Advances in Applied Mathematics 应用数学进展, 2020, 9(12), 2181-2187
Published Online December 2020 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2020.912254

Hans汉斯

基于ε-SVR的近红外无创血糖浓度回归预测研究

王朱宇[1]，杨新佳[1,2]，周林华[1,2*]，马文联[1*]

[1]长春理工大学理学院，吉林 长春
[2]省级数学实验教学示范中心(长春理工大学)，吉林 长春
Email: 1753239706 @qq.com, *zhoulh@cust.edu.cn, *mawl@cust.edu.cn

收稿日期：2020年11月21日；录用日期：2020年12月18日；发布日期：2020年12月25日

摘　要

为了提高近红外无创血糖浓度回归的预测精度，首先使用近红外光谱采集人体手指指尖光谱信号，同时用微创血糖仪检测对应血糖浓度；进一步基于高斯核函数与ε-支持向量回归机(ε-SVR)建立无创血糖浓度回归模型，并与偏最小二乘回归(PLSR)模型进行对比分析。通过ε-SVR模型对两名志愿者的实验结果表明：ε-SVR模型对测试集的预测精度有很大的提高，测试集的真实浓度与预测浓度的均方误差低，克拉克误差网格分析中位于A区的概率也均高于99%。

关键词

ε-SVR，无创血糖浓度预测，近红外光谱，高斯核函数

Regression Prediction of Non-Invasive Blood Glucose Concentration Based on ε-SVR

Zhuyu Wang[1], Xinjia Yang[1,2], Linhua Zhou[1,2*], Wenlian Ma[1*]

[1]School of Science, Changchun University of Science and Technology, Changchun Jilin
[2]Provincial Demonstration Center for Experimental Mathematics Education (Changchun University of Science and Technology), Changchun Jilin
Email: 1753239706 @qq.com, *zhoulh@cust.edu.cn, *mawl@cust.edu.cn

Received: Nov. 21st, 2020; accepted: Dec. 18th, 2020; published: Dec. 25th, 2020

*通讯作者。

文章引用: 王朱宇, 杨新佳, 周林华, 马文联. 基于 ε-SVR 的近红外无创血糖浓度回归预测研究[J]. 应用数学进展, 2020, 9(12): 2181-2187. DOI: 10.12677/aam.2020.912254

Abstract

In order to improve the prediction accuracy of near-infrared non-invasive blood glucose concentration regression, near-infrared spectroscopy technology is first used to collect individual fingertip spectral signals, and the corresponding blood glucose concentration is measured by a minimally invasive blood glucose meter, and then a Gaussian kernel function and ε-support vector regression machine (ε-SVR) are used to establish non-invasive blood glucose concentration regression model, and compared with partial least squares regression (PLSR) model. The experimental results of the two volunteers through the model show that the model has greatly improved the prediction accuracy of the test set, and the test set is true. The mean square error between the concentration and the predicted concentration is small, and the probability of being located in area A in the Clark error grid analysis is also higher than 99%.

Keywords

Support Vector Machine Regression, Noninvasive Blood Glucose Detection, Near-Infrared Spectroscopy, Gaussian Kernel Function

1. 引言

近年来，随着化学计量学、计算机科学和光学仪器的快速发展，红外光谱定性与定量分析的精度、准确性、可靠性都有了大幅提高，其在无创血糖检测领域得到了广泛应用[1] [2] [3] [4]。然而，信号微弱、测量条件复杂、人体生理背景难以定量等问题对光谱测量的影响，使得目前的实验研究和实际产品中，得到的血糖校正模型大都达不到临床要求，仍有许多关键问题需要解决。随着机器学习[5]的发展，基于神经网络、支持向量机等机器学习方法对生物光电信息进行智能分析的研究不断得到重视和发展。如胡命嘉[6]等人通过经粒子群优化算法(PSO)优化后的支持向量机(SVM)分类器对采集到的表面机电信号进行分类识别。李东明[7]等人提出了基于多光谱应用 BP 人工神经网络预测血糖；李庆波[8]等结合净信号预处理(NAP)算法和径向基偏最小二乘(RBFPIS)回归建立了一种适合于人体血糖测量的非线性建模方法 NAP-RYFPLS。针对单个个体每天的血糖变化规律的差异性，代娟[9]等人以血糖吸收最强的 1550 nm 近红外光吸光度为自变量、血糖浓度为因变量，结合粒子群(PSO)算法和人工神经网络(ANN)建立了一种无创血糖检测模型 PSO-2ANN 模型。徐楠[10]等人提出了基于 VIP 处理的 O-PLS 算法的无创血糖基础研究。张春霞[11]，马爽[12]等将支持向量机应用到无创血糖检测中，这些方法均在一定程度上对回归精度有所提升。

