文章编号:1671 - 6833(2019)02 - 0001 - 05

基于优化极限学习机的 CVD 预测模型研究

逯 鹏[1,2], 李奇航[1,2], 尚莉伽[3], 李新建[1], 张 微[1]

(1. 郑州大学 电气工程学院,河南 郑州 450001; 2. 互联网医疗与健康服务河南省协同创新中心,河南 郑州 450001; 3. 北京市东城区中小学卫生保健所,北京 100007)

摘 要: 利用机器学习算法,改变传统心血管疾病(CVD)预测模型的严格数理化公式,以增加危险因素的纳入、降低数据格式的要求. 首先提出利用基于单隐层前馈神经网络(SLFNs)的极限学习机(ELM)算法建立 CVD 预测模型;进一步通过五阶段连续变异方式建立增强领导粒子的粒子群算法(ELPSO),以粒子群(PSO)算法的优化策略,对 SLFNs 的隐层单元参数进行优化. 通过对 UCI 数据库 Statlog (heart) 数据集和 heart disease database 分析结果显示,所提 ELPSO-ELM 模型测试正确率分别达到 85.71%、84.00%,AUC(ROC 曲线下面积)分别达到 0.902 4、0.842 3,高于传统 CVD 预测模型,同时放松了数据线性化约束,能纳入更多的复杂危险因素.

关键词: 心血管疾病;风险预测;极限学习机;粒子群

中图分类号: TP2 **文献标志码**: A **doi**:10.13705/j.issn.1671 - 6833.2018.05.005

0 引言

CVD 已成为致死率第一的慢性病[1]. 定量预测发病风险,能有效降低 CVD 的发病率[2].

经典的 Framingham、SCORE、WHO 等模型[3]在 10 年发病风险预测中已取得较好的预测效果,主要建模方法有:基于疾病流行过程和特征做出趋势性推理的预测法[4];基于疾病危险因素、发病情况的随访队列横断面数据的数理预测法[5]. 以上方法的核心是 Logistic 回归和 Cox 比例风险回归[6]. 回归分析方法无法解决多分类因素的拟合问题[7],此外,概率型策略对自变量数据要求较高,在缺乏大型横断面研究资料时应用困难.

机器学习具有较强的自学习、自适应能力,能够有效处理医疗信息中模糊、非线性数据[8],为解决此类问题提供了新思路. 笔者以构建网络预测模型为切入点,改变传统公式化预测模型,建立机器学习模型,从而拓展模型预测因素,提高预测能力. 首先建立 ELM[9]模型;然后引入 ELPSO 优化算法[10],构建 ELPSO-ELM 模型;最后利用文献[14]、[15]数据集验证模型性能. 实验表明,所设计算法的 AUC 值均高于传统模型的 AUC 值(0.76 ~ 0.80),可见笔者算法能较好地实现 CVD 预测.

1 传统模型问题分析

以 Logistic 回归方法为例.

给定 n 组包含 k 个危险因素的观测样本(x_{i1}, x_{i2}, …, x_{ik}; y_i)($i = 1,2,\cdots,n$), $y_i = 0$ 表示未得病, $y_i = 1$ 表示得病,二元 Logistic 回归方程式为:

$$p_i = \frac{\exp^z}{1+\exp^z} = \frac{\exp^{(\alpha+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_k x_{ik})}}{1+\exp^{(\alpha+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_k x_{ik})}}. \tag{1}$$

对式(1)取 Logistic 变换,得 Logistic 回归预测模型:

$$\mathrm{Logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \alpha + \boldsymbol{\beta}_1 x_{i1} + \boldsymbol{\beta}_2 x_{i2} + \cdots + \boldsymbol{\beta}_k x_{ik}, \tag{2}$$

式中: α 为回归截距; $\boldsymbol{\beta}_i$ 为自变量的回归系数. 用最大似然函数法求各参数值.

在回归问题中,当危险因素必须以多分类定性变量才能完全表述时,一般顺序赋值为 1、

收稿日期:2018 - 05 - 13;**修订日期**:2018 - 09 - 11

基金项目:国家自然科学基金资助项目(60841004、60971110、61172152);郑州市科技攻关资助项目(112PPTGY219-8);河南省青年骨干教师资助计划(2012GGJS-005)

作者简介:逯 鹏(1974—),男,河南郑州人,郑州大学副教授,博士,主要从事复杂视觉图像信息处理、脑 - 机接口和医疗大数据方面的研究,E-mail:lupeng@zzu.edu.cn.

