刘秀峰,刘 芬. 基于遗传算法优化的 BP 神经网络大肠癌证型分类[J]. 中华医学图书情报杂志,2018,27(3):14-18.

DOI:10.3969/j.issn.1671-3982.2018.03.004

· 研究与探讨 ·

基于遗传算法优化的 BP 神经网络大肠癌证型分类

刘秀峰,刘 芬

[摘要]针对医学领域传统 BP 神经网络应用于诊断辅助构建模型过程存在的疾病输入特征维数繁多而导致网络训练时间长、效率低、诊断模型泛化能力弱等问题,提出了将基于遗传算法降维优化的 BP 神经网络(GABP)诊断模型用于大肠癌虚实证型的分类研究。利用遗传优化算法从大肠癌的 28 项体征输入指标中筛选出的 10 个指标作为 GABP 诊断模型的输入,并与传统的未经优化的 BP 诊断模型进行对比,发现优化后的神经网络模型建模所需时间由 5.1844 秒缩短至 0.1976 秒,虚证正判率由 76.4567% 提升至 89.1809%,实证正判率由 64.8441% 提升至 70.1170%。仿真结果表明,基于 GABP 的神经网络泛化能力更好,分类效果更佳,为大肠癌中医证型的临床判别提供了很好的计算机网络模型。

[关键词]大肠癌;遗传算法;BP 神经网络;降维优化

[中图分类号]TP18;R735.3 [文献标志码]A [文章编号]1671-3982(2018)03-0014-05

Classification of large intestine carcinoma syndromes based on genetic algorithm-optimized back propagation neural network

LIU Xiu-feng, LIU Fen

(School of Medical Information Engineering, Guangzhou University of Chinese Medicine, Guangzhou 510006, Guangdong Province, China)

[**Abstract**]A diagnosis model of genetic algorithm (GA) dimension reduction-optimized back propagation (BP) neural network was established since the long time, low efficiency and poor capability of diagnosis model due to the large number of disease input feature dimensions when the traditional BP neural network was used in establishing diagnosis-aided model, which has been used in classification of large intestine carcinoma syndromes. Ten input feature indications, screened from the 28 input feature indications of large intestine carcinoma, were used as the input of GABP diagnosis model. The optimized GABP model was compared with the unoptimized BP diagnosis model, which showed that the modeling time of optimized BABP model was reduced to 0.1976 s from 5.1844 s, the positive diagnosis rate of deficiency syndrome was increased to 89.1809% from 76.4567 and that of excess syndrome was increased to 70.1170% from 64.8441%. Simulation analysis showed that the general ability and classification efficiency of GABP-based neural network are better than those of traditional BP neural network, and can thus provide a better computer network model for the differential diagnosis of large intestine carcinoma syndromes in clinical practice.

[**Key words**]Large intestine carcinoma; Genetic algorithm; BP neural network; Dimension reduction optimization

大肠癌是常见的消化道恶性肿瘤,包括结肠癌和直肠癌。近年来,随着人们生活方式和生活环境的改变,大肠癌在我国的发病率和死亡率呈现快速上升的趋势[1]。大肠癌属本虚标实之证,既有脏腑气血亏虚,又有气滞、血瘀、痰凝、湿毒等标实的情

[基金项目]广州中医药大学薪火计划资助项目"基于深度神经网络进行多层特征学习的大肠癌患者证候模型研究"(XH20160105)

[作者单位]广州中医药大学医学信息工程学院,广东 广州 510006

[作者简介]刘秀峰(1973-),女,江西樟树人,硕士,教授,研究方向:中医药数据处理与分析、肿瘤信息学、移动医疗。

况[2]。目前中医证候诊断标准存在证名不统一、证型的诊断标准没有考虑到疾病的影响，构成证的基本元素模糊不清，标准制定主要依据中医理论、文献及专家咨询带有很大主观成分等诸多问题，而临床上病例往往诸证夹杂，临床医师治疗经验不同，很难统一“标准”，“标准”在临床实践上也难以应用实施[3]。因此，利用计算机技术对中医药信息进行规范化的研究十分必要。

随着人工智能的发展，神经网络技术也不断成熟。由于其分类能力强且具有智能性，人工神经网络在医学诊断、预后、生存分析、临床决策支持、模式识别等领域得到广泛应用[4]。而 BP（Back Propagation）算法是基于梯度下降的，自身的缺陷也不可避免。如网络权值和阈值随机设定，当解空间多个局部极小时容易陷入局部极小而无法跳出[5]。此外，网络输出自变量之间多数情况下并非独立而存在冗余信息，会导致网络训练陷入过拟合状态且延长建模时间。本文针对传统 BP 算法用于分类建模时的缺陷，拟采用遗传算法（Genetic Algorithm，GA）优化后的 GABP（Genetic Algorithm Back Propagation）模型构建大肠癌的虚实证型分类器，以期促进对大肠癌虚实证型分类的统一标准研究。通过对优化后的神经网络进行前后对比分析，提升大肠癌分类器的有效识别能力，对制定大肠癌中医临床诊断标准具有指导意义。

