

基于GA-BP神经网络的布鲁氏菌病的预测研究

万方芳

云南师范大学数学学院, 云南 昆明

收稿日期: 2024年7月5日; 录用日期: 2024年7月26日; 发布日期: 2024年8月8日

摘要

布鲁氏菌病在我国是一种法定乙类传染病, 近年来布鲁氏菌病患病人数量增多, 对布鲁氏菌病建立预测模型, 对该病的预防和控制有着重要的意义。本文基于2004年~2020年布鲁氏菌病的月发病率数据, 建立BP神经网络、遗传算法优化的BP神经网络模型, 对我国的布鲁氏菌病进行预测, 用MAE、MAPE、MSE、RMSE四个指标比较模型的预测精度。研究结果表明, 遗传算法优化的BP神经网络模型的预测精度优于BP神经网络, 所以遗传算法优化的BP神经网络模型对布鲁氏菌病的预测效果更好。

关键词

布鲁氏菌病, BP神经网络, 遗传算法, GA-BP神经网络

Study on the Prediction of Brucellosis Based on GA-BP Neural Network

Fangfang Wan

School of Mathematics, Yunnan Normal University, Kunming Yunnan

Received: Jul. 5th, 2024; accepted: Jul. 26th, 2024; published: Aug. 8th, 2024

Abstract

Brucellosis in China is a legal group B infectious disease, in recent years, the number of patients with brucellosis has increased, and the establishment of a prediction model for brucellosis has important significance for the prevention and control of the disease. In this paper, based on the monthly incidence data of brucellosis from 2004 to 2020, a model of BP neural network and BP neural network optimized by genetic algorithm was established to predict brucellosis in China. MAE, MAPE, MSE and RMSE were used to compare the prediction accuracy of the model. The results show that the prediction accuracy of BP neural network model optimized by genetic algorithm is better than

that of BP neural network, so the BP neural network model optimized by genetic algorithm has better prediction effect on brucellosis.

Keywords

Brucellosis, BP Neural Network, Genetic Algorithm, GA-BP Neural Network

Copyright © 2024 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

布鲁氏菌病是一种由布鲁氏菌引起的人畜共患病，在我国是乙类传染病。近年来，布鲁氏菌病患者人数增多，成为影响人民群众身体健康的重要卫生问题。因此，研究布鲁氏菌病发病率的预测方法，对预防和控制布鲁氏菌病有着重要的意义。

目前对布鲁氏菌病预测的研究主要集中在时间序列、机器学习方法方面。温福东等人[1]分别建立 SARIMA 模型和 prophet 模型对我国布鲁氏菌病发病人数进行预测，结果表明，prophet 模型对布鲁氏菌病的预测效果更好。张睿等人[2]用随机森林、SARIMA 模型预测我国布鲁氏菌病的发病率，发现两种模型的预测精度都较高，随机森林的预测效果更好一些。田德红等人[3]对中国布病建立 ARIMA、DES 模型、ARIMA-DES 混合模型，发现选取的最优模型为 ARIMA-DES 模型。郑洪淼[4]建立基于历史病例数和气象因素的随机森林预测模型，同时建立季节自回归移动平均模型，根据 R2 和 RMSE 等指标评价布鲁氏菌病的预测效果，发现随机森林模型效果更好。张璐[5]采用灰色 GM(1, 1)模型预测未来 10 年滨州市布鲁氏菌病的发病情况，为有关部门防控布病提供依据。米尔夏提江·阿力木[6]采用 ARIMA 模型、XGBoost 机器学习模型对布鲁氏菌病进行预测，结果发现，XGBoost 模型比 ARIMA 模型预测精度高。张翀[7]建立深度学习长短期记忆网络 LSTM 模型，并将预测结果与传统时间序列 ARIMA 模型进行比较，结果发现，LSTM 模型预测的准确性要优于 ARIMA 模型。袁磊等人[8]采用 GM(1, 1)模型建立我国布鲁氏菌病模型，并利用残差尾段对模型进行修正，建立了模型精度更好的布鲁氏菌病发病率模型。操龙挺[9]采用神经网络、支持向量机、随机森林三种机器学习算法对布病进行预测，结果发现，机器学习方法优于 ARIMA 模型，神经网络模型为最佳模型。黄璐等人[10]采用广义回归神经网络预测布鲁氏菌病月发病数，发现模型拟合较好，可为该病防控提供科学依据。