2、…、n,此时带来的问题包括:①无序多分类变量被视作连续变量;②有序多分类变量,强行规定无法准确衡量类别间差距的变量为等距.上述问题丢失了数据的真实信息,模型误差较大.Logistic 回归中,上述问题可通过人为地设置哑变量[11](记作 Dv-Logistic 的方法)将多分类转换成二分类解决.但对于包含较多复杂因素的医疗数据,该方法建模困难且易出错,无法自动建立预测模型.

2 基于极限学习机和粒子群算法的预测模型

2.1 极限学习机(ELM)

极限学习机(extreme learning machine,ELM)是基于单隐层前馈神经网络[12]的机器学习算法.给定 n 组训练样本 $(\boldsymbol{x}_i,\boldsymbol{y}_i)$,$\boldsymbol{x}_i=(x_{i1},x_{i2},\cdots,x_{il})^{\mathrm{T}}$、$\boldsymbol{y}_i=(y_{i1},y_{i2},\cdots,y_{im})^{\mathrm{T}}$ 分别为 l 维输入和 m 维输出,设定隐单元个数为 k,则网络模型为:

$$\tilde{\boldsymbol{y}}_j=\sum_{i=1}^{k}\boldsymbol{g}(x)\boldsymbol{\beta}_i(\boldsymbol{w}_i\cdot\boldsymbol{x}_j+\boldsymbol{b}_i),\tag{3}$$

式中:$\boldsymbol{g}(x)$ 为激活函数;$\boldsymbol{\beta}_i$ 为第 i 个隐单元的输出权重;$\boldsymbol{w}_i$ 为第 i 个隐单元的输入权重;$\boldsymbol{b}_i$ 为第 i 个隐单元的阈值.

极限学习机能够使输出值 $\tilde{\boldsymbol{y}}_j$ 以零误差逼近真实值 $\boldsymbol{y}_j$,即存在 $\boldsymbol{\beta}_i$、$\boldsymbol{w}_i$ 和 $\boldsymbol{b}_i$,使得

$$\boldsymbol{y}_j=\sum_{i=1}^{k}\boldsymbol{g}(\boldsymbol{x})\boldsymbol{\beta}_i(\boldsymbol{w}_i\cdot\boldsymbol{x}_j+\boldsymbol{b}_i).\tag{4}$$

表示为矩阵形式,为:

$$\boldsymbol{H}\boldsymbol{\beta}=\boldsymbol{Y},\tag{5}$$

式中:$\boldsymbol{H}$ 为隐层输出矩阵;$\boldsymbol{Y}$ 为模型输出矩阵.

$$\boldsymbol{H}=\begin{bmatrix} g(w_1\cdot x_1+b_1) & g(w_2\cdot x_1+b_2) & \cdots & g(w_k\cdot x_1+b_k)\\ g(w_1\cdot x_2+b_1) & g(w_2\cdot x_2+b_2) & \cdots & g(w_k\cdot x_2+b_k)\\ \vdots & \vdots & \vdots & \vdots\\ g(w_1\cdot x_n+b_1) & g(w_2\cdot x_n+b_2) & \cdots & g(w_k\cdot x_n+b_k)\end{bmatrix}_{n\times k};$$

$$\boldsymbol{\beta}=\begin{bmatrix}\beta_1^{\mathrm{T}} & \cdots & \beta_k^{\mathrm{T}}\end{bmatrix}_{k\times m}^{\mathrm{T}};\ \boldsymbol{Y}=\begin{bmatrix}y_1^{\mathrm{T}} & \cdots & y_n^{\mathrm{T}}\end{bmatrix}_{n\times m}^{\mathrm{T}}$$

式(5)的最小范式最小二乘解为:

$$\hat{\boldsymbol{\beta}}=\boldsymbol{H}^{+}\boldsymbol{Y},\tag{6}$$

式中:$\boldsymbol{H}^{+}$ 为隐层输出矩阵 $\boldsymbol{H}$ 的 Moore-Penrose 广义逆.

2.2 ELM 模型稳定性改进

2.2.1 问题分析

ELM 随机初始化隐单元参数,由式(3)知,随机化过程中可能出现影响较小或无效单元[9],需要设置大量的隐单元才能达到理想效果.而 ELM 测试复杂度为 $O(Nlk^2m)$,N(测试样本数)、l(输入特征维数)、m(输出标签维数)均确定,为降低复杂度则要求在误差允许范围内 k 尽可能小.

