文章编号: 1006-2475(2022)11-0069-06

基于改进遗传算法优化 BP 神经网络的糖尿病并发症预测模型

汪 敏,徐英豪,朱习军

(青岛科技大学信息科学技术学院,山东 青岛 266061)

摘要: BP 神经网络是在深度学习的研究中使用较为频繁的神经网络。本文提出一种改进遗传算法优化 BP 神经网络的算法(IGABP),利用遗传算法的全局搜索能力优化 BP 神经网络的初始结构。由于遗传算法易陷入局部最优解,影响自身的寻优能力,故对遗传算法进行改进,最后构建糖尿病并发症预测模型进而预测糖尿病并发症的发生。本文改进遗传算法的选择算子并改进自适应遗传算法的交叉及变异概率公式。通过构建预测模型,将改进后的 IGABP 与 BP、GABP、AGABP 进行比较。仿真实验结果表明,使用 IGABP 进行预测的准确率要明显优于 BP、GABP 与 AGABP,并且加快了网络的收敛速度。

关键词: 遗传算法; BP 神经网络; 自适应; 糖尿病预测; 数据预处理

中图分类号: TP301.6 文献标志码: A DOI: 10.3969/j.issn.1006-2475.2022.11.010

Prediction Model of Diabetic Complications Based on BP Neural Network Optimized by Improved Genetic Algorithm

WANG Min, XU Ying-hao, ZHU Xi-jun

(Deptartment of Information Science and Technology, Qingdao University of Science and Technology, Qingdao 266061, China)

Abstract: BP neural network is one of the most frequently used neural networks in deep learning research. In this paper, an improved genetic algorithm (IGABP) is proposed to optimize the initial structure of BP neural network. The genetic algorithm is easy to fall into local optimal solution, which affects its own optimization ability, so the genetic algorithm is improved, and finally the prediction model of diabetes complications is constructed to predict the occurrence of diabetes complications. The selection operator of genetic algorithm is improved, and the crossover and mutation probability formula of adaptive genetic algorithm is improved also. By building a prediction model, the improved IGABP is compared with BP, GABP and AGABP. The simulation results show that the prediction accuracy of IGABP is significantly better than that of BP, GABP and AGABP, and the convergence speed of the network is accelerated.

Key words: genetic algorithm; BP neural network; self adaptation; diabetes prediction; data preprocessing

0 引 言

2019 年,国际糖尿病联盟(IDF)[1] 发布了最新的全球糖尿病地图,地图显示我国是糖尿病人数最多的国家,且发病率不断增高。糖尿病是一种以身体内血糖含量升高为特点的慢性疾病,它的慢性并发症也是极其严重的一种疾病[2]。糖尿病并发症在早期不易被发现,所以很多人错过了最佳治疗时间[3-4]。糖尿病及其并发症已经成为严重威胁人们健康和生命的杀手,找到糖尿病最有可能引发的并发症和发病规律,对于患者及早做好预防措施,对于医生急时有针对性地给出治疗方案具有非常重要的意义。

近些年,我国对糖尿病及并发症应用于数据挖掘领域有了很大提升。陈淑良等人[5] 对某医院糖尿病患者的数据进行挖掘,建立 Logistic 回归模型以及多层感知器神经网络模型,试验得到 Logistic 回归模型

收稿日期: 2021-12-11; 修回日期: 2022-01-14

基金项目: 山东省产教融合研究生联合培养示范基地项目(2020-19)

作者简介: 汪敏(1996—),女,山东泰安人,硕士研究生,研究方向: 数据挖掘和图像处理,E-mail: 2334158648@qq.com; 徐英豪(1996—),男,硕士研究生,研究方向: 数据挖掘与图像增强,E-mail: xyh0609@163.com; 朱习军(1964—),男,教授,硕士生导师,博士,研究方向: 数据挖掘和图像处理,E-mail: zhuxj990@163.com。

