基于近红外光谱的SG-MSC-MC-UVE-PLS算法在全血血红蛋白浓度检测中的应用

孙代青[1,2]，谢丽蓉[1*]，周　延[2]，郭煜涛[1]，车少敏[2]

1. 新疆大学电气工程学院，新疆 乌鲁木齐　830047
2. 西安交通大学能源动力工程学院，陕西 西安　710049

摘　要　为提高全血血红蛋白浓度预测模型的预测精度，基于近红外光谱分析，首先对原始全血透射光谱数据分别进行均值中心化、标准化、标准正态变量变换(SNV)、多元散射校正(MSC)以及Savitzky-Golay(SG)卷积平滑结合MSC的预处理操作，最终选择预处理效果最好的SG-MSC方法作为数据预处理方法，其最大相关系数达到0.944 1。对SG平滑的平滑窗口宽度进行讨论，找出平滑效果最好的窗口宽度为27。数据预处理消除了全血吸收光谱的基线失真，提高了全血吸收光谱数据的信噪比。将190个样本(190个血红蛋白浓度对应的透射光谱数据)分为具有相近血红蛋白浓度分布的校正集和测试集，其中校正集为143个样本(对应血红蛋白浓度分布为10.6～17.3 g·dL^{-1})，测试集为47个样本(对应血红蛋白浓度分布为10.3～17.3 g·dL^{-1})，确保建立模型的适用性。对校正集数据预处理后利用蒙特卡洛无信息变量消除(MC-UVE)方法对其进行波长变量选择，剔除含信息量少的波长点，提高含信息量多的波长占比。设置蒙特卡洛迭代次数为1 000，最终从全血吸收光谱的700个波长变量中筛选出191个波长变量用于建立全血血红蛋白浓度偏最小二乘(PLS)回归模型。对比分析原始全血透射光谱全谱PLS模型、原始全血吸收光谱全谱PLS模型、预处理全血吸收光谱全谱PLS模型、SG-MSC-MC-UVE-PLS模型以及已有二阶导数PLS模型的模型效果，表明基于SG-MSC-MC-UVE-PLS算法的全血血红蛋白浓度预测模型效果较其他模型效果更优，预测相关系数由0.676 3提高到0.979 1，预测集均方根误差由0.898 1减小到0.220 3，最大绝对误差由2.426 1减小到0.411 2。同时，利用MC-UVE方法进行波长变量选择，在保证预测精度的前提下，筛选出建模的波长个数更少，有利于提高模型计算效率。研究结果表明，SG-MSC-MC-UVE-PLS方法能够提高全血吸收光谱信号的信噪比，简化模型结构，提高模型的预测精度和计算效率，对推动血红蛋白浓度检测技术的发展具有进步意义。

关键词　近红外光谱；全血血红蛋白浓度预测；光谱信号预处理；无信息变量消除
中图分类号：O657.33　　**文献标识码**：A　　**DOI**：10.3964/j.issn.1000-0593(2021)09-2754-05

引　言

血红蛋白(Hemoglobin)是生物化学和生物医学研究中最重要的成分之一[1-2]，它大约占红细胞的96%，承担着将氧气通过循环系统运输到器官的重要责任，同时血红蛋白浓度的测定也是临床上应用最广的检查项目。目前血红蛋白浓度检测方法主要分为两类，一种为有化学试剂类型[3]，另一种为无化学试剂类型[1-2,4-6]。有化学试剂的检测方法通常因为所使用的化学试剂对人体和环境有害而使其应用场所受到限制。无化学试剂方法测量精度很低，所需血液量较多(20 μL)[6]，而且其价格及其昂贵[7]。

近年来，近红外光谱技术不断发展，其应用领域也越来越广泛。基于近红外光谱技术能反映组织细胞生理病理信息的特点，对蓝莓果渣花色苷含量进行了测定[8]。同时，基于近红外光谱的无创、快速等特点，将近红外光谱与偏最小二乘(PLS)回归相结合用以检测血液中的不同成分含量的方法广受欢迎[5,9]。偏最小二乘回归是建立光谱信号和浓度关系