1　资料与方法

1.1　数据来源

本文所用数据共 338 例，其中 150 例来源于广东省中医院临床病案，其余 188 例来源于中国知网全文数据库、万方数据知识服务平台及维普全文数据库文献数据。纳入标准为经细胞学或病理学检查诊断为结直肠癌或者已通过肠镜报告检查诊断为大肠癌及经手术切除原发病灶的大肠癌患者，排除病理诊断不明确且具有多原发性肿瘤或既往有恶性肿瘤的患者。综合上述纳入、排除标准，最终筛选出符合条件的 218 例大肠癌患者数据，其年龄分布于 25～80 岁，男性患者 108 例，女性患者 110 例。

1.2　数据量化

1.2.1　症状量化

在专家指导下，综合临床与文献两部分数据，选出与大肠癌患者密切相关的 28 项体征，包括大便脓血、脉沉、舌苔白、纳呆等症状。为突出中医术语的严谨规范性，按照《中医症状鉴别诊断学》（第二版）[6]标准依据，对症状的分类进行初步划分，并将已经规范好的大肠癌常见症状进行赋值量化，有该症状为 1，无该症状为 0。为防止量化过程中人工录入产生错乱，数据录入采用双边独立录入并进行一致性检验。

1.2.2　证型量化

参照 2008 年中华中医药学会发布的《中医诊疗指南》中提出的诊断、辨证标准，同时参照 1992 年全国大肠中医科研协作会议制定的证型划分标准和 2004 年杨金坤主编的《现代中医肿瘤学》[7]，结合本文所收集的实际病例情况，在多位专家进行判定之后，决定将本文筛选出的 218 例大肠癌患者归为脾失健运、脾虚夹瘀、湿热内蕴、气滞血瘀、瘀毒内阻、气血亏虚、脾肾阳虚、肝脾不调 8 种证型。考虑到分属于这 8 种证型的样本数据量较小，可能会对后续建模的分类效果有一定影响，因而将这 8 种证型进一步分为虚证与实证两种类型，且将虚证赋为 1，实证赋为 2，同样在赋值后对录入的数据进行一致性检验。

1.3　研究方法

人工神经网络是由大量的功能和形式比较简单的神经元互相连接而构成的复杂网络系统，网络可以看作是从输入到输出的一个非线性映射[8]。BP 神经网络具有的非线性映射能力能够保证其成功实现各种简单或复杂分类，它将信息分布式存储于连结权系数中，使网络具有较高的容错性和鲁棒性[9]。遗传算法是模拟自然界优胜劣汰生物进化过程的全局优化搜索算法，采用群体进化的方式对目标函数空间进行并行式搜索，根据个体适应度大小选择个体，保留竞争力强的基因，是一种搜索效率高、鲁棒性强的优化方法[10]。遗传算法广泛应用在机器学习、自适应控制、组合优化等领域，并取得了很好的效果[11]。本文结合遗传算法与 BP 神经网络的特点，首先利用遗传算法对自变量进行优化筛选，利用遗传算法优化 BP 神经网络的权值和阈值提高筛选效率，最终通过优化 BP 神经网络的结构和系数构建大肠癌虚实证型分类器，探讨大肠癌证型分类的有效方法。

2 基于遗传算法优化的过程

利用遗传算法选择特征维必须经过输入变量编码、初始种群产生、适应度计算、交叉变异选择、优化输出等过程[12]。由于 BP 算法本质上为梯度下降法，所要优化的目标函数非常复杂，容易出现局部最优、收敛速度慢等问题，而遗传算法是多点搜索，能够避免局部最优。此外，利用遗传算法取代一些其他的优化算法，是因为遗传算法的寻优能力可以获取最佳权重。利用遗传算法进行自变量降维与网络初始权值与阈值的优化步骤如下。