对糖尿病的风险预测效果较好; 谭燕等人[6]选取 2 型糖尿病数据集作为研究对象,实验验证尿 IV 型胶原对糖尿病肾病的特异性诊断具有积极作用,成为预测早期 DN 的有价值的临床指标。苏萍等人[7]采用 Cox 比例风险回归构建糖尿病预测模型,以患者工作特征曲线下面积评价模型的预测效能,以十折法检验模型的稳定性,实验证明该模型在健康管理人群中具有较好的预测能力。

本文利用 BP 神经网络构建糖尿病并发症预测模型,采用遗传算法优化 BP 神经网络,通过改进遗传算法的选择方式、改进遗传算法的交叉及变异概率公式,以全局搜索的方式确定最佳的初始权值及阈值,从而在一定程度上提高了糖尿病并发症预测模型预测的准确率,最终通过仿真实验来对该模型进行评估。

1 数据预处理

1.1 原始数据分析

实际生活中,直接获取的原始数据并不能直接用于分析,因为这些数据大部分都是不完整不一致且极易受到噪声侵扰的。本文所使用的数据集为青岛市某三甲医院 HIS 系统中的糖尿病诊断数据,由于直接在医院获取的数据集中数据信息太多,故选取 1000 个样本,6 个特征属性,3 个标签属性,通过 jupyter notebook 对该样本进行数据预处理操作。选取详情见表 1,选取数据集整体分布如图 1 所示。

表 1 选取数据集属性

属性名称	属性描述
AGE	年龄
HEIGHT	身高
WEIGHT	体重
BLOOD_DRESS	血压
GLU	血糖
DIAGNOSIS	确诊病症
COMP_01	并发症 1
COMP_02	并发症 2
COMP_03	并发症 3

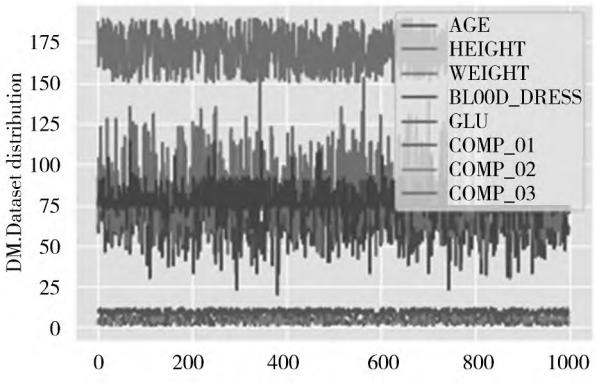

图 1 糖尿病数据集分布

所选属性中,特征属性分别为: 年龄、身高、体重、血压、血糖、确诊病症,标签属性为并发症 1、并发症 2、并发症 3。

关于标签属性的并发症,通过 jupyter notebook 工具,以分组计数的形式,统计所有并发症名称及数量,统计结果见表 2。

表 2 各类糖尿病并发症及数量统计

编号	并发症 1	数量
1	足病	57
2	周围血管病变	124
3	肾病	223
4	伴神经并发症	210
5	酮症酸中毒	47
6	伴眼并发症	146
7	心脏病	99
8	低血糖性昏迷	14

1.2 异常/缺失值处理

对于数据集中的异常及缺失值,利用插补法填补[8]。图 2 为数据集箱线图,通过设置数据标准来反映数据面貌。通过该图对数据集进行异常分析。

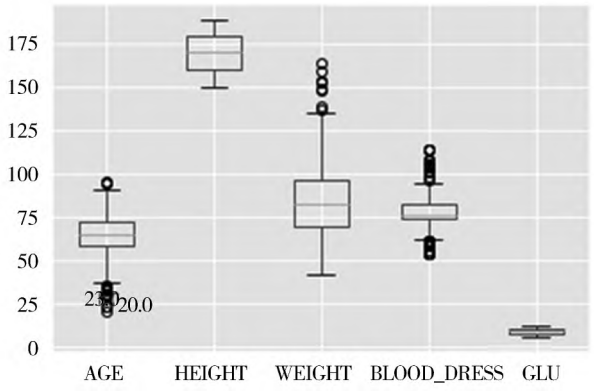

图 2 数据集箱线图

最后将数据集中不合理的数据或者是存在缺失值的数据(如年龄过小、血压为空等情况) ,以平均值的方式进行替代。

1.3 数据标准化

数据标准化(normalization) 是指将原始各指标数据按照比例缩放,去除数据单位限制,转化为无量纲的纯数值,以便于不同单位或量级的指标能够进行比较和加权[9]。对于该数据集,采用 min-max 标准化的方法,进行线性变化,使其落到区间 [0,1] 中。