收稿日期：2020-09-08，**修订日期**：2021-01-10
基金项目：国家自然科学基金项目(51667021)和新疆维吾尔自治区区域协同创新专项(2018E02072)资助
作者简介：孙代青，1995年生，新疆大学电气工程学院硕士研究生　　e-mail：1248429197@qq.com
*通讯作者　　e-mail：wzywwwxr@163.com

的最流行的方法[10]。其中也不乏对于血红蛋白浓度的检测研究，但是，目前基于这种方法建立的血红蛋白浓度检测模型都存在精度不高的问题，很难达到临床应用的标准，主要原因是所获取的近红外光谱数据可能包含很多背景信号，降低了光谱信号的信噪比。

为减小背景信号对光谱数据质量的影响，一阶导数[11]、二阶导数[12]、主成分分析[13]、多元散射校正(MSC)[14]等数据预处理方法被提出，然而缺少对于全血光谱数据的预处理方法、波长选择的研究。故基于近红外光谱分析，对全血光谱数据的预处理方法、波长筛选、以及全血血红蛋白浓度预测模型进行研究，为提高全血血红蛋白浓度预测精度提供一种新的思路。

1　实验部分

1.1　样本

数据集取自 Karl Norris[15]的文章。这组数据是使用 NIRSystems6500 光谱仪获得。仪器参数设置如下：波长变量为 1 100～2 498 nm，分辨率为 2 nm。样品池是带有石英窗口的直径 2 cm 的不锈钢圆柱体。将 200 μL 全血从移液管转移至样品池，使样品厚度为 0.6 mm，一共获得 190 组不同血红的蛋白浓度的全血透射光谱，所获透射光谱对应最小血红蛋白浓度为 10.3 g·dL^{-1}，最大血红蛋白浓度为 17.3 g·dL^{-1}。

1.2　样本数据集划分

为使得建立的模型具有普遍性，选用前 143 个血红蛋白浓度对应的透射光谱样本作为校正集，剩下 47 个作为验证集。经划分后的校正集透射光谱样本对应最小血红蛋白浓度为 10.6 g·dL^{-1}，最大血红蛋白浓度为 17.3 g·dL^{-1}，平均血红蛋白浓度为 13.68 g·dL^{-1}，标准差为 1.64 g·dL^{-1}；验证集透射光谱样本对应最小血红蛋白浓度为 10.3 g·dL^{-1}，最大血红蛋白浓度为 17.3 g·dL^{-1}，平均血红蛋白浓度为 13.94 g·dL^{-1}，标准差为 1.65 g·dL^{-1}。

1.3　数据预处理

首先对原始全血透射光谱取 $-\log(T)$，将其转换成吸收光谱数据，然后对原始全血吸收光谱分别进行均值中心化、标准化、SNV、MSC 以及 SG 卷积平滑结合 MSC 方法预处理操作。讨论卷积平滑与 MSC 的操作顺序对于预处理效果的影响，以及平滑窗口宽度对于 SG-MSC 平滑效果的影响，比较不同平滑窗口的降噪效果，选择降噪效果最好的一个窗口宽度作为卷积平滑窗口。对比以上几种预处理方法的降噪效果，选择表现最好的方法作为全血吸收光谱数据预处理方法。

1.4　波长筛选程序及血红蛋白浓度预测模型建立

蒙特卡洛无信息变量消除算法(Monte Carlo uninformative variable elimination，MC-UVE)是无信息变量消除方法的一种，它是基于模型变量稳定性值对无信息变量进行剔除的方法。稳定性值的绝对值越大，所对应的变量越重要，保留稳定性值大的变量，剔除稳定性值小的变量。利用此方法从预处理过的全血吸收光谱中选择出稳定性值较大的波长变量，以提高基于近红外光谱的全血血红蛋白浓度预测模型的预测精度和预测效率。