另外,样本可能存在复共线性问题,每次 ELM 求解的 Moore-Penrose 广义逆 $\boldsymbol{H}^{+}$ 可能不同,导致求出的隐层输出权重不同,使得 ELM 模型泛化能力和预测稳定性不足.

2.2.2 粒子群(PSO)优化算法

针对 ELM 随机初始化隐单元参数导致的问题,采用基于群体演化的 PSO 算法进行优化.

设 D 维搜索空间有 N 个粒子,第 i 个粒子位置和速度分别为 $\boldsymbol{X}_i=(x_{i1},\cdots,x_{id},\cdots,x_{iD})$ 和 $\boldsymbol{V}_i=(v_{i1},\cdots,v_{id},\cdots,v_{iD})$,每个粒子能记忆粒子自身历史最优位置 $\boldsymbol{P}_i=(p_{i1},\cdots,p_{id},\cdots,p_{iD})$、个体极值 p_{best}、群体历史最优位置 $\boldsymbol{P}_g=(p_{g1},\cdots,p_{gd},\cdots,p_{gD})$ 和全局极值 g_{best},每次迭代过程中,粒子速度和位置分别通过方程(7)、(8)更新,

$$v_{id}(t+1)=\omega v_{id}(t)+c_1r_1(p_{id}-x_{id})+c_2r_2(p_{gd}-x_{id});\tag{7}$$

$$x_{id}(t+1)=x_{id}(t)+v_{id}(t+1),\tag{8}$$

式中:ω 为惯性权重;c_1、c_2 为学习因子;r_1、r_2 为[0,1]上相互独立的随机数.

针对 PSO 易过早收敛陷入局部最优且无法自动跳进的问题,采用增强领导粒子的 PSO 算法(enhanced leader PSO,ELPSO),通过提高群体领导粒子的质量,增强 PSO 的搜索性能.

2.3 ELPSO-ELM 混合智能算法

以 ELM 的输入权值矩阵 $\boldsymbol{w}_i$ 和隐单元阈值 $\boldsymbol{b}_i$ 为 ELPSO 的粒子,设计 ELPSO-ELM 算法模型.

算法 1 ELPSO-ELM 算法

输入:数据集实例 $(\boldsymbol{x}_i,\boldsymbol{y}_i)$

输出:概率矩阵 $\boldsymbol{Y}$

Step1:初始化算法参数.粒子群规模 N 为 40,迭代次数 $t_{\max}=30$,惯性权重 $\omega_{\max}=0.9$、$\omega_{\min}=0.4$,$c_1=c_2=2$,r_1、r_2 为[0,1]上随机数;ELM 隐单元个数 k 设定为 20.PSO 维数计算公式为 $D=k(n+1)$,n 为每个粒子的维数.第 i 个粒子

表示为:$\boldsymbol{\theta}^i = [\omega_{11}^i, \cdots, \omega_{1k}^i \cdots \omega_{n1}^i, \cdots, \omega_{nk}^i, b_1^i, \cdots, b_k^i]$, ω^i、b^i 取值范围分别为[-1,1]、[0,1].

Step2:选择 sigmoid 函数为 ELM 算法、激活函数,根据式(3)~(5)计算输出权重矩阵,反推得到输出矩阵 $\boldsymbol{Y}'$,计算 $\boldsymbol{Y}'$ 与训练数据真实值 $\boldsymbol{Y}$ 的均方根误差(RMSE),以 RMSE 为粒子的适应度值,以具有较小适应度值的粒子为更优粒子[13].

Step3:根据式(7)、(8)更新粒子. ELPSO 算法中,每次迭代过程中对领导粒子 P_g 依次进行五阶段连续变异,分别为 Gaussian 变异、Cauchy 变异、P_g 各维 opposition-based 变异、P_g 整体 opposition-based 变异和 DE-based 变异,每次变异选取适应度值最小的粒子为种群领导粒子.

Step4:判断是否达到最大迭代数. 若达到,则输出全局最优粒子;若未达到,先利用式 $\omega = \omega_{\max} - (\omega_{\max} - \omega_{\min}) \times t/t_{\max}$ 更新惯性权重,再返回 Step2 继续执行.

Step5:最终 ELPSO 输出的粒子即为 $\boldsymbol{w}_i$ 和 $\boldsymbol{b}_i$ 的最佳取值,再用 ELM 建立预测模型,导入测试数据进行回归计算,得到真实数据集的预测结果.