用机器学习方法对布鲁氏菌病进行研究的文章很多，但是机器学习方法中，用 BP 神经网络、遗传算法优化的 BP 神经网络对布鲁氏菌病研究的文章很少。因此，本文先用 BP 神经网络对布鲁氏菌病发病率进行预测，由于 BP 神经网络极易陷入局部最优解，而遗传算法是一种全局搜索最优解的算法，所以用遗传算法来优化 BP 神经网络，对我国的布鲁氏菌病发病率进行预测。

2. 研究方法

2.1. BP 神经网络

BP 神经网络是一种按照误差反向传播算法的多层前馈型网络，它包括输入层、隐藏层、输出层。每一层有多个神经元，层与层之间的神经元通过权值连接。它主要分为信号的前向传播和误差的后向传播。前向传播是信号先进入输入层，然后传至隐藏层，最后传入输出层，如果输出层的期望值与真实值不相

符，就会进行误差的反向传播。误差后向传播是把误差通过隐藏层传至输入层，在这个过程中，会进行权值与阈值的更新修正。重复进行这两个过程，直到误差低于预先给定的误差或者达到预先给定的学习次数。

Step 1: 信号的前向传播

符号含义见表 1。

Table 1. Symbolic meaning
表 1. 符号含义

符号	含义
x_j	输入层第 j 个神经元的输入信号
ω_{ij}	输入层与隐藏层之间的权值
θ_i	隐藏层第 i 个神经元的阈值
$f_1(x)$	隐藏层的传递函数
ω_{ki}	隐藏层与输出层之间的权值
α_k	输出层第 k 个神经元的阈值
$f_2(x)$	输出层的传递函数
o_k	输出层第 k 个神经元的输出信号

因此，隐藏层第 i 个神经元的输入信号可表示为：

$$ceng_i = \sum_{j=1}^M \omega_{ij} x_j - \theta_i.$$

隐藏层第 i 个神经元的输出信号为：

$$y_i = f_1(ceng_i) = f_1\left(\sum_{j=1}^M \omega_{ij} x_j - \theta_i\right).$$

输出层第 k 个神经元的输入信号为：

$$ceng_k = \sum_{i=1}^q \omega_{ki} y_i - \alpha_k = \sum_{i=1}^q \omega_{ki} f_1\left(\sum_{j=1}^M \omega_{ij} x_j - \theta_i\right) - \alpha_k.$$

输出层第 k 个神经元的输出信号为：

$$o_k = f_2(ceng_k) = f_2\left(\sum_{i=1}^q \omega_{ki} y_i - \alpha_k\right) = f_2\left(\sum_{i=1}^q \omega_{ki} f_1\left(\sum_{j=1}^M \omega_{ij} x_j - \theta_i\right) - \alpha_k\right)$$

这样，神经网络就完成了第一步信号的向前传播。

Step 2: 误差的反向传播

对于单个样本 p ，用 E_p 表示神经网络的期望值 Q_k 与实际值 O_k 之间的误差。 E_p 用平方误差来计算：

$$E_p = \frac{1}{2} \sum_{k=1}^L (Q_k - O_k)^2.$$

因此，对于所有 P 个样本，用 E 表示总体误差，则它的公式为：

$$E = \frac{1}{2} \sum_{p=1}^P \sum_{k=1}^L (Q_k^p - O_k^p)^2.$$

用梯度下降法更新参数，求出问题的最优解。

修正的权值阈值分别为：

$$\Delta\omega_{ki} = -\eta \frac{\partial E}{\partial \omega_{ki}}, \Delta\omega_{ij} = -\eta \frac{\partial E}{\partial \omega_{ij}}, \Delta\alpha_k = -\eta \frac{\partial E}{\partial \alpha_k}, \Delta\theta_i = -\eta \frac{\partial E}{\partial \theta_i},$$

