考虑多种因素的近红外光谱血糖预测模型对比

王晓飞*，张欣怡，徐馨荷
北京信息科技大学仪器科学与光电工程学院，北京 100192

摘要 以血糖浓度为例，采用将动态光谱提取数据和非测量组分影响因素一同纳入预测模型的方式来提高血糖测量系统中的精度。通过支持向量算法建立血糖预测的模型，建模结果表明，考虑多因素模型的预测值优于未考虑非测量组分模型中的预测值。与后者相比，前者的相关系数达到 0.9627，提高了 14.23%；均方根误差为 0.13，减少了 43.12%；相对误差在 10%范围内的样本数量增加 8.33%。

关键词 医用光学；近红外光谱；血糖；动态光谱；支持向量机
中图分类号 Q819 **文献标识码** A **doi**：10.3788/LOP56.041701

Comparison of Multi-Factor-Considered Blood Glucose Prediction Models by Near-Infrared Spectroscopy

Wang Xiaofei*, Zhang Xinyi, Xu Xinhe
School of Instrumentation Science and Optoelectronic Engineering, Beijing Information Science and Technology University, Beijing 100192, *China*

Abstract Taking the blood glucose concentration as an example, the accuracy of the blood glucose measurement system is improved by means of the simultaneous incorporation of the extraction data form dynamic spectra and the influencing factors of non-measured components into the prediction model. The blood glucose prediction model is established through the support vector machine algorithm. The modeling results show that the prediction value from the multi-factor-considered model is superior to that from the non-measurement-component-considered model. The correlation coefficient of the former is 0.9627, higher by 14.23%, the root mean square error is 0.13, reduced by 43.12%, and the number of samples with a relative error in the range of 10% is higher by 8.33%, compared with those of the latter.

Key words medical optics; near-infrared spectroscopy; blood glucose; dynamic spectrum; support vector machine
OCIS codes 170.1470; 170.6510; 300.1030; 300.6170; 300.6340

1 引 言

基于光谱检测技术的血液成分检测[1-4]因其无创伤、快速和无污染等特点成为研究的热点，在血液无创检测中有广泛的应用。无创的血糖浓度测量有多种方法[5]，其中光谱法有较好的预测结果。将光源照射人体部位如手指、耳垂等，用光电探测器测得光强。由于心脏的搏动，动脉血管中的血液吸光度呈周期性变化，因此接收到的光强值呈现周期性变化。进一步处理光谱可以预测得到血糖浓度。在无创血液检测的过程当中，受到个体差异和血液中其他成分等多种因素影响，利用近红外光谱预测血液的模型达不到临床精度要求。基于包含多种因素的测量理论[6-8]，在近红外光谱无创血液测量过程中，需将多种影响因素考虑在内建立模型，以提高测量精度。本文以血糖浓度值为例，在测量过程中得到动态光谱数据，并且将甘油三酯、白蛋白、球蛋白、胆固醇、年龄等 5 种非测量组分数据考虑在内，用这些数据进行建模，得到血糖浓度的预测值。

收稿日期：2018-09-04；**修回日期**：2018-09-09；**录用日期**：2018-09-18
基金项目：北京市自然科学基金(7172035)
* **E-mail**：wangxiaofei@bistu.edu.cn

2 实验方案

实验方案如图 1 所示，光线照射手指，由光谱仪接收到近红外动态光谱信息并保存至计算机中。实验共有 239 位受试者，在采集数据前，受试者在情绪、心理等各方面均保持稳定状态。每个受试者的测量时间为 20 s，光谱仪积分时间为 10 ms，单波长采样点数为 2000 个点，将光谱仪采集到的数据保存到计算机中。在使用光谱仪采集数据的同时，抽取受试者的血液进行生化分析得到血糖真值，及对血糖有影响的甘油三酯、胆固醇、球蛋白、白蛋白等成分的浓度真值，并记录受试者的年龄。将获得的数据进行处理并建模，根据模型可得到血糖浓度的预测值。