3　实验分析

3.1　数据集

以 UCI 机器学习库的 Statlog (Heart)[14] 数据集和 Heart Disease Database[15] 验证模型,两数据集分别包含 270 和 820 组实例,每组实例包含 13 个属性和 1 个分类标签值. 数据的属性如表 1.

表 1　数据集共同的特征属性

Tab.1　The same attributes in two datasets

编号	属性名	类型	编号	属性名	类型
1	Age	连续	8	Thslach	连续
2	Sex	二分类	9	Exang	二分类
3	Cp	四分类	10	Oldpeak	连续
4	Trestbps	连续	11	Slope	有序三分类
5	Chol	连续	12	Ca	连续
6	Fbs	二分类	13	Thal	三分类
7	Restecg	三分类	14	标签值	1-不得病 2-得病

3.2　实验设计

实验一:利用二元 Logistic 回归分析训练数据,求得参数 $(\alpha, \beta_1, \cdots, \beta_{11})$ 值,根据式(1)建立 Logistic 预测方程,代入测试数据的属性值,对比所得预测值与真实标签值计算出预测正确率.

实验二:对多分类属性设置哑变量,然后对数据利用实验一方法进行 Logistic 回归分析.

实验三:利用 ELPSO-ELM 对原始数据进行实验,算法参数初始化如 Step1. 因 ELPSO 每次寻优参数不同,导致模型预测结果有所波动,因此选择 30 次预测结果的均值作为最终值.

实验四:为保证在同一计算代价下验证算法,组建 ELM 随机搜索组合算法(ELMs). ELPSO-ELM 单次运行适应度值的评价次数为(40+5)×30;组合 PSO-ELM 算法作对比,其粒子规模设置为 30、迭代 45 次;ELMs 算法中每次 ELM 运行次数为 45×30 次,选择其中训练误差最优的测试结果作为 ELMs 算法的输出.

3.3　实验结果

(1)统计 Logistic、Dv-Logistic、ELM、ELMs、PSO-ELM、ELPSO-ELM 模型各 30 组预测结果的真阳性(TP)、假阳性(FP)、真阴性(TN)、假阴性(FN),利用均值计算灵敏度、特异性及正确率,结果见表 2.

(2)ROC 曲线. 利用受试者工作特征曲线(ROC)[16] 评定预测模型性能,以 AUC 大小表示 ROC 曲线结果的好坏. AUC 值在 0.5~0.7 时具有较低的准确性,在 0.7~0.9 时具有一定的准确性,在 0.9 以上时具有较高的准确性. 各模型 ROC 曲线如图 1.

3.4　ELPSO-ELM 模型性能参数分析

3.4.1　ELPSO 参数分析

(1)惯性权重 ω. ω 一般取值在[0,1],稍大的 ω 利于扩大群体搜索范围,稍小的 ω 利于收敛到最优位置. 据此,使 ω 值随迭代次数线性变小,以保证 ELPSO 具有最优的效果.

(2)学习因子 c_1、c_2. 由式(7)知,c_1 过大则粒子较多地在局部范围搜索,c_2 过大则粒子过早收敛,结合大量实验,最终选定 $c_1 = c_2 = 2$.

(3)粒子群规模. 一般优化问题在[20,50]取值,粒子数量设置过小,对结果偏差影响较大. 综合考虑 ELPSO 的优化维度,平衡模型稳定性、正确率和快速性,将粒子数设置为 40.

3.4.2　ELM 参数及结果影响

由于 ELM 随机初始化隐单元参数,预测模型稳定性较低. 在各模型最佳隐单元个数下做30 次重复实验,预测结果以方差表示,如表 2.

对比各模型方差可知,ELM 结果波动较大,且正确率不高;ELMs 因选择多次运行中的最优值,结果变动不大,但正确率不高;ELPSO-ELM 在保证模型稳定性的同时较大地提高了正确率.