其中， η 是学习率。

Step 3: 重复进行这两个过程，直到误差低于预先给定的误差或者达到预先给定的学习次数，训练就完成了。

2.2. GA-BP

BP 神经网络因为其内部采用的是局部搜索算法和梯度下降法，极易陷入局部最优解，而遗传算法是一种全局搜索最优解的算法，所以可以用遗传算法来优化 BP 神经网络。

遗传算法的流程一般包括染色体的编码、初始群体的生成、适应度函数的设计、选择、交叉、变异操作、终止条件。

2.2.1. 染色体的编码方式

遗传算法是把参数空间中的解通过一定的方式编码成字符串进行运算的，一个解对应一条染色体，这些编码后的染色体可以进行基因的交叉、变异操作。染色体的编码方式有很多种，例如二进制编码、符号编码、浮点数编码等，其中最常用的是二进制编码。

2.2.2. 初始群体的生成

染色体编码后会形成一些字符串，一个字符串就对应一个个体，所有的个体构成一个种群。一般是通过电脑编码生成随机数，因为个体是随机生成的，所以它们适应环境的能力不同，有好有坏，也就是适应度有高有低。

2.2.3. 适应度函数的设计

随机生成的个体适应度有高有低，适应度低的个体会被淘汰。所以，我们需要一个判断标准，来判断个体的优劣。这就是适应度函数。本文中用均方误差作为适应度的评估值，可以表示为：

$$F = MSE(Q - O) = \frac{1}{n} \sum_{i=1}^n (q_i - o_i)^2.$$

其中， q_i 和 o_i 分别为第 i 天的期望值与实际值。

2.2.4. 选择

选择是指从种群中选择出适应环境能力强的个体，让它们产生下一代种群，适应环境能力差的个体直接淘汰。通常适应度高的个体，更容易生存下来，适应度低的个体，更容易被淘汰。一般采用轮盘赌选择法。公式为：

$$p_i = \frac{f_i}{\sum_{i=1}^n f_i}$$

其中， f_i 为第 i 个个体适应度值， p_i 为个体被选择的概率， n 为种群规模。

2.2.5. 交叉

交叉对应遗传学中的基因重组，是指把通过交叉概率选出来的个体，互相交换自身染色体上基因的

一部分，产生新的两条染色体，即新的个体。基因不断发生重组，产生新的个体，有可能使优良的基因结合在一起，使种群向更适应环境的方向进化。交叉概率的取值一般为 0.9~1。