2.1 输入变量编码和种群初始化

本文在利用遗传算法进行自变量筛选的过程中，染色体编码采用二进制法，即将个体编码为一个二进制串。根据研究中的 28 项体征，此处染色体编码长度为 28，染色体的每一位对应一个输入特征维，若染色体某一位值为 1，说明该位置上相对应的输入自变量参与建模，若为 0 则不参与。而在利用遗传算法进行 BP 网络权值和阈值的优化过程中，我们对初始种群采取实数编码，每个个体由一组实数串组成。当 BP 网络结构确定时，个体编码长度也随权值和阈值个数确定。本文将种群大小初始化为 20，最大进化代数设置为 100，遗传算法以这 20 个个体作为初始点进行迭代。

2.2 适应度函数计算

BP 神经网络的误差绝对值越小越好，而在遗传算法中，适应度值越大越好[13]，因此采用以 BP 神经网络目标函数的倒数作为适应度函数。本文选取测试集数据误差平方和的倒数作为适应度函数，以便使遗传算法朝向适应度函数增大的方向进化。

2.3 选择交叉变异

选择交叉变异是遗传算法的核心。本文选择操作采用轮盘赌法，首先产生 0 与 1 之间的随机数确定种群中个体被选中次数，然后根据上述适应度函数计算，选择适应度大的进入下一代种群。对于交叉操作，降维过程选择单点交叉算子，变异点位置是随机产生的，且该位置上的基因值由 0 变成 1 或 1 变为 0。在权值和阈值优化过程中，交叉操作采用交叉算子，利用一对个体根据给定的概率重组产生新的种群后代。

2.4 优化结果输出

按照遗传算法进行层层迭代之后，当迭代次数达到最大进化代数时，进化终止，最终输出的此种群中适应度最好的个体对应输入变量的基因编号为 1、4、5、9、13、14、18、20、22、27，分别对应于大肠癌的腹泻、里急后重、大便秘结、脉沉、脉细、脉滑、舌紫、舌胖、舌苔白、纳差等 10 项体征，筛选出来参与建模的自变量不到原来总输入的一半，即收集统计的 28 项体征输入自变量之间并非独立，存在冗余信息，经遗传算法降维是非常必要的。遗传算法优化降维的过程如图 1 所示。由图 1 可知，最优个体是在种群进化到第 35 代后得到的。由于遗传算法存在一些随机因素，故种群的平均适应度函数均值在稳定中出现微小波动。

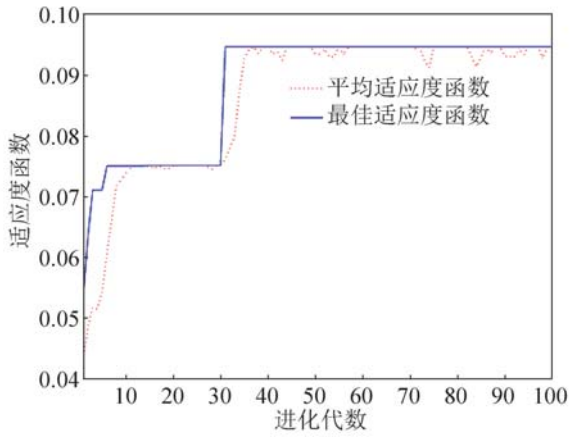

图 1 种群适应度函数进化曲线

3 建立基于遗传算法优化的 BP 神经网络

3.1 遗传算法优化 BP 神经网络流程

BP 神经网络对网络初始权值和阈值的依赖性比较强，以往大多随机选取易陷入局部最优。遗传算法是一种全局优化的自适应概率搜索算法，可以得到全局最优[14]。本文利用遗传算法对神经网络的初始权值和阈值进行优化寻优，基本流程如图 2 所示。

3.2 BP 神经网络结构及隐层设置

BP 神经网络是信号沿着前项传播，误差沿着反方向传播，不断调整权值与阈值使网络趋于收敛构建的过程。由于由输入层、隐藏层、输出层构成的 3 层 BP 神经网络可以逼近任意函数，故本文选取 BP 的 3 层拓扑结构进行大肠癌虚实分类器的构建。输入层由经遗传算法降维优化的10项体征设定，输出