人体的复杂性对近红外光谱无创血糖检测造成了很大的影响，实际情况下采集到的近红外光谱信号和血糖浓度值之间存在复杂的非线性关系。本文在此基础上选择 ε-SVR 结合高斯核函数作为回归模型，基本思想是通过事先确定的非线性映射将输入向量映射的一个高维特征空间(Hilbert 空间)中，然后在此高维空间中再进行线性回归，从而取得在原空间非线性回归的效果。

2. ε-支持向量机回归原理

在支持向量机回归中，输入样本 x 首先通过非线性映射 $\phi(x)$ 映射到一个高维的特征空间，然后在这个特征空间中建立一个线性模型来估计回归函数，公式如下：

$$f(x,\omega)=\omega\cdot\phi(x)+b \tag{1}$$

其中，ω 为权重向量；b 为阈值。

对于给定的训练数据集 $\{(x_i,y_i)|i=1,2,\cdots,n\}$，采用 ε-不敏感损失函数，对应的支持向量机称为 ε-支持向量机，则其约束优化问题可表示为：

$$\begin{aligned}&\min_{\omega}\frac{1}{2}|\omega|^2+C\sum_{i=1}^{n}\xi_i+\xi_i^*,i=1,2,\cdots,n\\&\text{s.t.}\begin{cases}y_i-\omega\cdot\phi(x)-b\le\varepsilon+\xi_i^*;\\\omega\cdot\phi(x)+b-y_i\le\varepsilon+\xi_i;\\\xi_i,\xi_i^*\ge 0.\end{cases}\end{aligned} \tag{2}$$

式(1)的优化问题可通过引入拉格朗日函数将其转化为对偶问题，通过解对偶问题得到式(2)的解：

$$f(x)=\sum_{i=1}^{n_{SV}}(\alpha_i-\alpha_i^*)K(x_i,x)+b \tag{3}$$

其中，$\alpha_i,\alpha_i^*(i=1,2,\cdots,l)$ 为拉格朗日乘子，α_i,α_i^* 只有一小部分不为 0，它们对应的样本就是支持向量(Support Vector, SV)；n_{SV} 为支持向量的个数；$K(x_i,x)$ 为核函数，通常采用高斯核函数：

$$K(x_i,x)=\mathrm{e}^{\left(-\lambda\|x-x_i\|^2\right)} \tag{4}$$

其中，λ 为核参数。

3. 数据准备

本文采用近红外光谱成像仪在 950 nm 到 1700 nm 的波长范围内采集两名志愿者在葡萄糖摄入后随时间变化血糖浓度对应的光谱信号。光谱分辨率为 5 nm，每测量一次，光谱仪对着同一区域扫描 100 次，得到 100 帧数据，每一帧的大小为 320 × 1。总共测量了 68 次，得到光谱数据 Data_1 和 Data_2。