对序列 $x_1, x_2, \cdots, x_n$ 进行变换:

$$y_i = \frac{x_i - \min\{x_j\}}{\max\{x_j\} - \min\{x_j\}} \tag{1}$$

新序列 $y_1, y_2, \cdots, y_n \in [0,1]$ 并且是无量纲的。(注: $\max\{x_j\}$ 为样本属性的最大值,$\min\{x_j\}$ 为样本属性的最小值) 。

1.4　特征相关性分析

特征相关性分析主要分析特征两两间的关系[9]。本文数据集相关性分析使用 jupyter notebook，通过 corr 函数，得到皮尔逊相关系数(Pearson correlation coefficient)。图 3 为糖尿病数据集各个特征相关关系热力图。

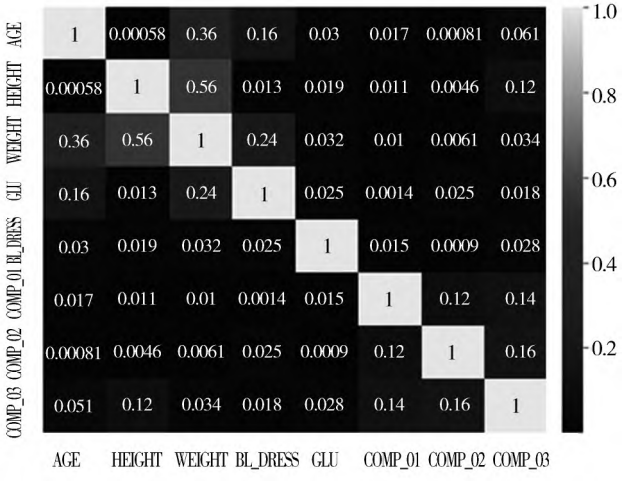

图 3　特征关联分析热力图

由图 3 可知：血糖与并发症 1 这 2 个特征间的关联程度较大；血糖与体重这 2 个特征间的关联程度较大等。

通过数据预处理，获得一个新的糖尿病数据集，将该数据集用于后续 BP 神经网络的建模工作，数据集划分为训练集与测试集，来验证构建预测模型的预测准确度。

2　改进遗传算法优化 BP 神经网络

BP 神经网络是基于误差反向传播的多层前馈神经网络[10-11]，在机器学习方面已经比较成熟。图 4 为 BP 神经网络的基本结构。它是由信息正向传播及误差反向传播 2 个过程组成。

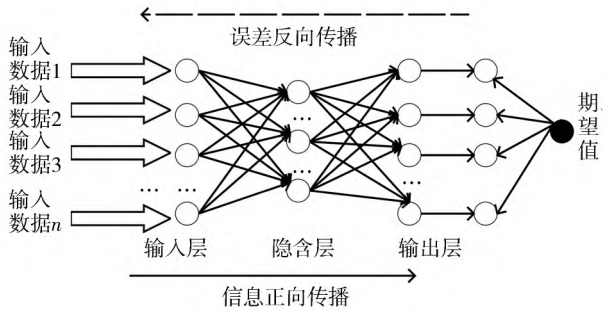

图 4　BP 神经网络基本结构

由于 BP 神经网络对于非线性数据集的预测效果较为理想，有非线性逼近以及自适应学习能力，而本文所使用到的糖尿病数据集为非线性数据集，故选用 BP 神经网络对糖尿病并发症进行预测。

传统 BP 神经网络具有极易陷入局部极小值的缺点，为了提高该模型的训练准确度，使用遗传算法的全局搜索能力确定 BP 神经网络最优初始权值及阈值。

2.1　改进遗传算法

1992 年，Holland[12] 提出模拟自然界遗传机制和生物进化论机制的遗传算法(Genetic Algorithms)。该算法具有全局搜索能力，通过遗传学中的选择、交叉、变异操作，对个体进行筛选，最终收敛到全局最优解。

