基于支持向量机的无创血糖光谱算法①

马 爽[1,2], 蒲宝明[1]

[1](中国科学院 沈阳计算技术研究所, 沈阳 110168)
[2](中国科学院大学 计算机与控制学院, 北京 100049)

摘 要: 为了实现无创血糖浓度检测, 提出了基于支持向量机回归模型的无创血糖光谱算法. 该算法使用光电容积脉搏波(PPG)设备对志愿者指端红光、红外光交替采样得到 PPG 信号, 然后通过微创血糖仪测得血糖浓度. 对采集到的 PPG 信号进行处理提取特征组成特征矩阵, 分别运用不同机器学习模型对特征矩阵和实时血糖浓度进行回归训练, 得到特征矩阵与血糖浓度间的关系, 并对训练得到的函数关系进行验证, 选取出高斯核支持向量机模型为最佳训练模型. 实验证明, 与偏最小二乘回归进行对比, 本文提出的运用核函数为高斯核的支持向量机算法的预测准确度能提升 10%～15%, 预测的高低血糖正确率达到 98%.

关键词: 支持向量机; 无创血糖检测; 近红外光谱; 光电容积脉搏波

Non-Invasive Blood Glucose Measurement Based on Support Vector Machine Algorithm

MA Shuang[1,2], PU Bao-Ming[1]

[1](Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China)
[2](School of Computer and Control Engineering, University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: To achieve sensing non-invasive blood glucose concentration, this paper proposes a non-invasive blood glucose spectrum algorithm based on support vector machine (SVM) model. The algorithm firstly uses a PPG device to sample red and infrared signal of volunteers to obtain the PPG signal, and then extracts the blood glucose concentration by a minimally-invasive glucometer. Compared with traditional near-infrared spectrum and regression method, the proposed algorithm could effectively improve the prediction accuracy by adopting the SVM model to find regression relationship between signal and glycemic. And the SVM model uses the Gauss-function as the kernel function. The algorithm is independently of the individual and environmental factors. The experimental results demonstrate that compared to partial least squares regression, the proposed algorithm could improve the predictive accuracy by 10%–15%. And the algorithm's forecast accuracy could reach to 98%.

Key words: support vector machines; non-invasive blood glucose measuring; near-infrared spectrum; photoplethysmograph

据世界卫生组织统计, 2014 年世界糖尿病患者达到 3.89 亿, 预计截止到 2035 年, 世界糖尿病患者可能达到 5.9 亿[1]. 血糖浓度控制不当易引发多种并发症, 危及生命. 糖尿病患者需要每天多次检测血糖浓度. 目前临床采用有创、微创方式检测血糖, 准确度高, 但是给患者带来病痛的同时, 增加感染几率, 而且花费较高, 给患者带来经济负担. 无创血糖实时检测可以克服以上缺点, 近年来成为国际学术界研究热点[2].

目前无创血糖检测方法有多种, 如偏振光测定[3]、能量代谢守恒测定[4]、中红外光谱测定[5]、近红外光谱测定[6]等. 其中, 近红外光谱法是目前最有前景的无创血糖检测方法, 是目前无创血糖检测的主流方法. 但是受到人体皮肤和血液吸收、个体特异性、外界环境产生的随机噪声影响, 郎伯比尔定律线性关系被严重破坏, 多元线性回归、偏最小二乘回归、主成分回归等以理论导向为基础的回归算法在预测血糖浓度中

① 基金项目: "核高基"专项(2012ZX01029_001_002)
收稿时间:2015-11-25;收到修改稿时间:2015-12-31 [doi:10.15888/j.cnki.csa.005253]

表现不佳[7]. 引入核技巧的非线性支持向量机模型由Boser、Guyon、Vapnik[8]提出, 特别适用于先验理论不清、实验环境复杂的系统, 是近年来机器学习领域的研究热点, 已经在特征选择、图像处理、模式识别等诸多领域广泛使用并取得了优异的成绩.