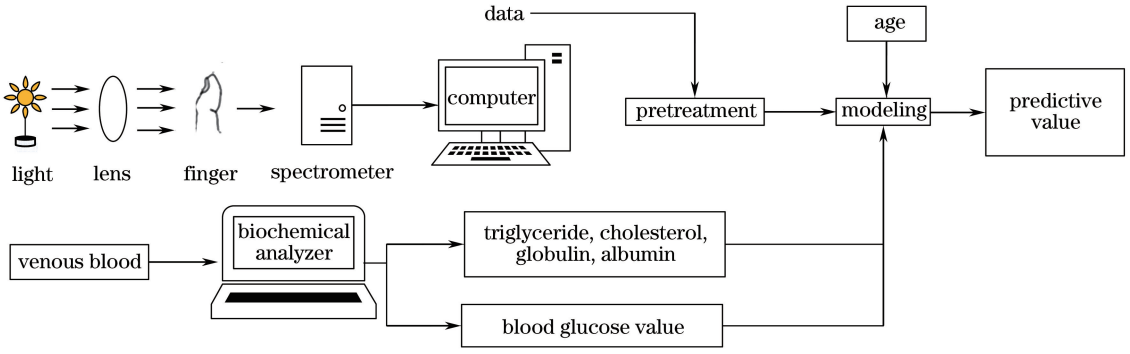

图 1 血糖测量实验方案

Fig. 1 Experiment scheme for measuring blood glucose

血糖浓度预测模型的建立基于支持向量机(SVM)[9]的方法，SVM 在解决小样本、非线性以及高维模式识别中表现出特有的优势。可根据有限的样本信息在模型的复杂性和学习能力之间寻求最佳折中，以求获得最好的推广能力。模型建立基于 Libsvm-3.22 工具箱，Libsvm 是台湾大学林智仁(Lin Chih-Jen)教授等设计开发的一个易于使用且快速有效的 SVM 模式识别与回归的软件包。

3 实验过程及结果

在建立模型的过程中将非测量组分影响因素也作为自变量输入模型，以减小其对测量系统的影响。采用非线性算法 SVM 来建立将非测量组分考虑在内的血糖预测模型，并与未将非测量组分考虑在内的血糖预测模型进行对比。

将实验测得的样本进行筛选[10]，剔除无效数据，并对光谱数据进行预处理[11-12]，得到有效样本数 192 组，每个样本的有效光谱数据为 606 个。建立模型时，血糖真值数据作为因变量矩阵 $\boldsymbol{Y}=[\boldsymbol{y}_1\ \ \boldsymbol{y}_2\ \cdots\ \boldsymbol{y}_{192}]$输入模型。若不考虑非测量组分，模型的自变量输入仅有光谱数据 $\boldsymbol{X}=[\boldsymbol{x}_{i1}\ \boldsymbol{x}_{i2}\cdots\boldsymbol{x}_{i606}]$，其中 i 表示样本数量，$i=1,2,\cdots,192$。考虑非测量组分时，将甘油三酯、胆固醇、白蛋白、球蛋白和年龄 5 个影响因素和动态光谱数据作为自变量矩阵 $\boldsymbol{X}=[\boldsymbol{x}_{i1}\ \boldsymbol{x}_{i2}\cdots\ \boldsymbol{x}_{i611}]$，其中 $i=1,2,\cdots,192$；因变量输入均为血糖数据；模型输出均为血糖预测值。

将建模数据按血糖浓度进行从小到大排序，以保证训练集血糖样本浓度覆盖预测集的血糖样本浓度。按照 3∶1的比例划分训练集和预测集，选取 144 例样本进行建模，48 例样本进行预测。考虑和未考虑非测量组分的模型，均按此标准选择训练集和预测集样本数据。

3.1 SVM 建立考虑非测量组分的血糖校正模型

SVM 模型建立过程如下。

第一步：对输入的自变量和因变量数据进行归一化操作。将其包含数据的概率分布统一归纳到上述区间，使其处于同一数量级，提高训练效率。归一化公式为

$$y_i^*=\frac{y_i-y_{\min}}{y_{\max}-y_{\min}},i=1,2,\cdots,192。\tag{1}$$

第二步：寻找合适的惩罚因数 c 和核函数参数 g。核函数参数 g 可由函数的宽度参数 σ 表示，即

$$g=\frac{1}{2\sigma^2}。\tag{2}$$

对于血糖模型来说，无法在训练模型之前得知血糖模型的最优参数值。因此需通过网格法来寻找最优的参数。网格法参数寻优的原理是通过对所有可能的参数在一定的范围内进行网格划分，并且遍历网格中所有点进行穷举，一一进行实验，找到分类准确率最高时所对应的参数值。根据该方法，得到 $c=128$，$g=0.0078125$。