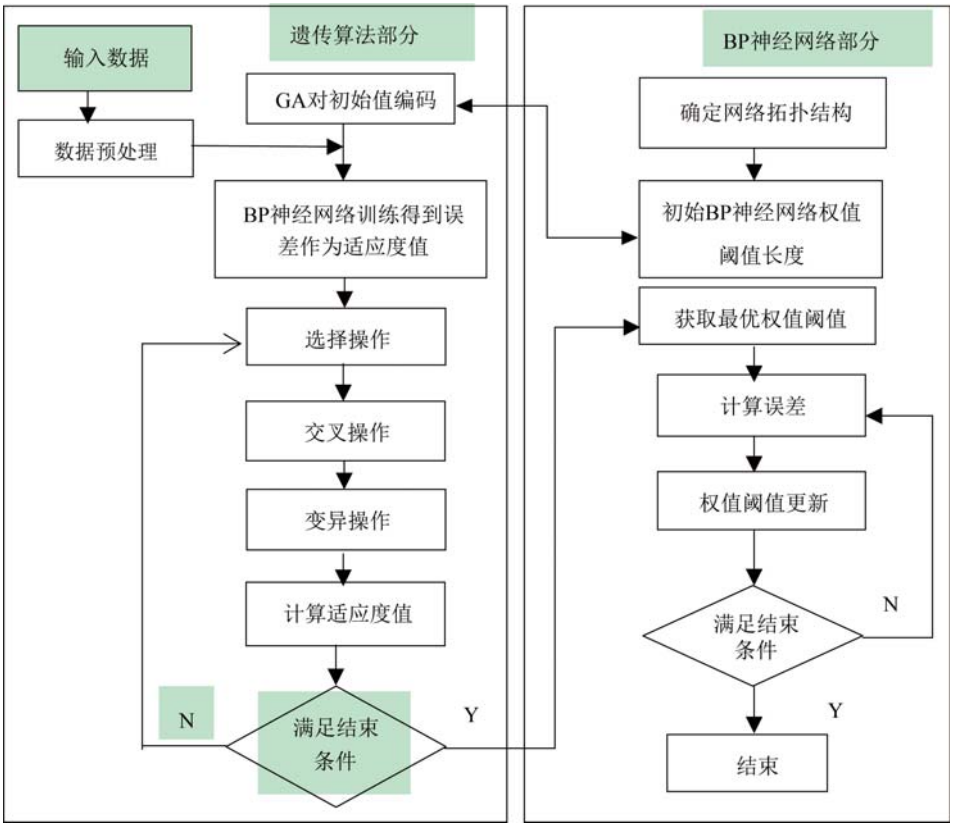

图 2　遗传算法优化神经网络流程

层由虚实两种证型作为输出。由于 BP 神经网络隐藏层节点通常由经验公式及误差对比调整得出，本文考虑到样本量偏少，为充分利用数据集，利用 MATLAB（matrix&laboratory）中的 crossvalind 函数采用五折交叉验证确定 GABP 神经网络建模过程中的最佳隐含层神经元个数，程序根据数据集实际情况设定最佳神经元个数在 10-28 之间循环寻找。结果显示当隐含层神经元个数为 15 时，网络的分类效果最好，故 GABP 神经网络模型的隐层个数确定为 15 个。因此本文中 BP 神经网络的基本模型确定为 10-15-2，且隐含层和输出层的传递函数都取 s 函数。

3.3　BP 神经网络输出方式

在进行网络训练之前，将虚实证型分别标记为“1”“2”。由于经 BP 神经网络输出的为非二值结果，即网络的期望输出只能取 1 或 2 作为已标记的大肠癌虚实证型分类，故在程序中进行四舍五入的设定，即当网络预测输出值介于 0.5～1.5 之间时视为虚证，若该值介于 1.5～2.5 之间时则视为实证，统计预测结果并输出正判率。

3.4　BP 神经网络训练及测试

将经遗传算法优化计算后筛选的输入自变量对应的数据提取出来建立新的 BP 神经网络。为了使训练的网络不失一般性，随机选取 165 例数据作为训练集，剩余的 53 例作为测试集，设定最大迭代次数为 1000，允许误差界值设为 0.001。当 2 次迭代结果的误差小于该值时，系统结束迭代计算，输出结果，并统计正判率。三次仿真结果的虚实两证型的正判率及建模时间如表 1 所示。

表 1　传统 BP 与 GABP 的 3 次仿真分类结果

模型	第一次仿真			第二次仿真			第三次仿真		
识别	虚证正判率/%	实证正判率/%	建模时间/秒	虚证正判率/%	实证正判率/%	建模时间/秒	虚证正判率/%	实证正判率/%	建模时间/秒
BP	77.142	61.1111	4.3524	76.4706	68.4211	6.8796	75.7576	65.0000	4.3212
GABP	91.4286	66.6667	0.1872	88.2353	73.6842	0.1872	87.8788	70.0000	0.2184