在此背景下, 本文将支持向量机回归模型引入无创血糖光谱检测中. 本算法根据提 PPG 信号提取光谱特征, 与实时血糖真实值组成矩阵, 输入到不同的机器学习模型中进行回归训练并验证, 选取出高斯核支持向量机模型为最佳机器学习模型, 即基于支持向量机回归模型的无创血糖光谱算法. 受到个体差异性限制, 本算法目前不能够对人群中所有个体建立统一校正模型, 但可以实现对单个个体进行个体校正.

1 无创血糖检测模型及其算法

本文建立的无创血糖检测算法由三个主要部分组成, 如图 1.

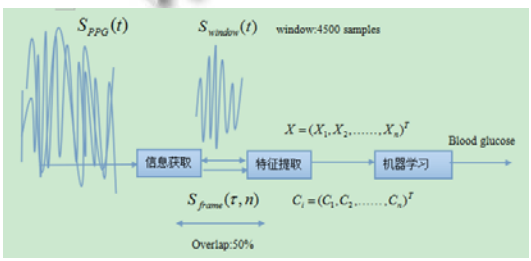

图 1 算法结构图

采用红光、红外光交替采样, 拟定采样频率为 75 样本/秒, 将采集的光电信号转化成数字信号, 得到序列 $S_{PPG}(t)$,作为信息获取模块的输入; 信息获取模块从 $S_{PPG}(t)$中选取 4500 个连续样本作为一个窗口, 再将窗口中的样本数据进行分片, 该模块最终输出:

矩阵 $S_{window}(t)$, 由一分钟的采集信号组成(样本长度 L_{window}=4500).

二维数组 $S_{frame}(\tau,n)$, 从 $S_{window}(t)$中分离出红光、红外光信号 $S'_{window}(t)$和 $S''_{window}(t)$, 分别拆分出若干个 $S'_{frame}(\tau,n)$和 $S''_{frame}(\tau,n)$, τ 代表每个 *Frame* 中的样本编号, n 代表每个 *Window* 中 *Frame* 编号. 每个 *Frame* 由 5 秒(375 个)样本组成, 相邻两个 *Frame* 重叠率为 50%, 每个 *Frame* 中样本数量即 L_{frame}=375, 每个 *window* 中 *Frame* 数量即 N_{frame}=24.

$S_{window}(t)$和 $S_{frame}(\tau,n)$作为特征提取模块的输入, 输出矩阵 $X_i=(X_{i1},X_{i2},\ldots,X_{in})$, 此矩阵中包含支持向量机回归模型所需要的多维特征, 将矩阵 X_i 结合血糖浓度矩阵 C_i 输入到支持向量机回归模型中训练, 得到血糖浓度与特征矩阵间的函数关系, 最终输出血糖浓度.

1.1 信息获取

信息获取的主要功能是去除 $S_{PPG}(t)$中的干扰信号(信号失真、抖动, 噪声等), 将交替的红光、红外光信号分离, 输出矩阵 $S'_{window}(t)$、$S''_{window}(t)$和数组 $S'_{frame}(\tau,n)$、$S''_{frame}(\tau,n)$. $S_{window}(t)$选为一分钟、4500 个采样点, 保证了计算样本充足; 对于每个 *Frame* 周期选为 5 秒和 50%的重叠保证每个 *Frame* 中至少有 4～5 次脉搏搏动, 5 秒的时间小于呼吸周期, 最大程度减少由于呼吸或其它生理因素对波形造成的干扰.

1.2 特征提取

特征提取模块从获取的 PPG 信号中提取特征, 根据信息获取模块输出的 $S'_{window}(t)$、$S''_{window}(t)$提取多维特征矩阵作为机器学习模块的输入.

1.2.1 *Kaiser-Teager* 能量特征

Kaiser-Teager 能量特征计算公式[8]为:

$$KTE(t)=x(t)2-x(t+1)x(t-1) \tag{1}$$

根据该式可得到分片实时能量值:

$$KTE_n(t)=S_{frame}(t,n)-S_{frame}(t+1,n)S_{frame}(t-1,n) \tag{2}$$

其中, t=1,2,…L_{frame}-1.

由 $KTE_n(t)$可得单个分片的均值 KTE_n^{μ},方差 KTE_n^{σ}, 四分间距 KTE_n^{iqr}, 斜度 KTE_n^{skew}; 由单个分片的上述特征可得到分片所在窗口的能量均值 KTE^{μ},方差 KTE^{σ}, 四分间距 KTE^{iqr}, 斜度 KTE^{skew}.