2　结果与讨论

2.1　原始全血透射光谱数据及吸收光谱

将 170 个不同全血血红蛋白浓度的近红外透射光谱数据导入 Matlab R2017a 计算原始全血透射比与全血血红蛋白浓度之间的相关系数，其相关系数曲线如图 1(a)所示。再将透射光谱数据取 $-\log(T)$，计算全血吸收度与全血血红蛋白浓度的相关系数 R^2，其曲线如图 1(b)所示。

分析图 1 可知，原始透射光谱的 R^2 最大值仅为 0.003 5，在波长 1 954 nm 处取得。相应的原始吸收光谱的 R^2 最大值也仅为 0.005 0，且只有少量的信号对应于较大(此处指大于 0.005 0)的 R^2 值。由此可见，利用原始信号建立全血血红蛋白浓度预测模型是比较困难的。

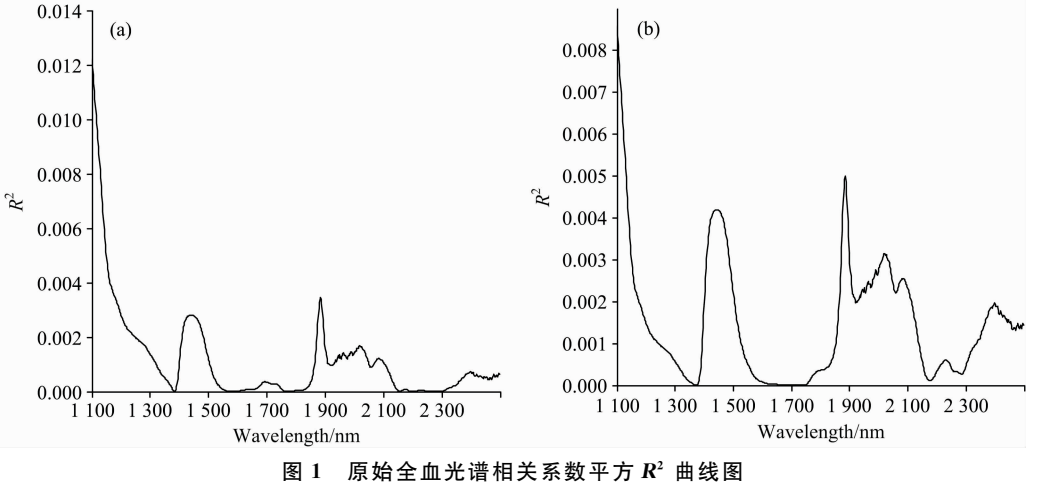

图 1　原始全血光谱相关系数平方 R^2 曲线图

(a)：原始透射；(b)：原始吸收

Fig. 1　R^2 curves for the original spectra of whole blood signals

(a)：Transmission spectrum；(b)：Absorbance spectrum

2.2 数据预处理

表 1 中展示了分别使用均值中心化、标准化、SNV、MSC 以及 SG 卷积平滑结合 MSC 对原始全血吸收光谱分别进行预处理后相关系数平方最大值 R^{2*} 的变化情况。其中，中心化、标准化、SNV 这几种预处理方法对于全血吸收光谱的平滑处理效果都不明显；单独使用 MSC 时，其处理效果也不理想，R^{2*} 值仅为 0.105 2，但在结合 SG 卷积平滑后降噪效果迅速提升，最大相关系数平方值 R^{2*} 迅速提升至 0.944 1，这是因为 MSC 在处理浆状物透射近红外光谱方面具有很好的效果，血液样本正好符合这一特征。除此之外，二阶导数能消除光谱采集过程中由于检测环境和仪器状态等因素引起的基线平移，平滑处理能消除光谱中的随机误差，提高信号的信噪比。

表 1 不同预处理方法 R^{2*} 值比较

Table 1 Comparison of R^{2*} values of different preprocessing methods

Method	原始透射	原始吸收	中心化	标准化	SNV	MSC	SG-MSC
R^{2*}	0.003 5	0.005 0	0.005 0	0.005 0	0.070 6	0.105 2	0.944 1