4 基于传统 BP 与经遗传算法优化的 GABP 神经网络分类模型的结果对比

经传统 BP 神经网络模型和遗传算法优化改进的 BP 神经网络模型(GABP)实现的的大肠癌虚实证型分类器结果平均值见表 2。两种模型样本总量均为 218 例,其中训练样本选取总样本的 75%(本文中选取 165 例),测试样本占总样本的 25%(本研究中选取 53 例)。从表 2 可以看出,经遗传算法优化改进后的 BP 神经网络分类效果明显优于传统 BP 神经网络,这是因为经遗传算法优化改进后,不仅克服了传统 BP 网络存在的缺陷,而且利用遗传降维有效提取了输入维度的信息,从而实现了高效的分类。此外,在利用两种模型构建分类器时,实证正判率明显低于虚证正判率,这是因为在 218 例总样本中,虚证有 143 例,而实证仅有 75 例,且在测试构建好的网络模型时,随机选取的测试集虚实证型的不均匀分布也会影响分类的最终效果。如果可以克服样本证型数量分布不平衡,实验分类效果可能更佳。

表 2 分类模型结果对比

模型类别	虚证正判率/%	实证正判率/%	建模时间/秒
BP	76.4567	64.8441	5.1844
GABP	89.1809	70.1170	0.1976

5 结语

实验结果表明,BP 神经网络分类器可在一定程度上克服中医症状与证型之间的不确定性与模糊性,能获得较高的分类正判率。引入遗传算法后对证型分类的效果有进一步的提高,说明人工神经网络与遗传算法相结合能有效将大肠癌表面体征与内在证候特征有机结合起来对大肠癌证型进行预测分类,具有较强的研究价值与应用意义,对大肠癌中医的临床诊断有一定的借鉴意义。不足之处在于,本文是通过遗传算法将自变量降维和优化 BP 神经网络的初始权值和阈值来提高网络性能的,也可利用遗传算法过程中不同编码方式达到优化效果。此外,为验证经遗传算法优化改进后的 BP 神经网络的分类效果,我们也构建了 LVQ(Learning Vector Quantization)神经网络用来进行对比。利用 LVQ 神经网络无需对数据进行归一化,而只需通过计算输入向量与隐含层神经元间的距离,便可完成模式识别构建大肠癌分类器。实验结果显示,LVQ 神经网络得到的虚证正判率达到 81.5118%,实证正判率达到 67.4537%,略优于传统 BP 的分类效果,但依旧低于经遗传算法优化的 BP 神经网络分类效果。对比实验很好地表明了虽然 LVQ 神经网络也是一种很好的分类算法,但 GABP 模型用于大肠癌虚实证型的分类拥有更高的分析速度与准确度。

【参考文献】

[1] 张 华,王晶囡,宋丹丹,等. 大肠癌的贝叶斯预测[J]. 高师理科学刊,2018,38(1):13-17.

[2] 陈 叶,刘金涛,朱 源,等. 大肠癌中医辨证及治疗概况[J]. 中国肿瘤,2015,24(4),319-324.

[3] 郭秋生,吕仙梅. 大肠癌中医辨证分型概况[J]. 临床合理用药杂志,2012,5(5):178-180.

[4] 辛基梁. 人工神经网络在中医临床辨证模型研究中的应用[D]. 福州:福建中医药大学,2017.

[5] 张静波. 以遗传算法改进为核心的神经网络优化[J]. 中国新通信,2018,20(1):42.

[6] 姚乃礼. 中医症状鉴别诊断学[M]. 北京:人民卫生出版社,2000:27-35.

[7] 杨金坤. 现代中医肿瘤学[M]. 上海:上海中医药大学出版社,2004.

[8] 罗 新,牛海清,林浩然,等. BP 和 RBF 神经网络在气隙击穿电压预测中的应用和对比研究[J]. 电工电能新技术,2013,32(3):110-115.

[9] 钟淑瑛,李陶深. 基于 MATLAB 的 BP-LVQ 神经网络组合分类模型[J]. 计算机技术与发展,2006,16(2):114-116.

[10] 孙文兵. 遗传算法降维优化的 BP 模型及葡萄酒质量预测[J]. 邵阳学院学报:自然科学版,2017,14(1):23-29.

[11] 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):29-32,37.

[12] 张秋云,张 营,李 臣. 遗传算法优化 BP 神经网络在中医按摩机器人中的应用[J]. 应用科技,2017,44(2):73-77.

[13] 马满芳,陆惠玲,王媛媛,等. 基于遗传算法:BP 神经网络的乳腺肿瘤辅助诊断模型[J]. 软件导刊,2016,15(11):144-148.

[14] Xie SC, Zhou H, Zhao JJ, *et al*. Energy-absorption forecast of thin-walled structure by GA-BP hybrid algorithm[J]. Journal of Central South University, 2013, 20(4): 1122-1128.

[收稿日期:2018-02-20]

[本文编辑:黄思敏]