1.2.2 心率特征

通过采集波形可得心脏跳动间隔时间, 从而得出窗口心率均值 HR^{μ}、方差 HR^{σ}、四分间距 HR^{iqr}、偏度 HR^{skew}.

1.2.3 光谱熵特征

首先对 $S_{frame}(\tau,n)$进行快速傅立叶变换, 即:

$$X_n \leftarrow FFT(S_{frame}(\tau,n),L_{FFT}) \tag{3}$$

其中,L_{FFT}=512.

然后对 X_n 进行正则化, 即:

$$P_X^n[k] \leftarrow \frac{|X_n[k]|^2}{\sum_{j=1}^{L_{FFT}}|X_n|j||^2},k=1\ldots L_{FFT} \tag{4}$$

最后, 根据 P_X^n 求熵, 即:

$$H \leftarrow P[k]Log(P[k]) \tag{5}$$

通过 H_n^s 可计算出单个分片的均值 H_s^{μ}、方差 H_s^{σ}、四分间距 H_s^{iqr}、偏度 H_s^{skew}.

1.2.4 光谱能量对数特征

根据光谱能量对数公式

$$LogE \leftarrow Log(S(\tau,n)) \quad (6)$$

计算出分片所在窗口的光谱能量对数方差 $LogE^{\sigma}$, 四分位差 $LogE^{iqr}$.

将上述特征组成特征矩阵: X_i=(KTE^{μ}, KTE^{σ}, KTE^{iqr}, KTE^{skew}, HR^{μ}, HR^{σ}, HR^{iqr}, HR^{skew}, Hs^{μ}, Hs^{σ}, Hs^{iqr}, Hs^{skew}, $LogE^{\sigma}$, $LogE^{iqr}$)作为机器学习模块的输入.

1.3 机器学习

机器学习模型有很多种, 支持向量机回归模型是一种适用于实验环境复杂、先验理论不清的多维矩阵回归方法[9]. 其中, 线性支持向量机最早由 Cortes 与 Vapnik 提出并应用于统计领域, 并成为线性分析的最有效方法之一. 但是, 对于非线性模型, 线性支持向量机回归效率低下、错误率高.

本算法根据采集得到的 PPG 信号提取出 1.2 中的 14 维特征, 并使用微创血糖仪测得对应的血糖浓度, 将多组特征和其对应的血糖浓度组成矩阵, 然后用 MATLAB 读取矩阵中的数据作为高斯核支持向量机模型、线性支持向量机模型、线性偏最小二乘、非线性偏最小二乘的接口的参数, 并分别进行回归训练, 得出运用四种不同机器学习算法下的特征矩阵 X 与血糖浓度的关系.

2 实验

为了探究特征矩阵和血糖浓度的关系, 设计实验, 通过比较不同机器学习模型的预测结果, 选取最佳机器学习模型.

2.1 实验过程设计

实验硬件计算机一台, 采用 WINDOWS 操作系统, MATLAB 开发平台; MSP430 开发板一台, 血氧指夹采用 AFE4403 集成模拟前端, 怡成血糖仪一台.

实验选取一名健康志愿者, 连续一周, 每天连续五小时, 每隔 10 分钟用 PPG 设备采集指端一分钟 PPG 信号, 同时, 用怡成血糖仪记录实时血糖值.

由采集的 PPG 信号, 根据 2.1 中编写的分片算法和 1.2 中特征提取算法, 提取出 14 维特征. 随机选取采集信号中 80%数据作为训练集, 调用 MATLAB 中四种不同的机器学习算法, 进行回归训练, 余下 20%数据作为测试集合, 根据学习到的回归函数计算对应的血糖值, 并使用克拉克误差网格程序[10]分析算法准确度.

2.2 实验结果

实验结果分别通过线性偏最小二乘回归、线性支持向量机回归、非线性偏最小二乘回归、非线性支持向量机回归(高斯核)方法获得. 不同方法性能首先通过克拉克误差网格误差分析进行评估, 然后分别给出算法整体性能比较和个体性能比较.