第三步：寻找合适的核函数。在实验过程中，分别使用线性核函数和径向基核函数（RBF 核函数）对考虑非测量组分的数据建立矫正模型，模型结果如图 2、图 3 所示。

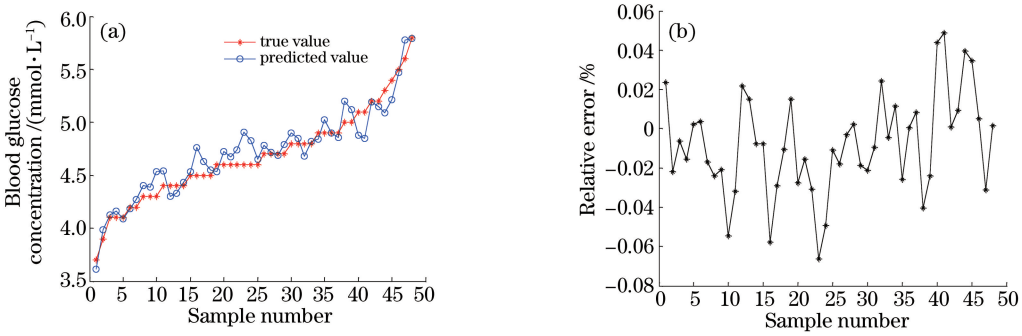

图 2 考虑非测量组分的 RBF 建模结果。(a)预测值和真值；(b)相对误差

Fig. 2 RBF modeling results with non-measurement-component considered. (a) Predicted values and true values; (b) relative errors

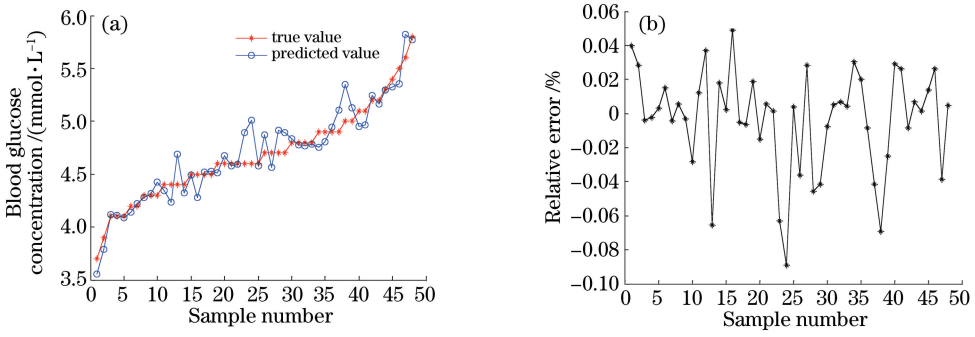

图 3 考虑非测量组分线性核函数建模结果。(a)预测值和真值；(b)相对误差

Fig. 3 Linear kernel function modeling results with non-measurement-component considered. (a) Predicted values and true values; (b) relative errors

线性核函数建立的模型预测值和血糖真值的相关系数为 0.9437，均方根误差为 0.14。计算得出 RBF 核函数建立的模型预测集相关系数为 0.9627，预测集均方根误差（RMSEP）为 0.13。相较线性核函数，预测集相关系数增加了 2.01%，预测集均方根误差减小了 7.14%。对比这些指标可知，对于将非测量组分考虑在内的血糖模型来说，RBF 核函数建立的模型质量更高，稳健性更好。

3.2 SVM 建立未考虑非测量组分的血糖校正模型

根据上述数据，对于血糖模型来说，RBF 核函数建立模型稳健性更好，所以在不考虑非测量组分数据时也使用 RBF 核函数建立校正模型。重复上述建模步骤，选出惩罚系数 $c=32$，$g=0.000976$，得出模型预测值。模型结果如图 4 所示。