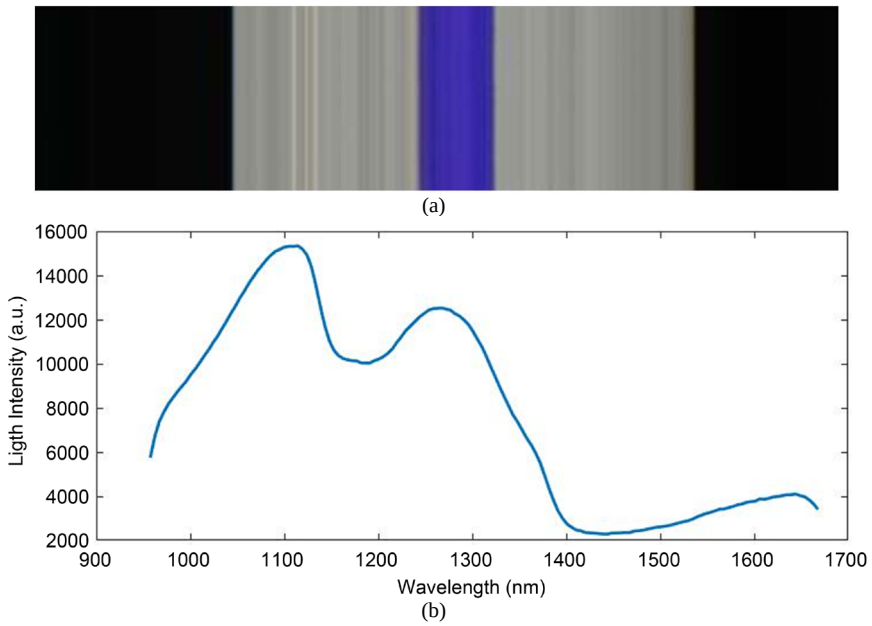

Figure 1. Image map of raw data and near infrared spectrum graph of fingertips. (a) Image map of detection of local blood glucose concentration at fingertips by near infrared imaging spectrometer; (b) Near infrared spectrum of fingertips

图 1. 原始数据成像图及其指尖近红外光谱曲线图。(a) 近红外成像光谱仪对指尖局部血糖浓度探测的成像图；(b) 指尖近红外光谱曲线

数据形式如图 1 所示，图 1(a)为测量中近红外成像光谱仪对指尖局部血糖浓度探测 100 帧的图像，其像素大小为 320 × 100。其中蓝色部分为手指所在位置，由于不是对整个手指进行采集，只是采集了指尖某个区域的数据，所以图像不会呈现手指形状。读取图 1(a)中的信息，获得图中所有点处的光谱曲线，其中指尖处的光谱曲线如图 1(b)所示。

Data_1 和 Data_2 都是测量了 68 次所得到的光谱数据，但它们分别得到了 39 和 30 个浓度。由此可见同一个浓度会有多个光谱样本与之对应，其中每个光谱样本包含 100 条光谱信号，每条光谱的维度为 150。以志愿者 1 为例，为了保证数据的均衡性，当浓度的批次数量大于等于 2 时，随机选取其中一批，使得每个浓度只有一批数据。由于血糖变化给光谱所带来的影响非常的小，为了使光谱能更好的反映血糖浓度，需要提取出手指所在位置的光谱数据和进一步处理。由于手指所在位置的光谱曲线波动性较大，与手指以外地方的光谱曲线有明显区别，故选择标准差作为评判标准，确定手指所在位置。具体步骤如下：

Step1：对第一帧数据中的 320 条光谱分别求出标准差 S。

Step2：寻找最大标准差所在的位置并提取对应的光谱，最终 100 帧数据提取到 100 条光谱。

Step3：去除异常光谱。对 100 条有效光谱应用马氏距离法和 3σ 法则筛选出异常光谱，首先设置一个阈值。这个阈值是根据计算出的 100 个马氏距离设置的，阈值 $D_t = 3\sigma_D$，其中 $3\sigma_D$ 是马氏距离的标准差。然后由 3σ 进行判断，$\left|D_i - \bar{D}\right| \le D_t$，则第 i 条光谱为正常光谱；反之，若 $\left|D_i - \bar{D}\right| > D_t$，则第 i 条光谱为异常光谱，将其剔除并随机选择一条正常光谱进行填充，其中 $\bar{D}$ 为马氏距离的平均值。

Step4：按照上述操作，对浓为度 $C_k\ (k = 1,2,\cdots,39)$ 的光谱进行提取。

Step5：对每条光谱取对数变换获得总数据集 Ω，

$$\Omega = \left\{ \left(x_{n,1}^k, x_{n,2}^k, \cdots, x_{n,150}^k \right) \middle| k = 1,2,\cdots,39; n = 1,2\cdots,100 \right\}。$$