2.2.1 克拉克网格误差分析

克拉克网格误差分析通过比较参考血糖值和实际血糖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.

图 2、图 3 分别给出了运用克拉克误差分析对实验中四种机器学习方法进行血糖浓度预测结果图, 其中, ‘+’代表运用偏最小二乘回归, ‘*’代表运用支持向量机回归.

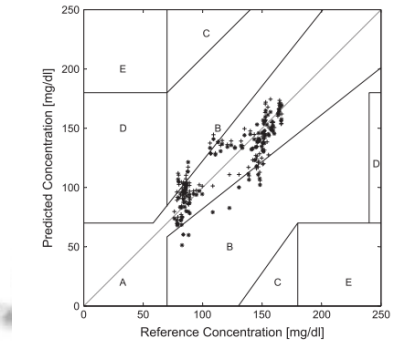

图 2　偏最小二乘和支持向量机回归克拉克误差网格结果

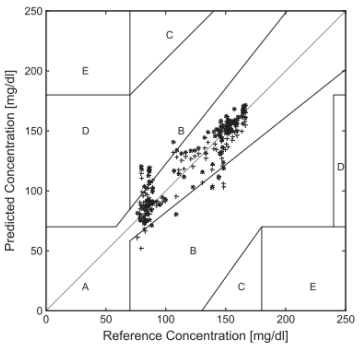

图 3　非线性偏最小二乘和支持向量机回归克拉克误差网格结果

图 2 中, 运用线性偏最小二乘回归, 79.31%的点落在 A 区域; 运用线性支持向量机回归算法, 84.48%的点落在 A 区域. 图 3 中, 运用线性偏最小二乘回归, 83.62%的点落在 A 区域; 运用非线性支持向量机回归, 88.79%的点落在 A 区域.

从图 2、图 3 中可以看出, 同样的预测数据, 使用支持向量机回归方法比使用偏最小二乘回归更接近克拉克网格对角线, 可见支持向量机回归方法比偏最小二乘回归具有更高的精准度.

2.2.2 算法整体性能比较

表 1 给出了线性和非线性最小二乘回归、线性和非线性支持向量机回归算法的预测结果. 这几种算法的整体性能通过分别计算训练集均方根误差(RMSEC)和测试集合的校正均方根误差(RMSEP)分析, 校正均方根误差公式如下[11]:

$$RMSEC=\left\{\sum_{m=1}^{M}(y_m-\hat{y}_m)^2/(M-1)\right\}^{1/2} \quad (7)$$

$$RMSEP=\{\sum_{n=1}^{N}(y_n-\hat{y}_n)/(N-1)\}^{1/2} \quad (8)$$

训练集中, 使用支持向量机回归算法 RMSEC 为 0.7856mmol/L,比使用偏最小二乘算法提升 20.11%; 测试集中, 使用支持向量机回归算法 RMSEP 为 0.8682mmol/L,比使用偏最小二乘算法效率提升 7.88%. 使用非线性偏最小二乘和非线性支持向量机回归算法可以得到类似的结论.

训练集和测试集的预测血糖和真实血糖之间相关系数用 rC 和 rP 来表示. 在测试集合中, 用线性偏最小二乘算法相关系数[12]为 0.8476, 用线性支持向量机回归相关系数为 0.8785; 运用非线性偏最小二乘回归相关系数为 0.8981, 用非线性支持向量机回归相关系数为 0.9152.

表 1 算法整体性能

数据 \ 方法		偏最小二乘	支持向量机	非线性偏最小二乘	非线性支持向量机
训练集	RMSEC(mmol/L)	0.9834	0.7856	0.7597	0.6477
	rC	0.8385	0.9003	0.9204	0.9353
测试集	RMSEP(mmol/L)	0.9425	0.8682	0.7778	0.7370
	rP	0.8476	0.8785	0.8981	0.9152

表 2 算法个体性能(训练集)

数据 \ 方法	观察者数量			
	偏最小二乘	支持向量机	非线性偏最小二乘	非线性支持向量机
>10%	101	73	53	46
5%-10%	36	41	48	46
2.5%-5%	16	22	41	39
1%-2.5%	14	22	41	39
0%-1%	7	16	21	21