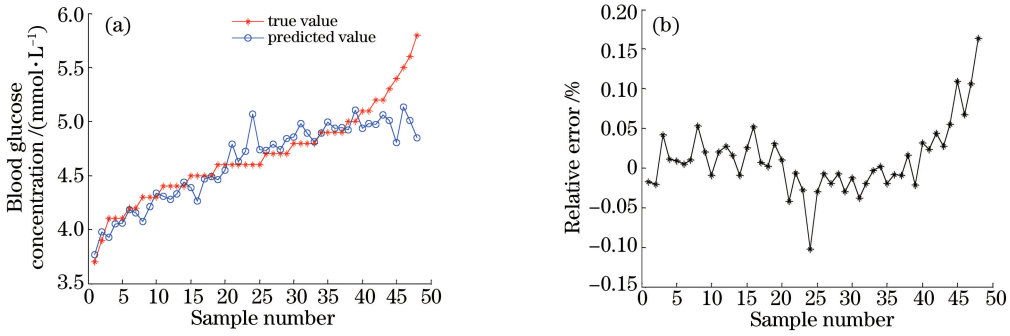

图 4 未考虑非测量组分的模型结果。(a)预测值和真值；(b)相对误差

Fig. 4 Modeling results without non-measurement-component considered. (a) Predicted values and true values; (b) relative errors

根据模型预测结果计算可得，未将非测量组分考虑在内建立的模型预测值和血糖真值的校正集相关系数为 0.9344；校正集均方根误差（RMSEC）为 0.17；预测集相关系数为 0.8655；预测集均方根误差为0.23。从图 4(b)可看出，误差在 10%范围内的有 44 个样本点，为总样本数的 91.67%。

3.3 结果分析

对于建立的校正模型，均采用相关系数 R、校正集均方根误差、预测集均方根误差和相对误差 E 这 4 个指标来评价模型，其中相关系数反映了预测值和理论值的相似程度，均方根误差和相对误差反映了模型精度。

所建模型相对误差如图 5 所示，考虑非测量组分建模时，预测值和真实值相对误差在 10%范围内的样本点数有 48 个，为总样本数的 100%。未考虑非测量组分建模时，预测值和真实值相对误差在 10%范围内的样本点数有 44 个，为总样本数的 91.67%。

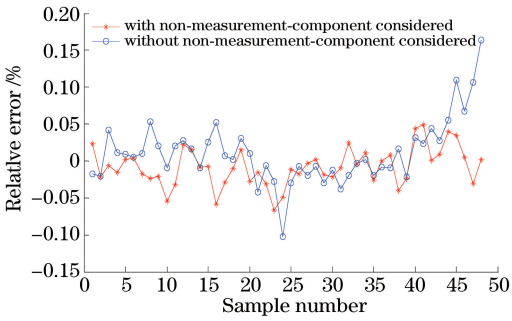

图 5 SVM 建立模型的相对误差

Fig. 5 Relative errors of SVM model

非测量组分考虑在内的模型和未将非测量组分考虑在内的模型均使用 SVM 建模，各模型指标如表 1 所示。

表 1 各模型指标

Table 1 Model parameters

Modeling	Training set		Prediction set	
	R	RMSEC	R	RMSEP
With non-measurement-component considered	0.9993	0.02	0.9627	0.13
Without non-measurement-component considered	0.9344	0.17	0.8655	0.23

对比将胆固醇、甘油三酯、白蛋白、球蛋白、年龄 5 种非测量组分考虑在内和未将非测量组分考虑在内建立的模型，前者的预测结果均优于后者的预测结果。前者比后者预测集相关系数提高 14.23%，预测集均方根误差减少 43.12%，相对误差在 10%范围内的样本数量多 8.33%。相较于仅使用光谱数据建模，将 5 种非测量组分考虑在内的血糖测量系统的预测精度显著提高。

4 结 论

通过对 192 个样本的光谱数据以及非测量组分进行分析，并结合生化分析结果，使用 SVM 的方法，分别建立了考虑和未考虑非测量组分的血糖浓度预测模型。通过对比模型建立结果可知，将非测量组分考虑在内的预测结果相关系数增大，均方根误差减小，相对误差减小，预测精度优于未将非测量组分考虑在内的预测结果。在血液成分的无创测量中可采用此种方法来提高测量精度。

参 考 文 献

[1] Yang X, Ji Z, Yang L, *et al*. The principles and research status of noninvasive glucose detection based on near-infrared spectrum[J]. Journal of Biomedical Engineering, 2013, 30(1): 204-207.
杨星, 季忠, 杨力, 等. 基于近红外光谱法的无创血糖检测原理与研究现状[J]. 生物医学工程学杂志, 2013, 30(1): 204-207.