图 2(a)中展示了不同平滑窗口宽度下的 SG-MSC 预处理的 R^{2*} 值变化情况，图中 R^{2*} 表示最大相关系数平方值，将平滑窗口的宽度依次从 1 变化到 61，最大 R^{2*} 值为 0.944 1，此时窗口宽度为 27，得到最佳平滑参数是 2 阶导数平滑，二阶多项式和 27 个平滑点。图 2(b)展示了在上述参数设置下的 SG-MSC 方法对全血原始吸收光谱处理后各波长点处的相关系数平方值 R^2。与图 1(b)相比，R^2 迅速上升，并且较大(此处指 R^2 值大于 0.6)R^2 个数也明显增多。

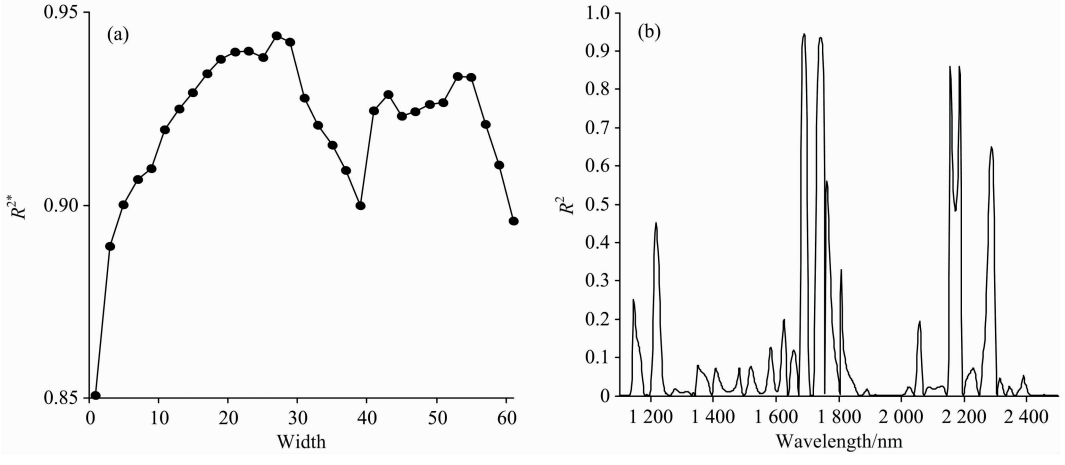

图 2 平滑窗口宽度对 SG-MSC 预处理方法 R^{2*} 值的影响

(a)：窗口宽度 *vs.* R^{2*}；(b)：SG-MSC 处理后 R^2

Fig. 2 R^{2*} *vs.* the width of the SG-MSC method

(a)：R^{2*} *vs.* the width；(b)：R^2 after preporcessing by SG-MSC

2.3 波长变量选择与结果分析

为了进一步提高预测模型的预测精度和预测效率，对预处理过的全血吸收光谱的 700 个波长进行蒙特卡洛无信息变量消除，剔除 509 个波长，剩余 191 个波长用于建立全血血红蛋白浓度回归模型。为建立稳健性好、预测能力强的血红蛋白浓度预测模型，比较分析了原始全血透射光谱 PLS 模型、原始全血吸收光谱 PLS 回归模型、SG-MSC-PLS 回归模型、SG-MSC-MC-UVE-PLS 回归模型以及二阶导数 UVE-PLS 回归模型[12]，各模型指标结果如表 2 所示，表中 NW (number of wavelengths)为筛选出的波长变量个数。

由表 2 中结果分析可知，原始全血透射光谱 PLS 模型的 R^2 比较小，且 RMSEP 很大，这也说明了直接利用原始全血透射光谱进行建模不可取。加入 SG-MSC 预处理以后，所建 PLS 模型的预测集 R^2 相比于原始透射光谱数据提高了 0.296 5，RMSEP 下降了 0.669 1，MAE 减小了 1.931 8。证明 SG-MSC 对于全血光谱数据的平滑降噪能力非常强。在此基础上，对预处理过的光谱数据进行波长选择，建立 SG-MSC-MC-UVE-PLS 模型，与 SG-MSC-PLS 模型指标相比，其 R^2，RMSEP，MAE，MRE 均优于未筛选波长的 PLS 模型，且与前人所提二阶导数 UVE-PLS 模型相比，其具有更高的 R^2 和更低的 RMSEP 值。进一步说明 SG-MSC-MC-UVE-PLS 算法可以有效降低噪声、筛选更具有价值的波长