表 3 算法个体性能(测试集)

数据 \ 方法	观察者数量			
	偏最小二乘	支持向量机	非线性偏最小二乘	非线性支持向量机
>10%	62	46	40	30
5%-10%	28	32	28	24
2.5%-5%	15	20	27	31
1%-2.5%	8	11	12	16

2.2.3 算法个体性能比较

表 2 和表 3 给出测试集和训练集中每个个体的预测结果. 如表 3 所示, 在训练集合中, 使用线性支持向量机回归算法 18 个观测数据误差低于 2.5%, 使用线性偏最小二乘 11 个观测数据误差低于 2.5%; 使用非线性支持向量机回归算法 62 个观测数据误差低于 5%, 使用非线性偏最小二乘算法 48 个观测数据误差低于 5%, 由此可见, 使用线性和非线性支持向量机回归算法均比偏最小二乘算法性能有明显改善.

3 结论

为了实现无创血糖监测, 本文提出了应用光谱分析中的 PPG 技术提取信号特征, 结合真实血糖值组成矩阵, 输入机器学习模块中训练, 得到目标决策函数, 最后通过克拉克网格误差分析、计算训练集和测试集相关系数、计算校正均方根误差比较并选取最佳机器学习模型.

实验证明: 使用支持向量机回归模型进行训练, 核函数采用高斯核, 比线性、非线性偏最小二乘回归以及线性支持向量机回归算法整体准确度好, 整体预测高低血糖正确率达到 98%; 个体误差范围低, 较其他机器学习算法准确度提升 10%–15%, 误差范围在 20%以内, 达到美国 FDA 检测标准.

后期研究中, 还需要改进指夹前端, 选取其他波长的光及精度合适的传感器, 以减少输入量误差; 改

进支持向量机回归模型, 使用不同核函数, 提高算法精度; 增加临床受试者数量, 特别是增加糖尿病病患受试者, 对更丰富的样本数据进行回归训练, 提高疾病检出率.

参考文献

1 Wen YY, et al. Prevalence of Diabetes among Men and Women in China. NEngl J Med, 2010, 362: 1090–1101.

2 陈文亮,徐可欣,等.人体无创血糖检测技术.仪器仪表学报, 2003,24(4):258–261.

3 王洪,蒋明峰,崔建国,等.基于光学旋光法的血糖浓度测量. 激光,2006,27(1):80–81.

4 Cameron BD, Gorde HW, Satheesan B, Cote GL. The use of polarized laser light through the eye for Nonivasive glucose monitorinol. Diabetes Technol, Ther. 1999,1: 135–143.

5 Borchert MS, Storrie-Lombardi MC, Lambert JL. A nonivasive glucose monitor: preliminary results in rabbits. Diabetes Technol.Ther. 1999,1: 145–151.

6 Maruo K, Tsurugi M, Chin J, Ota T, Arimoto H, Yamada Y, Tamura M, lshii M, Ozaki Y. Nonivasive blood glucose assay using a newly developed near-infrared system. IEEE J Sel Top.Quantum Electron., 2003, 9(2): 322–330.

7 Ooi ET, Zhang XQ, Chen JH, Soh PH, Ng K, Yeo JH. Nonivasive blood glucose measurement using multiple laser diodes. SPIE Conference Series, 2007, 6445.

8 Heise HM, Bittner A. Multivariate calibration for Near-infraed spectroscopic assays of blood substrates in human plasma based on variable selection using pls-regression vector choices. Fresenius J. Anal. Chem, 1998, 362(1): 141–147.

9 Ham FM, Kostanic IN, Cohen GM, Gooch BR. Determination of glucose concentrations in an aqueous matrix from nir spectra using optimal time-domin filtering and partial least-squares regression. IEEE Trans. Biomed, Eng. 1997, 44(6): 475–485.

10 Clarke WL. The original clarke error grid analysis(ega). Diabetes Technol. Ther., 2005, 7(5): 776–779.

11 刘光达,蔡靖,孙茂林,等.基于光电容积脉搏波的无创血糖测量研究.吉林大学学报,2015,1:1671–5876.

12 李航,等.统计学习方法.北京:清华大学出版社,2015.