前人已经对遗传算法进行了多种改进，像自适应概率、父子代竞争、保留最优父代等，这些改进在寻找最优解的问题上相对于传统的遗传算法有一些进步，但是基本都不会有很明显的下一代非常优秀于上一代的情况。主要原因便在于遗传算法与实际生物界生物进化相比，它在选择过程中并没有引导交叉变异，交叉是随机进行的交叉，变异是随机进行的变异，实际收敛速度并没有很快[13-19]。

本文对遗传算法的选择方式以及对自适应遗传算法交叉、变异概率公式进行改进，提高遗传算法寻找最优解的能力。

2.1.1　编码

一般用的比较多的编码方法有 2 种，分别为二进制编码和实数编码。由于 BP 神经网络的权值与阈值介于 -1 到 1 之间，故采用实数编码较为合适。

2.1.2　适应度

好的适应度对预测结果起积极作用，反之则起消极作用，故误差与适应度应为反比的关系，所以采用均方误差的倒数作为适应度函数。公式(2)为适应度函数公式，x 表示种群个体。

$$F(x) = \frac{1}{E(x)} \tag{2}$$

2.1.3　改进选择方式

本文对于选择操作，采用通过计算每一个个体的适应度值，然后将这些个体通过适应度值的大小进行顺序排序，适应度值最大的前 2 个个体直接遗传至下一代，剩余个体中根据适应度值均等分为一级、二级、三级 3 个等级。一级的复制 2 份遗传至下一代，二级的复制一份遗传至下一代，三级的直接舍去不使用。该选择方法使种群整体的平均适应度得到改善，并且维持了种群的多样性。以下是对该种方法的图形化展示：

1) 确定一个种群个数为 14 的初始种群。

5	32	16	14	30	7	29	4	11	35	8	17	13	19
1	2	3	4	5	6	7	8	9	10	11	12	13	14

2) 计算种群中每一个个体的适应度值，并且依

据适应度值由大到小进行顺序排序。

35	32	30	29	19	17	16	14	13	11	8	7	5	4
10	2	5	7	14	12	3	4	13	9	11	6	1	8

3) 将排序后处于前 2 名的个体直接遗传至下一代。

35	32
优秀	

4) 将剩余个体均分为 3 等份，分别定义为一级、二级与三级。

35	32	30	29	19	17	16	14	13	11	8	7	5	4
优秀		一级				二级				三级			

5) 将一级的复制 2 份，二级的复制一份，三级的直接淘汰。

30	29	19	17	30	29	19	17	16	14	13	11	8	7	5	4
一级的复制 2 份								二级的复制一份				三级的直接舍去			

2.1.4 改进自适应遗传算法交叉/变异概率公式

传统的遗传算法采用固定的交叉、变异概率，概率过大过小都会影响算法性能，Srinivas 等人[20-22]提出了自适应遗传算法(Adaptive Genetic Algorithm，AGA)，可以自适应调整交叉、变异概率。当个体适应度大于种群平均适应度时，得到一个较小的交叉、变异概率；当个体适应度小于种群平均适应度，得到一个较大的交叉概率、变异概率。较小的变异概率可以将优秀的个体保留下来，较大的变异概率可以加速变异得到新个体。具体公式如下：

1) 交叉概率：

$$P_c=\begin{cases}\dfrac{k_1(f_{\max}-f')}{f_{\max}-f_{\text{avg}}}, & f'\geqslant f_{\text{avg}}\\ k_2, & f'<f_{\text{avg}}\end{cases} \tag{3}$$

2) 变异概率：

$$P_m=\begin{cases}\dfrac{k_3(f_{\max}-f')}{f_{\max}-f_{\text{avg}}}, & f\geqslant f_{\text{avg}}\\ k_4, & f<f_{\text{avg}}\end{cases} \tag{4}$$