2.2.6. 变异

变异对应遗传学中的基因突变，是指把通过变异概率选出来的个体，改变其染色体上的某些基因。变异发生的概率很小，取值一般为 0.001~0.1。

2.2.7. 终止条件

当误差低于预先给定的误差或者达到预先给定的遗传代数时，训练就完成了。

遗传算法是通过调节 BP 神经网络的初始权重和阈值，来优化 BP 神经网络的。GA-BP 算法的思想是：

Step 1: 确定 BP 神经网络的结构并初始化参数，然后对参数进行编码操作，也就是形成种群中的个体。

Step 2: 进入遗传算法的流程，当达到遗传算法预先设定的目标时，就得到了 BP 神经网络参数的最优解。

Step 3: 对具有最优参数的 BP 神经网络进行训练，就得到用遗传算法优化的 BP 神经网络。

GA-BP 算法流程图如下图 1 所示。

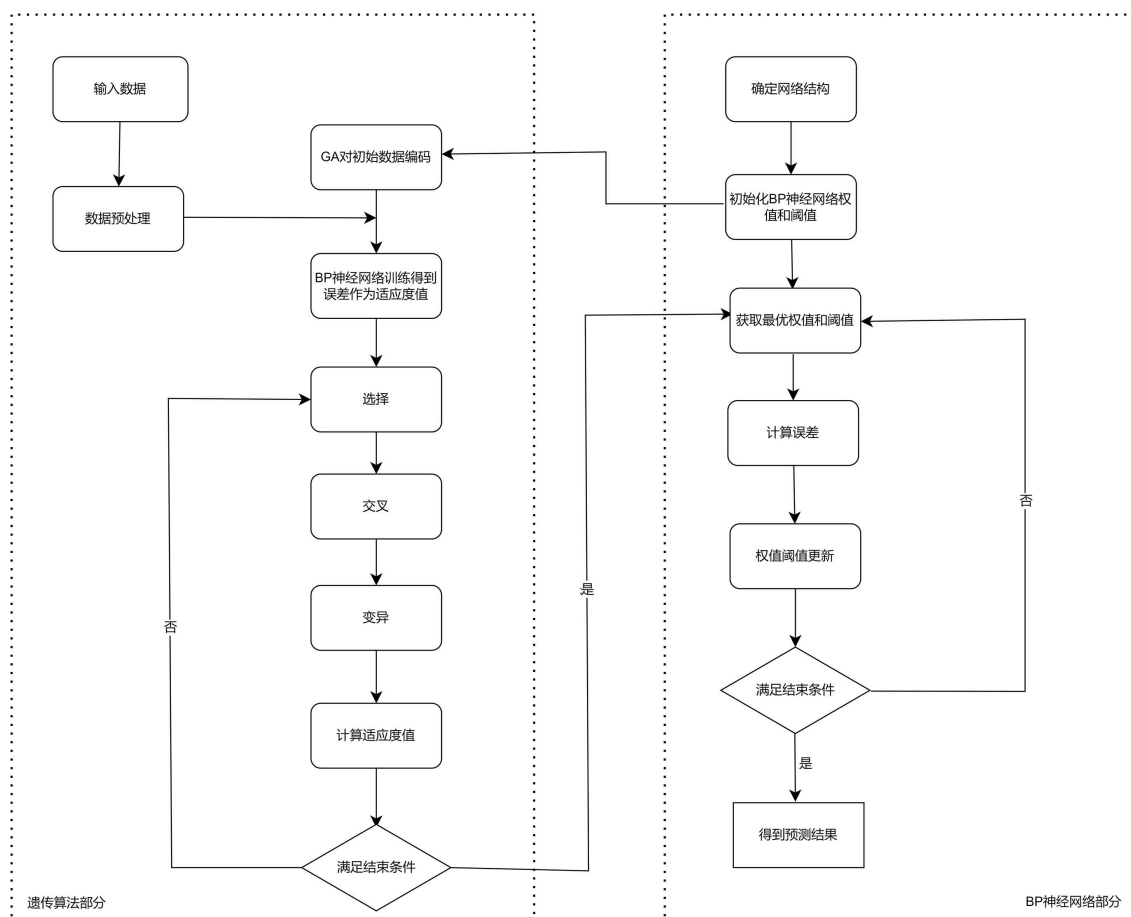


Figure 1. GA-BP algorithm flow chart

图 1. GA-BP 算法流程图

3. 评价模型预测精度的指标

用平均绝对误差(MAE)、平均绝对百分误差(MAPE)、均方误差(MSE)、均方根误差(RMSE)来评估模型的精度。

1) 平均绝对误差(MAE)是指真实值与预测值之间的偏差绝对值的算术平均值。公式为:

$$MAE = \frac{1}{n} \sum_{t=1}^n |y_t - \hat{y}_t|,$$

其中, y_t 是真实值, $\hat{y}_t$ 是预测值, n 是预测次数。

2) 平均绝对百分误差(MAPE)是指真实值与预测值之间的偏差与真实值比值的绝对值的算术平均值。公式为:

$$MAPE = \frac{1}{n} \sum_{t=1}^n \left| \frac{y_t - \hat{y}_t}{y_t} \right|,$$

其中, y_t 是真实值, $\hat{y}_t$ 是预测值, n 是预测次数。

3) 均方误差(MSE)是指真实值与预测值之间的偏差的平方和除以预测次数。公式为:

$$MSE = \frac{\sum_{t=1}^n (y_t - \hat{y}_t)^2}{n},$$

其中, y_t 是真实值, $\hat{y}_t$ 是预测值, n 是预测次数。

4) 均方根误差(RMSE)是指真实值与预测值之间的偏差的平方和除以预测次数的平方根。公式为:

$$RMSE = \sqrt{\frac{\sum_{t=1}^n (y_t - \hat{y}_t)^2}{n}},$$

其中, y_t 是真实值, $\hat{y}_t$ 是预测值, n 是预测次数。

4. 实证分析

4.1. BP 神经网络预测模型

下面构建布鲁氏菌病月发病率的 BP 神经网络预测模型, 将 2004~2019 年的数据作为训练集, 2020 年的数据作为测试集。考虑到布鲁氏菌病月发病率的周期性和自相关性的特点, 用 2004 年 1 月、2004 年 2 月、……、2004 年 11 月、2004 年 12 月的数据来预测 2005 年 1 月的数据。训练集和测试集的具体划分方式如下表 2 和表 3 所示。

