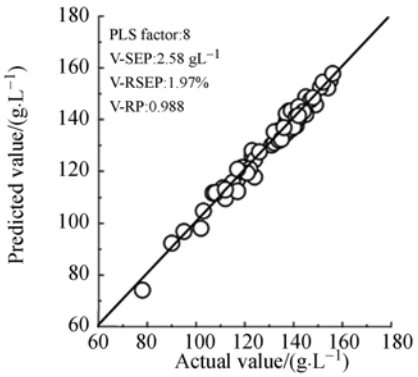

图 6　70 个检验样品预测值与实测值比较

Fig. 6　Comparison of predicted values and the actual values for 70 validation samples

4　结　论

利用可见-近红外光谱和 MWPLS 方法建立人类全血 HGB 的无试剂定量分析模型,并进行了具有稳定性的、高信噪比的波段优选。基于随机性、相似性和稳定性,提出了一种新的模型评价体系。对 MWPLS 方法进行了改进,不仅筛选了全局最优波段、局部最优波段(相对于波段位置和波长个数),还得到包含 77 个等效波段的模型空间,可以解决由于仪器设计的实际条件约束(如成本、材料性能等)对波段位置和波长个数的限制。实验结果表明,可见-短波近红外波段 400～1 100 nm 可以作为人体全血 HGB 的信息波段,从波段 400～1 100 nm 中选出全局最优波段为 492～890 nm,其检验效果 V-SEP、V-RP 和 V-RSEP 分别为 2.58 g·L^{-1}、0.988 和 1.97%,样品的 HGB 预测值与临床实测值有很高的吻合精度,可望应用于临床。

本文提出的模型评价体系避免了模型评价的失真,获得了稳定可靠的模型。其方法框架和算法平台具有普遍性,可以应用于其他的光谱分析。

参考文献:

[1] 陈星旦. 近红外光谱无创生化检验的可能性[J]. 光学 精密工程,2008,16(5):759-763.
CHEN X D. Possibility of noninvasive clinical biochemical examination by near infrared spectroscopy [J]. *Opt. Precision Eng.*, 2008, 16(5):759-763. (in Chinese)

[2] 谢军,潘涛,陈华舟,等. 血糖近红外光谱分析的 Savitzky-Golay 平滑模式与 PLS 因子数的联合优选[J]. 分析化学 ,2010, 38(3):342-346.
XIE J, PAN T, CHEN H ZH, *et al.*. Joint optimization of Savitzky-golay smoothing models and partial least squares factors for near-infrared spectroscopic analysis of serum glucose [J]. *Chinese Journal of Analytical Chemistry*, 2010, 38(3): 342-346. (in Chinese)

[3] 陈洁梅,潘涛,陈星旦. 二阶导数光谱预处理在用 FTIR/ATR 方法定量测定葡萄糖-6-磷酸和果糖-6-磷酸中的应用[J]. 光学 精密工程, 2006, 14(1): 127.
CHEN J M, PAN T, CHEN X D. Application of second derivative spectrum prepares in quantification measuring glucose-6-phosphate and fructose-6-phosphate using a FTIR/ATR method [J]. *Opt. Precision. Eng.*, 2006,14 (1) :127. (in Chinese)

[4] ISTVAN V N, KAROLY J K, JANOS M J, *et*

al.. Application of near infrared spectroscopy to the determination of haemoglobin [J]. *Clinica Chimica Acta*, 1997, 264(1):117-125.

[5] LEE Y, LEE S, IN J Y, *et al.*. Prediction of plasma hemoglobin concentration by near infrared spectroscopy [J]. *J Korean Med Sci*, 2008, 23(4): 674-677.

[6] SHAN X Q, CHEN L G, YUAN Y, *et al.*. Quantitative analysis of hemoglobin content in polymeric nanoparticles as blood substitutes using Fourier transform infrared spectroscopy [J]. *J Mater Sci*, 2010, 21(1): 241-249.

[7] 曹璞，潘涛，陈星旦. 小型近红外玉米蛋白质成分分析仪器设计的波段选择[J]. 光学 精密工程, 2007, 15(12):1952-1958.
CAO P, PAN T, CHEN X D. Choice of wave band in design of minitype near-infrared corn protein content analyzer [J]. *Opt. Precision Eng.*, 2007, 15(12):1952-1958. (in Chinese)

[8] 黄富荣，潘涛，张甘霖，等. 应用近红外漫反射光谱快速测定土壤锌含量[J]. 光学 精密工程, 2010, 18(3): 586-592.
HUANG F R, PAN T, ZHANG G L, *et al.*. Rapid measurement of zinc contents in soils by near-infrared diffuse reflectance spectroscopy [J]. *Opt. Precision Eng.*, 2010, 18(3): 586-592. (in Chinese)

[9] PAN T, HASHIMOTO A, KANOU M. Development of a quantification system of ionic dissociative metabolites using an FT-IR/ATR method [J]. *Bioprocess and Biosystems Engineering*, 2003, 26(2): 133-139.

[10] 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]. *Analytical Chemistry*, 2002, 74(14):3555-3565.

[11] CHEN H Z, PAN T, CHEN J M, *et al.*. Waveband selection for NIR spectroscopy analysis of soil organic matter based on SG smoothing and MW-PLS methods [J]. *Chemometrics and Intelligent Laboratory Systems*, 2011, 107(1):139-146.

[12] PAN T, CHEN Z H, CHEN J M, *et al.*. Near-infrared spectroscopy with waveband selection stability for the determination of COD in sugar refinery wastewater [J]. *Analytical Methods*, 2012, 4(4):1046-1052.

作者简介：

刘振尧（1983－），男，山东潍坊人，博士研究生，2010 年于西南大学获得硕士学位，主要从事近红外光谱应用以及相关算法方面的研究。E-mail: q183333348@gmail.com

导师简介：

潘　涛（1964－），男，广西宜州人，教授、博士生导师，日本生物信息工学博士，主要从事应用光谱、生物医学工程、应用数学等方面的研究。E-mail: tpan@jnu.edu.cn