式中，$f_{\max}$为种群中的最大适应度值，f_{avg}为种群中的平均适应度值，f'为交叉的较大适应度值，f为个体变异适应度值。使用交叉、变异概率函数不断自适应调整，加速遗传算法的收敛性，有效避免遗传算法陷入局部最优解，更快地找到全局最优解。

本文引入种群适应度的平均值 EX 及适应度值的离散程度 DX 来优化自适应遗传概率与变异概率公式：

$$\text{EX}=f_{\text{avg}}=\frac{f_1+f_2+\cdots+f_M}{M} \tag{5}$$

$$\text{DX}=\frac{f_1^{\,2}+f_2^{\,2}+\cdots+f_M^{\,2}}{M}-f_{avg}^{\,2} \tag{6}$$

$$\partial=\frac{\text{EX}+1}{\sqrt{\text{DX}}} \tag{7}$$

其中，f_i代表种群中适应度的值。随着不断进化，种群的平均适应度值不断增大，种群间的差异越来越小，所以种群的整体离散程度是不断减小的。综上，自定义系数 ∂ 是不断增大的。

最后通过自定义系数对原始的自适应遗传算法做改进，如下为改进后的公式：

$$P_c=\begin{cases}\dfrac{k_1}{1+e^{\frac{k_2}{\partial}}}\times\dfrac{f_{\max}-f'}{f_{\max}-f_{\text{avg}}}+k_3, & f'\geqslant f_{\text{avg}}\\ k_4, & f'<f_{\text{avg}}\end{cases} \tag{8}$$

$$P_m=\begin{cases}\dfrac{k_5}{1+e^{\frac{k_6}{\partial}}}\times\dfrac{f_{\max}-f}{f_{\max}-f_{\text{avg}}}+k_7, & f\geqslant f_{\text{avg}}\\ k_8, & f<f_{\text{avg}}\end{cases} \tag{9}$$

当 $f'\geqslant f_{\text{avg}}$ 时，$\dfrac{f_{\max}-f'}{f_{\max}-f_{\text{avg}}}\in[0,1]$。

由于 ∂ 是不断增大的，故 $\dfrac{k_1}{1+e^{\frac{k_2}{\partial}}}\in\left(0,\dfrac{k_1}{2}\right)$，故 $P_c\in\left(k_3,\dfrac{k_1}{2}+k_3\right)$，同理可得 $P_m\in\left(k_7,\dfrac{k_5}{2}+k_7\right)$。

该改进有效避免了种群在初始进化时形成局部收敛，陷入局部最优解的问题，提升了该算法的寻优性能。

2.2 改进的遗传算法优化 BP 神经网络

使用改进后的遗传算法优化 BP 神经网络，确定最佳初始权值及阈值，解决传统神经网络随机给定初始权值及阈值以及极易陷入局部最小值等缺点。如图 5 所示为改进后的遗传算法优化 BP 神经网络的算法流程图。