表2 各模型预测结果对比

Tab. 2 Results contrast of each prediction model

数据集	算法	隐单元数	TP	FP	TN	FN	灵敏度/%	特异性/%	正确率/%	AUC	预测结果方差/%
Statlog (Heart)	Logistic	—	20	3	28	19	51.28	90.32	68.57	0.763 4	—
	Dv-Logistic	—	28	1	30	11	71.80	96.77	82.86	0.878 4	—
	ELM	130	21	4	27	18	53.85	87.10	70.00	0.784 1	64.299
	ELMs	130	26	3	28	13	66.67	90.32	77.14	0.832 1	5.736
	PSO-ELM	50	27	6	25	12	69.23	80.65	74.28	0.795 7	31.162
	ELPSO-ELM	20	30	1	30	9	76.92	96.77	85.71	0.902 4	12.032
Heart Disease Database	Logistic	—	60	31	107	52	53.57	77.55	66.80	0.670 0	—
	Dv-Logistic	—	81	19	119	31	72.32	86.23	80.00	0.799 9	—
	ELM	150	59	35	103	53	51.79	74.64	64.80	0.642 3	59.670
	ELMs	150	76	28	110	36	67.85	79.71	74.40	0.720 1	3.446
	PSO-ELM	60	73	18	120	39	65.18	86.96	77.20	0.736 7	28.327
	ELPSO-ELM	20	87	15	123	25	77.68	89.13	84.00	0.842 3	7.051

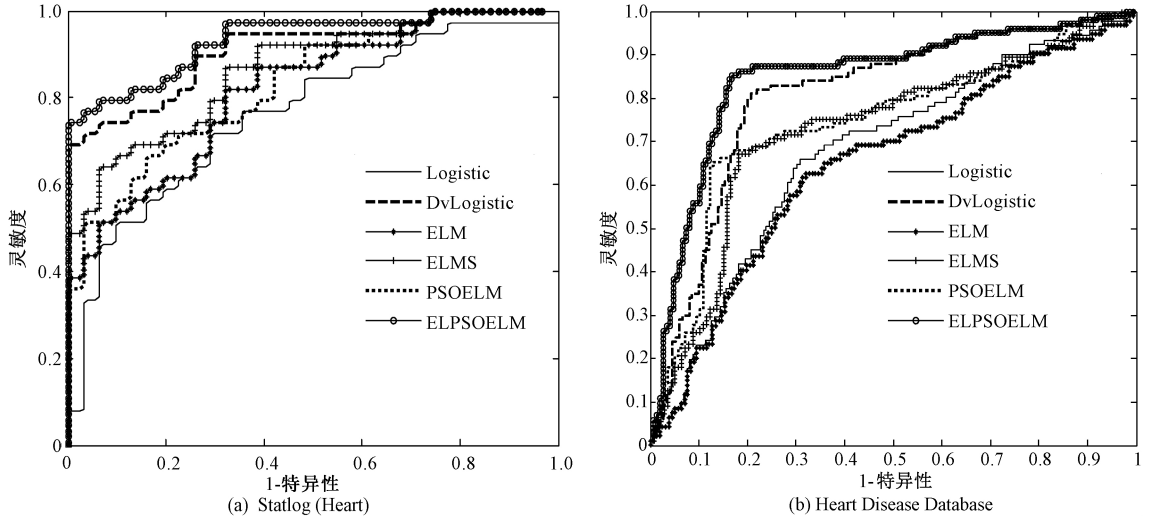

图1 ROC曲线图对比

Fig. 1 Comparison of ROC curves

以Statlog (Heart)数据集为例. 在相同的计算代价下,ELM隐单元个数对模型的影响如表3.

表3 隐单元个数对模型性能影响

Tab. 3 Influence of models performance by the number of hidden neurons

隐单元个数	准确率/%			运行时间/s		
	ELMs	PSO-ELM	ELPS-OELM	ELMs	PSO-ELM	ELPSO-ELM
10	70.00	70.00	82.86	36.71	29.37	35.69
20	72.86	70.00	85.71	37.79	31.05	40.18
40	71.43	71.43	84.28	38.69	35.82	42.89
50	72.86	74.28	84.28	39.82	37.47	49.12
130	77.14	70.00	81.42	47.27	48.60	79.34
150	74.28	68.57	78.57	50.88	52.09	85.08

4 结论

(1)对多分类变量设置哑变量能较大地提高Logistic回归模型的正确率.

(2)提出的ELPSO-ELM模型,在避免设置哑变量出现人为误差和较大工作量的同时,提高了模型预测正确率,AUC值大于或接近0.9,准确率较高.

(3)利用ELPSO优化ELM,相对于ELM方法和标准PSO优化方法,减少了隐层单元数目,提高了正确率,进一步证明所提方法的有效性.

(4)实验中,粒子群优化的计算维度达到240维,提高了数据处理复杂性,且ELPSO五阶段连续变异策略需进行大量迭代搜索,在提高模型稳定性和预测正确率的同时一定程度上降低了ELM的速度.