Step6：将采集到的光谱 Ω 随机分成两个数据集。从每个样本浓度对应的光谱中随机选出 80%条光谱构成训练数据集，其余 20%条光谱则作为测试数据集，分别用于校准和测试模型。

Step7：对训练集数据进行归一化处理，保留最大值最小值用于测试集归一化。

4. 结果与讨论

4.1. 模型训练

首先设置训练条件：1) 采用 ε-支持向量回归机进行回归；2) 高斯函数作为核函数；3) 通过传统网格搜索法进行试验，网格大小为 10 × 10，移动步长为 0.5，并且采用 5 折交叉验证所得的误差作为优化参数的标准；4) 预测误差采用均方误差 MSE 表示：

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - f_i\left(x\right)\right)^2 \tag{5}$$

其中 y_i 实际血糖浓度值；$f_i\left(x\right)$ 表示血糖光谱回归预测的浓度值。

然后将训练集光谱数据与其对应的标签向量作为输入，对模型进行训练，得到最优参数带标签的训练集进行模型训练，得到最优参数 w,b 和回归函数 $f\left(x,w\right)$，至此模型训练结束。最后将测试集光谱数据作为输入，进行浓度预测。

4.2. 模型评价

本节分别通过偏最小二乘回归(PLSR)与 ε-支持向量机回归机(ε-SVR)获得预测结果，并对模型的性能进行评估。采用决相关系数(R^2)、均方误差(MSE)和克拉克误差网格分析对血糖浓度回归模型进行评价。

对模型进行 5 折交叉验证，并计算出测试集均方误差、决定系数和克拉克误差网格 A 区概率的均值，用于作为模型评价的标准，分别用符号 *MSECV*、R^2、*P* 来表示。两名志愿者分别建立模型得到验证结果如表 1 所示。

Table 1. Results of validation of models established by two volunteers
表 1. 两名志愿者分别建立的模型的验证结果

志愿者	模型	*MSECV*/mmol·L^{-1}	R^2/%	*P*/%
志愿者 1	PLSR	1.9961	53.17	55.62
	ε-SVR	0.001	99.73	99.92
志愿者 2	PLSR	1.0334	35.92	86.67
	ε-SVR	0.0004	99.84	99.97

由表 1 可见，两名志愿者通过 PLSR 分别建模所得到的模型测试集的真实浓度和预测浓度的平均相关系数相对较低，平均误差 *MSECV* 的值分别在 2 mmol·L^{-1} 和 1 mmol·L^{-1} 左右，模型的预测效果较差，精度不高；与之相比通过 ε-SVR 建模所得到的模型相关系数集的真实浓度和预测浓度的平均相关系数均高于 99%，*MSECV* 的值分别在 0.001 mmol·L^{-1} 和 0.0004 mmol·L^{-1} 左右，克拉克误差网格分析的结果几乎 100% 落在 A 区。同时从两个志愿者的 *MSECV* 值和 *P* 值也可以看出不同个体所建的模型效果也有所区别。

图 2 是关于志愿者 1 的数据在不同模型中参考浓度与预测浓度的拟合情况。图 2(a)和图 2(b)分别表示了使用 PLSR 模型和 ε-SVR 模型进行回归预测中参考浓度与预测浓度的拟合程度。将参考浓度与预测浓度构成有序实数对，这些离散的点越靠近直线 $y = x$ 说明其预测结果越好。所以从图 2 中两个模型的结果对比来看，ε-SVR 模型优于 PLSR 模型。