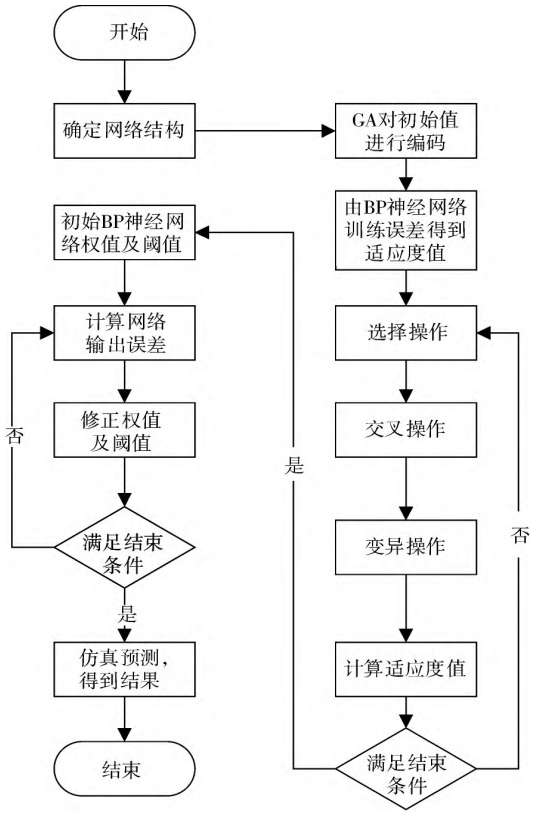

图 5 改进遗传算法优化 BP 神经网络算法流程图

3　仿真实验

为了验证改进后的遗传算法对糖尿病并发症有良好的预测效果，因此进行了仿真实验。数据集为上述预处理后的数据集，在 Matlab2016a 中对 BP 神经网络-糖尿病并发症预测模型进行训练，实验结果表明改进后的遗传算法优化 BP 神经网络，训练模型的预测结果要好于未改进的。

在本文仿真实验中，预处理后的糖尿病数据集共 1000 个样本，前 980 个样本作为训练集，后 20 个样本作为测试集，通过年龄、身高、体重、血压、血糖来预测未来可能会患有的并发症（确诊病症已筛选全部为糖尿病，故在这里不作为预测属性出现）。所以输入层节点为 5 个，输出层节点为 1 个。关于隐含层在这里不固定，自己进行手动设置，依据经验公式[23]：

$$h = \sqrt{m+n} + a \tag{10}$$

其中，m 为输入层节点数，n 为输出层节点数，a 为 1～10 之间取常数。

采用传统的 BP 神经网络进行糖尿病并发症的预测，在 Matlab 中设置如下基本参数：

1) net. trainParam. epochs = 100；为迭代次数。

2) net. trainParam. lr = 0. 1；为学习率。

3) net. trainParam. goal = 0. 001；为训练目标误差。

4) net. trainParam. show = 200；为训练结果显示。

使用 Matlab 对 BP 神经网络进行训练，预测结果如图 6 所示。

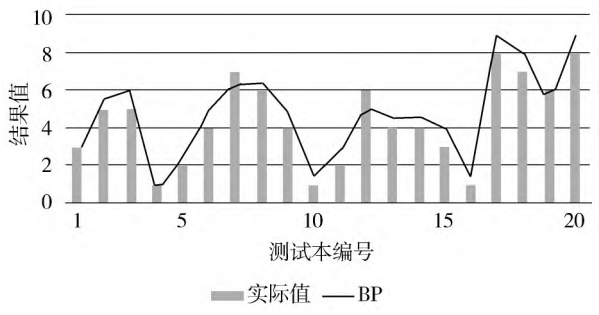

图 6　BP 神经网络预测值与实际值对比图

分别使用遗传算法（GA）、自适应遗传算法（AGA）、改进遗传算法（IGA）优化 BP 神经网络，在 Matlab 中对 GABP、AGABP、IGABP 神经网络进行训练，预测结果与实际值比较结果如图 7～图 9 所示。