[2] Wang X F, Zhao W J. Measurement of multi-wavelength pulse oxygen saturation based on dynamic spectroscopy [J]. Spectroscopy and Spectral Analysis, 2014, 34(5): 1323-1326.
王晓飞, 赵文俊. 基于动态光谱法的多波长脉搏血氧饱和度测量[J]. 光谱学与光谱分析, 2014, 34(5): 1323-1326.

[3] Zhang Y, Ni J S, Zhang Y Z, *et al*. Tissue intrinsic fluorescence spectrum recovery algorithm and its application in diabetes screening[J]. Chinese Journal of Lasers, 2018, 45(7): 0707001.
张洋, 倪敬书, 张元志, 等. 组织固有荧光光谱复原算法及其在糖尿病筛查中的应用研究[J]. 中国激光, 2018, 45(7): 0707001.

[4] Li D M, Jia S H. Application of BP artificial neural network in blood glucose prediction based on multi-spectrum[J]. Laser & Optoelectronics Progress, 2017, 54(3): 031703.
李东明, 贾书海. 基于多光谱应用 BP 人工神经网络预测血糖[J]. 激光与光电子学进展, 2017, 54(3): 031703.

[5] Sun K, Zhou H, Yang Y K, *et al*. Research advances in blood glucose monitoring system [J]. Chinese Journal of Lasers, 2018, 45(2): 0207003.
孙凯, 周华, 杨膺琨, 等. 血糖监测系统的研究进展[J]. 中国激光, 2018, 45(2): 0207003.

[6] Li G, Li Z, Wang X F, *et al*. Evolution of measure mode and proposition of "$M+N$" theory[J]. Journal of Beijing Information Science & Technology University, 2013, 28(2): 9-13.
李刚, 李哲, 王晓飞, 等. 测量模式的演进与"$M+N$"理论的提出[J]. 北京信息科技大学学报(自然科学版), 2013, 28(2): 9-13.

[7] Xu X H, Wang X F. Near-infrared spectra noninvasive measurement method of glucose based on the "$M+N$" theory[J]. Research and Exploration in Laboratory, 2018, 37(2): 5-9, 14.
徐馨荷, 王晓飞. 基于"$M+N$"理论的近红外光谱血糖无创测量方法[J]. 实验室研究与探索, 2018, 37(2): 5-9, 14.

[8] Li L, Wang X F, Lu K. Near-infrared spectra noninvasive measurement method of blood oxygen saturation based on the "$M+N$" theory[J]. Journal of Biomedical Engineering, 2016, 33(5): 885-889.
李丽, 王晓飞, 卢恺. 基于"$M+N$"理论的近红外光谱血氧饱和度无创测量方法[J]. 生物医学工程学杂志, 2016, 33(5): 885-889.

[9] Zhou Z H. Machine learning[M]. Beijing: Tsinghua University Press, 2016: 121-145.
周志华. 机器学习[M]. 北京: 清华大学出版社, 2016: 121-145.

[10] Li G, Wang H Q, Zhao Z, *et al*. The quality evaluation of dynamic spectrum data [J]. Spectroscopy and Spectral Analysis, 2010, 30(10): 2802-2806.
李刚, 王慧泉, 赵喆, 等. 动态光谱数据质量的评价[J]. 光谱学与光谱分析, 2010, 30(10): 2802-2806.

[12] Li G, Xiong C, Wang H Q, *et al*. Single-trial estimation of dynamic spectrum [J]. Spectroscopy and Spectral Analysis, 2011, 31(7): 1857-1861.
李刚, 熊婵, 王慧泉, 等. 动态光谱的单拍提取[J]. 光谱学与光谱分析, 2011, 31(7): 1857-1861.

[13] Lin L, Li Y C, Wang M J, *et al*. D-value estimation of dynamic spectrum based on the statistical methods [J]. Spectroscopy and Spectral Analysis, 2012, 32(11): 3098-3102.
林凌, 李永城, 王蒙军, 等. 基于统计方法的动态光谱差值提取[J]. 光谱学与光谱分析, 2012, 32(11): 3098-3102.