Table 2. Training set partitioning
表 2. 训练集划分

模型输入				模型输出
2004.01	2004.02		2004.12	2005.01
2004.02	2004.03		2005.01	2005.02
.....				
2018.11	2018.12		2019.10	2019.11
2018.12	2019.01		2019.11	2019.12

Table 3. Test set partitioning

表 3. 测试集划分

模型输入			模型输出	
2019.01	2019.02		2019.12	2020.01
2019.02	2019.03		2020.01	2020.02
.....				
2019.11	2019.12		2020.10	2020.11
2019.12	2020.01		2020.11	2020.12

输入变量为过去一年布鲁氏菌病月发病率数据，输出变量为下一个月的月发病率数据。所以，BP 神经网络的输入层的神经元节点数为 12，输出层的神经元节点数为 1。隐藏层神经元的节点数一般用公式 $n_1 = \sqrt{n+m} + a$ 来计算，其中， n_1 表示隐藏层节点数， n 表示输入层节点数， m 表示输出层节点数， a 是常数， $1 \leq a \leq 10$ 。所以本文隐藏层神经元节点数的初步确定为[4, 13]。不同隐藏层节点数的网络训练结果见下表 4。

Table 4. Network training results with different number of hidden layer nodes

表 4. 不同隐藏层节点数的网络训练结果

隐藏层节点数	均方误差
4	0.0015859
5	0.00076066
6	0.0007444
7	0.00094901
8	0.00075339
9	0.0014375
10	0.00087411
11	0.00082324
12	0.00084835
13	0.00093013

由上表可以看出，最佳隐藏层节点数为 6，相应的均方误差为 0.0007444。

确定最佳隐藏层节点后，可以构建最佳隐藏层节点数的 BP 神经网络，对 2004 年~2019 年的数据进行训练，具体的参数设置如下表 5 所示。

Table 5. Network parameter value

表 5. 网络参数取值

参数	取值
神经元节点数目	12、13、1
隐含层传递函数	Tan-sigmoid 型
输出层传递函数	Purelin 型

续表

训练函数	trainlm
最大迭代次数	1000
目标误差	10^{-6}
学习效率	0.01

BP 神经网络训练得到的布鲁氏菌病预测值与真实值的对比图见下图 2。

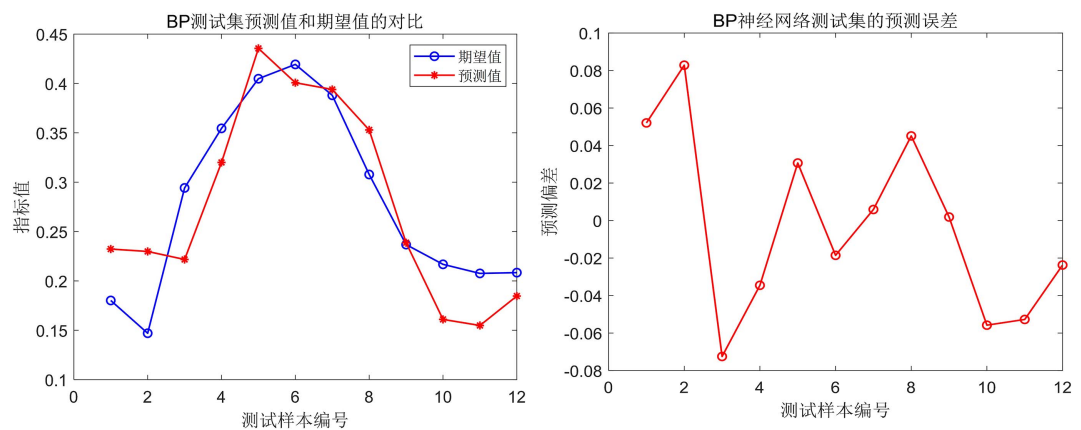


Figure 2. Comparison of true and predicted monthly incidence of brucellosis and error graph
图 2. 布鲁氏菌病月发病率真实值与预测值对比及误差图