表 2 PLS 模型预测结果

Table 2 Determination results for PLS models

Signals	NW	R^2	RMSEP	MAE	MRE
原始透射 PLS	700	0.676 3	0.898 1	2.426 1	0.140 24
原始吸收 PLS	700	0.896 1	0.464 2	1.63	0.094 2
SG-MSC- PLS	700	0.972 8	0.229 0	0.494 3	0.028 6
SG-MSC-MC-UVE-PLS	191	0.979 1	0.220 3	0.411 2	0.023 8
二阶导数+UVE+PLS	209	0.967 6	0.300 3	—	—

变量、提高预测能力和预测效率。

3 结　论

将获取的原始全血透射光谱转换成全血吸收光谱，应用偏最小二乘法建立全血血红蛋白浓度回归模型，针对原始数据相关性低的问题，对原始数据进行了光谱数据预处理；针对原始数据中无用信息成分较多问题采用了蒙特卡洛无信息变量消除方法对波长进行筛选；比较了原始数据、预处理数据、波长选择数据建立的 PLS 模型效果，得到以下结论：

(1)针对全血吸收光谱数据，通过比较均值中心化、标准化、标准正态变量变换、多元散射校正、SG 卷积平滑结合多元散射校正对全血光谱数据的预处理效果，得到最佳预处理方法为 SG 卷积预处理＋多元散射校正方法，其 R^2 为 0.944 1。

(2)对 SG-MSC 预处理方法的平滑窗口宽度对于平滑效果的影响进行研究，得到最佳参数设置为窗口宽度为 27，二阶导数平滑，二阶多项式拟合。与先进行多元散射校正再进行 SG 卷积平滑(相关系数平方值为 0.942 4)相比，卷积平滑之后再对数据进行多元散射校正处理，其相关系数平方值更大，为 0.944 1。

(3)MC-UVE 可以实现对全血吸收光谱波长变量的筛选，且其筛选的波长变量个数仅为 191 个，在模型效果更优的情况下，筛选出的波长变量更少，可以大大简化模型，提高模型效率。

(4)在全血血红蛋白浓度回归模型中，将 SG 卷积平滑、多元散射校正以及 MC-UVE 组合建立的 PLS 模型具有最优的模型效果，相比于原始全谱以及未经波长选择的 SG-MSC-PLS 模型，SG-MSC-MC-UVE-PLS 模型的模型精度更高，且筛选出的波长点更少，其模型指标 R^2 为 0.979 1，RMSEP 为 0.220 3，MAE 为 0.411 2，MRE 为 0.023 8。该模型效果与前人所提方法相比有所提高。

References

[1] Zhang S Z, Li G, Wang J X, et al. Scientific Reports, 2018, 8: 1.

[2] Wang Y Y, Li G, Wang H Q, et al. Applied Spectroscopy Reviews, 2019, 54(9): 736.

[3] YE Cui-qing, LIANG Qi-long, HUANG Jie-wen, et al(叶翠清，梁其隆，黄洁雯，等). Shenzhen Journal of Integrated Traditional Chinese and Western Medicine(深圳中西医结合杂志), 2019, 29(10): 74.

[4] Li G, Xu S J, Zhou M, et al. Spectroscopy Letters, 2017, 50(3): 164.

[5] Liu H Y, Peng F L, Hu M L, et al. Journal of Electrical and Computer Engineering, 2020: ID3034260.

[6] Yuan J Z, Lu Q P, Wang J L, et al. Chinese Journal of Analytical Chemistry, 2017, 45(9): 1291.

[7] Lee J, Song J, Choi J-H, et al. Scientific Reports, 2020, 1: 10.