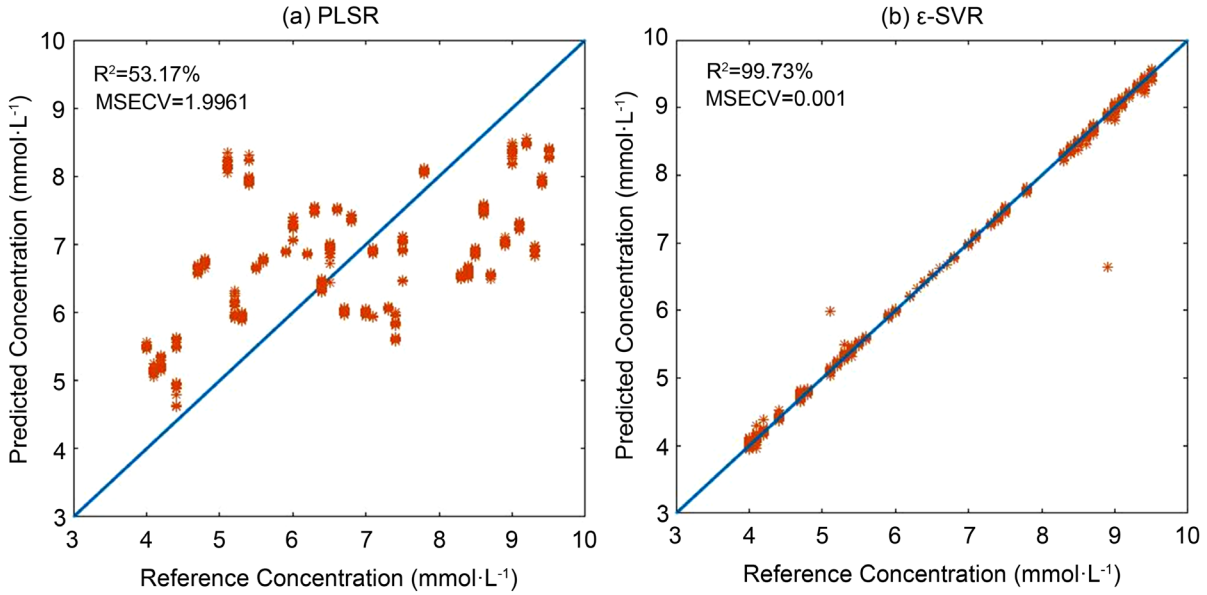

Figure 2. Fitting results of reference and predicted concentrations in different models (Volunteer 1)
图 2. 不同模型中参考浓度与预测浓度的拟合结果(志愿者 1)

克拉克误差网格分析通过比较参考血糖浓度值和预测血糖浓度值间的关系来评估血糖算法准确度。网格分为五个区域，标记为 A、B、C、D、E，落在 A、B 区域理论上可接受，落在 C、D、E 区域会造

成潜在的危险，造成临床误诊。注意 1 mmol·L^{-1} = 18 mg·dL^{-1}，为了更清晰的观查其误差变化，所以将参考浓度和预测浓度进行单位换算。图 3(a)和图 3(b)分别表示志愿者 1 的数据在不同模型下真实浓度与预测浓度的克拉克误差网格分析结果分布图。从图中可以看出，PLSR 模型预测的结果分布较散，散落在了 A 区以外，而 ε-SVR 模型的预测结果几乎全部落在 A 区。同时结合表 1 的结果也可以看出 ε-SVR 的线性性更强，模型预测效果较好，精度更高，具有较强的泛化能力。