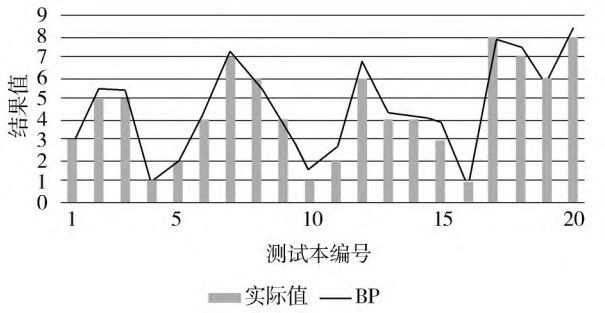

图 7　GABP 预测值与实际值对比图

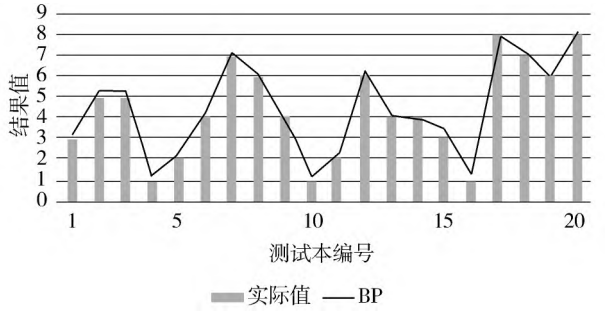

图 8　AGABP 预测值与实际值对比图

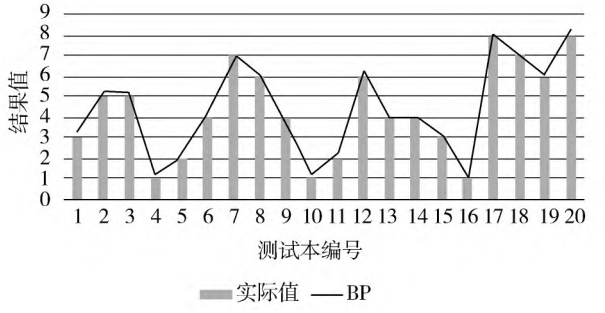

图 9　IGABP 预测值与实际值对比图

从图中可以清晰看到，这 4 种算法的预测结果依次逐渐接近于实际结果。即对于预测结果的好坏，BP < GABP < AGABP < IGABP。

图 10 为 4 种算法预测输出的误差百分比折线图。由图可知，对于传统的 BP 神经网络，其预测误差波动范围比较大，GABP 的波动范围其次，AGABP 与 IGABP 的误差波动范围均比较接近 0 值，在 0 附近小范围波动，且 IGABP 比 AGABP 更接近于 0 值。

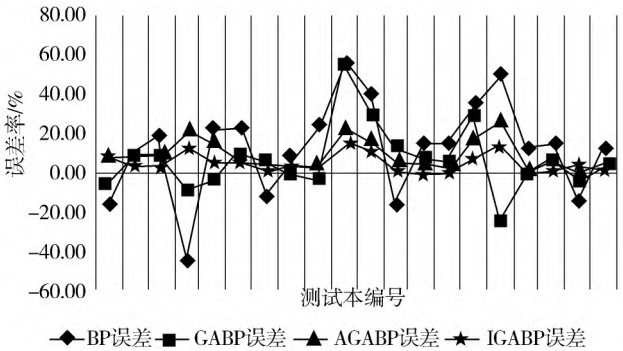

图 10　4 种预测误差对比图

预测结果如表 3 所示。

表 3 预测结果对比

预测模型	平均绝对误差 MAE/%	均方根误差 RMSE
BP	22.31	0.264342
GABP	11.84	0.172177295
AGABP	8.60	0.115457834
IGABP	4.84	0.067460092

由表 3 中统计数据可知，IGABP 神经网络预测模型的仿真效果最好，平均绝对误差 MAE 达到 4.84%，均方根误差 RMSE 也最低；传统的 BP 神经网络预测结果误差最大，平均绝对误差 MAE 为 22.31%，均方误差 RMSE 最高；GABP 神经网络预测模型与 AGABP 神经网络预测模型与传统的 BP 神经网络预测模型相比在预测结果上有了一定的改善，但均没有达到 IGABP 神经网络预测模型的精度，因此 IGABP 在预测的精度上要优于其他 3 种模型。

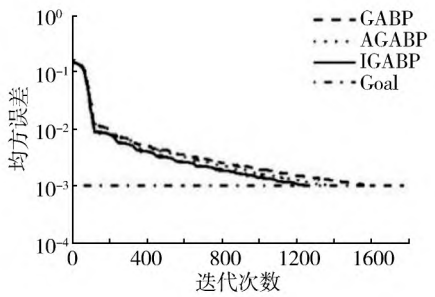

图 11 网络训练误差曲线

如图 11 所示，为 GABP、AGABP 与 IGABP 的网络训练误差曲线，由图中曲线可知 IGABP 的收敛速度要明显好于 GABP 与 AGABP，故 IGABP 的收敛性更好。

4 结束语

本文改进遗传算法优化 BP 神经网络构建预测模型来提高预测准确度，使用数据预处理后的数据集对糖尿病并发症进行预测。仿真实验，结果表明改进后的算法优化 BP 神经网络所构建模型的预测准确率最高。

参考文献：

[1] SAEEDI P, PETERSOHN I, SALPEA P, et al. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9 th edition [J]. Diabetes Research and Clinical Practice, 2019,157: 107843.