[8] ZHANG Li-juan, XIA Qi-le, CHEN Jian-bing, et al(张丽娟，夏其乐，陈剑兵，等). Spectroscopy and Spectral Analysis(光谱学与光谱分析), 2020, 40(7): 2246.

[9] Abd Rahima I M, Rahim H A, Ghazali R, et al. Jurnal Teknologi, 2016, 78(7-4): 85.

[10] Zifarelli A, Giglio M, Menduni G, et al. Analytical Chemistry, 2020, 11035: 11043.

[11] Beumers P, Engel D, Brands T, et al. Chemometrics and Intelligent Laboratory Systems, 2018, 172: 1.

[12] Zhou Y, Zheng C L, Cao H, et al. Biochemical and Biophysical Research Communications, 2012, 420(1): 205.

[13] LI Shang-ke, DU Guo-rong, LI Pao, et al(李尚科，杜国荣，李　跑，等). Food Research and Development(食品研究与开发), 2020, 41(17): 144.

[14] QIU Yan, ZHANG Xue-qin, GUO Yu-jun, et al(邱　彦，张血琴，郭裕钧，等). High Voltage Engineering(高电压技术), 2019, 45(11): 3587.

[15] Kuenstner J, Norris K. Journal of Near Infrared Spectroscopy, 1995, 3(1): 11.

Application of SG-MSC-MC-UVE-PLS Algorithm in Whole Blood Hemoglobin Concentration Detection Based on Near Infrared Spectroscopy

SUN Dai-qing[1,2], XIE Li-rong[1*], ZHOU Yan[2], GUO Yu-tao[1], CHE Shao-min[2]
1. School of Electrical Engineering, Xinjiang University, Urumqi 830047, China
2. School of Energy & Power Engineering, Xi'an Jiaotong University, Xi'an 710049, China

Abstract In order to improve the accuracy of the whole blood hemoglobin (Hb) concentration prediction model, the original whole blood transmission spectrum signals were first preprocessed by using centering, auto scaling, standard normal variate (SNV), multiplicative scatter correction (MSC), and Savitzky-Golay (SG) smoothing combined with MSC. And the best preprocessing effect was obtained with a R2 value of 0.9441 by using SG smoothing combined with MSC. The width of the SG smoothing window was discussed, and the optimal width is 27. The baseline shift of the whole blood absorbance signals was eliminated, and the signal-to-noise ratio was improved after data preprocessing. The 190 samples were divided into a calibration set (corresponding Hb concentrations from 10.6 to 17.3 $g \cdot dL^{-1}$) of 143 samples and a validation set (corresponding Hb concentrations from 10.3 to 17.3 $g \cdot dL^{-1}$) of 47 samples. The model's applicability was ensured when two sets have a similar distribution and range of Hb concentrations. And then, the Monte Carlo uninformative variable elimination (MC-UVE) was used to select the informative wavelength, which simplified the model structure and increased the proportion of useful wavelengths. When the Monte Carlo iteration number was 1000, 191 wavelength points were selected from the 700 wavelengths of the whole blood absorbance spectrum to build the whole blood Hb concentration partial least squares (PLS) model. Finally, a comparison was performed among the model based on the original whole blood transmission spectrum, the model based on the whole blood absorbance spectrum, the SG-MSC-PLS model, the SG-MSC-MC-UVE-PLS model and an existing model. In addition to this, the number of selected wavelengths based on MC-UVE was much smaller than the total number, but the predictive effect was much better, which was beneficial to improve the calculation efficiency of the model. The results indicate that the SG-MSC-MC-UVE-PLS method effectively increases the signal-to-noise ratio of the whole blood absorption spectrum signal and simplifies the model. Besides, our procedure's prediction accuracy and calculation efficiency of the model was improved by our procedure, which has reference significance for the development of hemoglobin concentration detection technology.

Keywords Near-infrared spectroscopy; Whole blood hemoglobin concentration detection; Spectral signals preprocessing; Uninformed variable elimination

(Received Sep. 8, 2020; accepted Jan. 10, 2021)

* Corresponding author