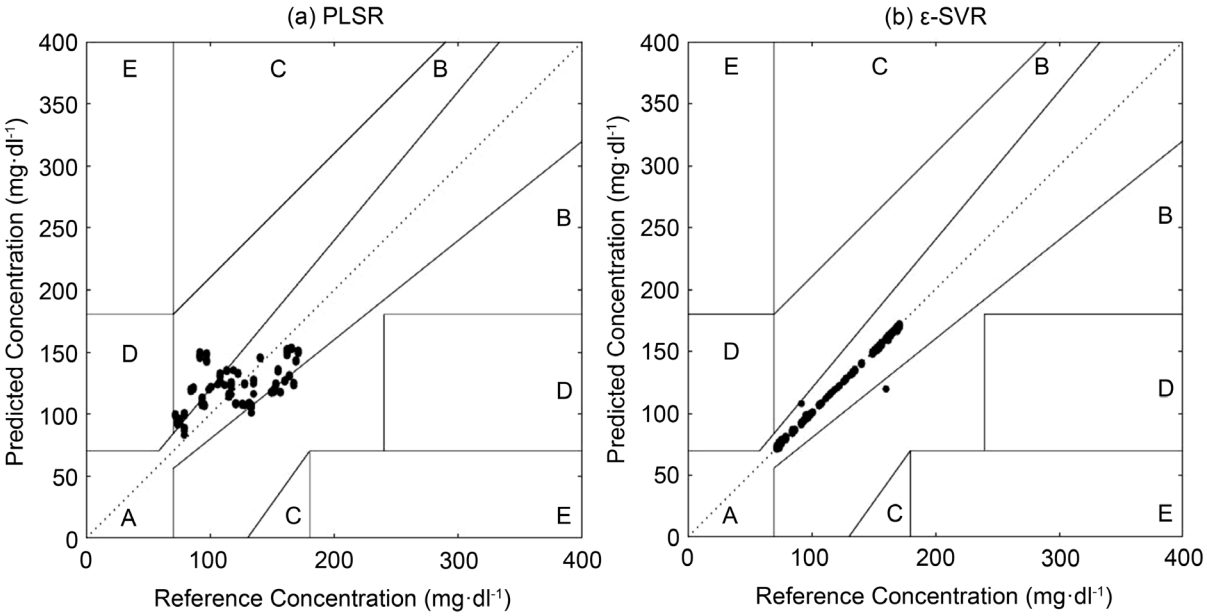

Figure 3. Distribution of Clarke error grid analysis results for different models (Volunteer 1)
图 3. 不同模型的 Clarke 误差网格分析结果分布图(志愿者 1)

5. 结论

支持向量机回归模型有效地减弱了干扰因素对预测模型的影响，能拟合血糖浓度与近红外光谱之间的非线性关系，从而提高了预测精度。相比 PLSR 模型，ε-SVR 模型测试集的克拉克网格误差分析结果均达到 99%；*MSECV* 的值分别为 0.001 mmol·L^{-1} 和 0.0004 mmol·L^{-1} 左右。同时对比两个志愿者的克拉克误差分析结果也可以看出不同个体所建的 PLSR 模型效果区别明显，但针对 ε-SVR 模型，两个个体模型的克拉克误差分析结果差别不大，也进一步说明了 ε-SVR 是适合于人体血糖浓度测量建模的非线性方法。除此之外，对比志愿 1 和志愿者 2 的均方误差和克拉克误差网格分析落入 A 区的概率，可以发现：在同样的模型下，志愿者 2 的结果比志愿者 1 的要更好，这也进一步验证不同个体之间存在差异性。

参考文献

[1] 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): 56-72.

[2] 刘蓉, 徐可欣, 陈文亮, 等. 光学无创血糖检测中的主要问题及研究进展[J]. 中国科学(G 辑: 物理学·力学·天文学), 2007(S1):124-131.

[3] 王慧泉. 动态光谱法血液成分无创检测若干关键技术研究[D]: [博士学位论文]. 天津: 天津大学, 2014.

[4] 王晓飞, 张欣怡, 徐馨荷. 考虑多种因素的近红外光谱血糖预测模型对比[J]. 激光与光电子学进展, 2019, 56(4): 209-213.

[5] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016: 121-145.

[6] 胡命嘉, 宫玉琳, 王锋. 基于 PSO-SVM 的手势识别方法研究[J]. 长春理工大学学报(自然科学版), 2019, 42(4):

102-107.

[7] 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): 244-249.

[8] 李庆波, 黄政伟. 净信号预处理结合径向基偏最小二乘回归在血糖无创检测中的应用[J]. 光谱学与光谱分析, 2014, 34(2): 494-497.

[9] 代娟, 季忠, 杜玉宝. 基于粒子群和人工神经网络的近红外光谱血糖建模方法研究[J]. 生物医学工程学杂志, 2017, 34(5): 713-720.

[10] 徐楠, 邓琛, 江潇潇, 等. 基于 VIP 处理的 O-PLS 算法的无创血糖基础研究[J]. 测控技术, 2019, 38(3): 108-111.

[11] 马爽, 蒲宝明. 基于支持向量机的无创血糖光谱算法[J]. 计算机系统应用, 2016, 25(8): 120-124.

[12] 张晓霞. 基于支持向量机的无创血糖检测电极影响研究[D]: [硕士学位论文]. 成都: 成都理工大学, 2017.