[2] 杨文英. 中国糖尿病的流行特点及变化趋势 [J]. 中国科学(生命科学), 2018,48(8):812-819.

[3] 韦哲,石栋栋,王能才,等. 基于思维进化算法优化的 BP 神经网络对糖尿病并发症的预测研究 [J]. 中国医学装备, 2020,17(10):1-4.

[4] 刘文婷,吴艳艳,殷丽,等. 糖尿病和空腹血糖受损对缺血性脑卒中复发的影响 [J]. 中国卫生工程学, 2021,20(3):518-519.

[5] 陈淑良,常红,王冬平,等. 基于数据挖掘的 2 型糖尿病风险预测模型的建立和应用 [J]. 糖尿病新世界, 2019,22(4):1-3.

[6] 谭燕,杨永年,张志刚,等. 尿Ⅳ型胶原:一种早期糖尿病肾病的预测指标 [J]. 中华医学杂志, 2002(3):389-394.

[7] 苏萍,杨亚超,杨洋,等. 健康管理人群 2 型糖尿病发病风险预测模型 [J]. 山东大学学报(医学版), 2017,55(6):82-86..

[8] 李仪,林建君,朱习军. 基于改进 DNN 的糖尿病预测模型设计 [J]. 计算机工程与设计, 2021,42(5):1418-1424.

[9] 徐敏,王科,戴浩然,等. 基于电子病历的乳腺癌群组与治疗方案可视分析 [J]. 浙江大学学报(理学版), 2021,48(4):391-401.

[10] 圣文顺,赵翰驰,孙艳文. 基于改进遗传算法优化 BP 神经网络的销售预测模型 [J]. 计算机系统应用, 2019,28(12):200-204.

[11] 凌晨,张骅月. 基于 BP 神经网络的黄金价格预测分析 [J]. 天津科技, 2014,41(1):68-73.

[12] HOLLAND J H. Adaptation in Nature and Artificial Systems [M]. MIT Press, 1992.

[13] 宋超,宋娟. 基于遗传算法优化和 BP 神经网络的短期天然气负荷预测 [J]. 工业控制计算机, 2012,25(10):82-84.

[14] 王倩,李风军. 改进的自适应遗传算法及应用 [J]. 重庆师范大学学报(自然科学版), 2021,38(2):14-19.

[15] 白鹏,王浩. 改进遗传算法优化的 BP 神经网络高炉煤气预测 [J]. 机械工程与自动化, 2021(2):77-79.

[16] 倪渊,李子峰,张健. 基于 AGA-BP 神经网络的网络平台交易环境下数据资源价值评估研究 [J]. 情报理论与实践, 2020,43(1):135-142.

[17] 席亮,王瑞东. 基于自适应遗传算法的神经网络结构优化算法 [J]. 哈尔滨理工大学学报, 2021,26(1):39-44.

[18] 汪静,罗维平,陈永恒. 基于神经网络的房价预测与分析 [J]. 襄阳职业技术学院学报, 2021,20(2):112-115.

[19] 刘森. 基于遗传算法改进的 BP 神经网络模型的高等教育人才影响预测研究 [J]. 科技经济导刊, 2021,29(12):149-151.

[20] SRINIVAS M, PATNAIK L M. Adaptive probabilities of crossover and mutation in genetic algorithms [J]. IEEE Transactions on Systems Man & Cybernetics, 2002,24(4):656-667.

[21] 刘芳,马玉磊,周慧娟. 基于种群多样性的自适应遗传算法优化仿真 [J]. 计算机仿真, 2017,34(4):250-255.

[22] 吴陈,王和杰. 基于改进的自适应遗传算法优化 BP 神经网络 [J]. 电子设计工程, 2016,24(24):29-32.

[23] 曹镓玺,王鑫,雷光春. 基于遗传算法优化 BP 神经网络的青藏高原海北高寒湿地 CO_(2) 通量模拟及其影响因子 [J]. 山东大学学报(理学版), 2021,56(5):33-50.