参考文献:

[1] 陈伟伟,高润霖,刘力生,等. 中国心血管病报告 2013 概要[J]. 中国循环杂志,2014,29(7):487 – 491.

[2] COHNJ N. Prevention of cardiovascular disease [J]. Trends in cardiovascular medicine, 2015,25(5):436 – 442.

[3] ZHAO D,LIU J,XIE W X, et al. Cardiovascular risk assessment: a global perspective [J]. Nature reviews cardiology, 2015, 12(5):301 – 311.

[4] HERVAS R, FONTECHAJ, AUSIN D, et al. Mobile monitoring and reasoning methods to prevent cardiovascular diseases[J]. Sensors, 2013, 13(5):6524 – 6541.

[5] GAMMON K. Mathematical modelling: forecasting cancer [J]. Nature, 2012,491(7425):S66 – 67.

[6] KLEBER ME, DELGADO G, GRAMMER T B, et al. Uric acid and cardiovascular events: a mendelian randomization study [J]. Journal of the American society of nephrology, 2015, 26(11):2831 – 2838.

[7] ALEXANDRE CM, FARDN. Binary logistic regression and PHM analysis for reliability data[J]. International journal of reliability, quality and safety engineering, 2014, 21(5):1 – 30.

[8] DEO RC. Machine learning in medicine [J]. Circulation, 2015, 132(20):1920 – 1930.

[9] HUANG G B, ZHOU H, DING X, et al. Extreme learning machine for regression and multiclass classification[J]. IEEE transactions on systems, man and cybernetics, Part B, 2012, 42(2):513 – 529.

[10] JORDEHI A R. Enhanced leader PSO (ELPSO): a new PSO variant for solving global optimization problems [J]. Applied soft computing, 2015, 26:401 – 417.

[11] TSOUAC M, CHIS P, HUANG D Y. EDLRT: Entropy-based dummy variables logistic regression tree [J]. Intelligent data analysis, 2010, 14(6):683 – 700.

[12] MAN Z H, KEVIN L, WANG D H, et al. A new robust training algorithm for a class of single-hidden layer feed-forward neural networks [J]. Neurocomputing, 2011, 74(16):2491 – 2501.

[13] 王杰, 苌群康, 彭金柱. 极限学习机优化及其拟合性分析[J]. 郑州大学学报(工学版), 2016,37(2):20 – 24.

[14] Statlog (Heart) Data Set [B/OL]. (2004)[2016-05-20]. http://archive. ics. uci. edu/ml/datasets/Statlog + Heart.

[15] Heart Disease Data Set [B/OL]. (1988-7)[2016-05-20]. http://archive. ics. uci. edu/ml/datasets /Heart + Disease.

[16] KHREICHW, GRANGERE, MIRI A, et al. Adaptive ROC-based ensembles of HMMs applied to anomaly detection [J]. Pattern recognition, 2012, 45(1): 208 – 230.

A CVD Prediction Model Based on Optimized Extreme Learning Machine

LU Peng[1,2], LI Qihang[1,2], SHANG Lijia[3], LI Xinjian[1], ZHANG Wei[1]

(1. School of Electrical Engineering, Zhengzhou University, Zhengzhou 450001, China;
2. Collaborative Innovation Center of Internet Medical and Healthcare in Henan, Zhengzhou 450001, China;
3. Primary and Secondary School Health Care in Beijing Dongcheng District, Beijing 100007, China)

Abstract: In order to increase the risk factors that could be accepted and reduce the data format requirements in cardiovascular disease (CVD) prediction models, machine learning algorithms were used to change the strict mathematical formulas of traditional CVD prediction models. Firstly, a CVD prediction model by extreme learning machine (ELM) algorithm based on single hidden layer feed-forward neural network (SLFNs) was proposed. Further more, an enhanced leader particle swarm optimization (ELPSO) through a five-staged successive mutation method was used, and the optimized strategy of PSO was also used to optimize the SLFNs hidden layer units parameters. The analysis results on Statlog (Heart) dataset and Heart Disease Dataset of UCI database indicated that the test accuracy of proposed ELPSO-ELM model could reach 85.71% and 84.00% respectively, the AUC (The area under the ROC curve) could reach 0.902 4 and 0.842 3 respectively. They were higher than conventional CVD prediction models. The proposed model relaxed the linear constraints of data format and more complex risk factors could be accepted.

Key words: cardiovascular disease; risk assessment; extreme learning machine; particle swarm optimization