由上图可以看出，测试集的真实值和预测值非常接近，且误差在-0.08 和 0.09 之间，说明 BP 神经网络模型的预测效果很好。

4.2. GA-BP 神经网络预测模型

BP 神经网络因为其内部采用的是局部搜索算法和梯度下降法，极易陷入局部最优解，而遗传算法是一种全局搜索最优解的算法，所以可以用遗传算法来优化 BP 神经网络。仍然将 2004~2019 年的数据作为训练集，2020 年的数据作为测试集，最大迭代次数、目标误差、学习效率也和 BP 神经网络设置的一样，分别为 1000、 10^{-6} 、0.01。GA-BP 的其它参数设置如下表 6 所示。

Table 6. GA-BP neural network parameter value
表 6. GA-BP 神经网络参数取值

参数	取值
种群规模	20
遗传代数	20
交叉概率	0.8
变异概率	0.1

GA-BP 神经网络的回归拟合图见下图 3。

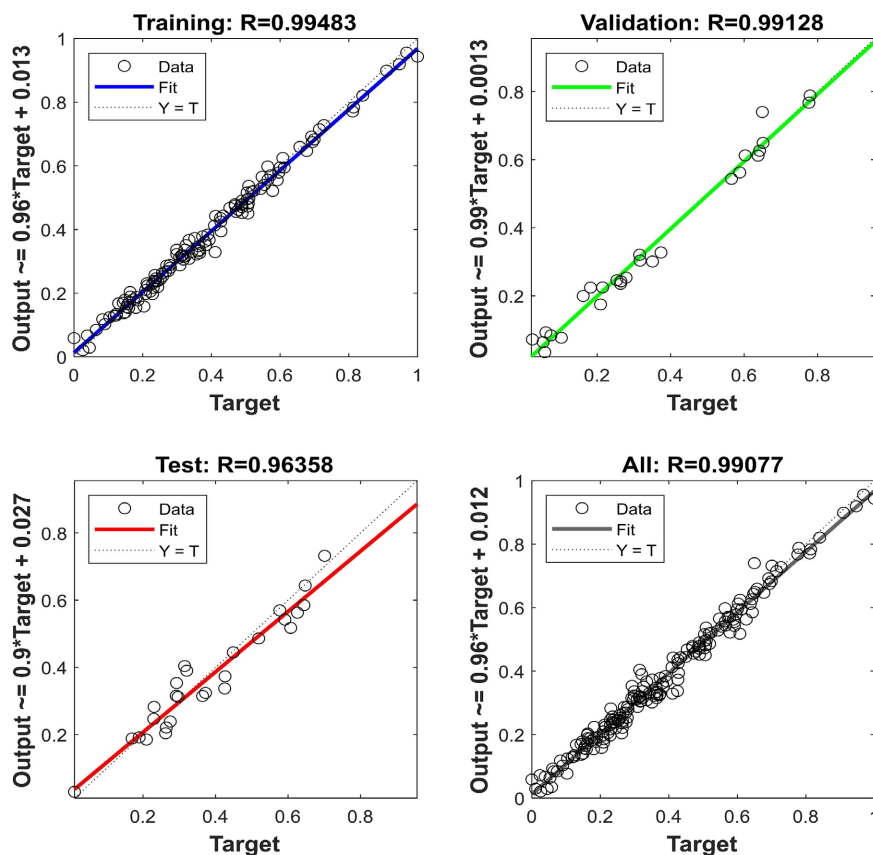


Figure 3. Regression fitting diagram of GA-BP neural network

图 3. GA-BP 神经网络的回归拟合图

由上图可以看出, GA-BP 神经网络的训练集的 $R=0.99483$, 也就是布鲁氏菌病发病率真实值与预测值的拟合度为 0.99483。真实值曲线与拟合曲线可以看到是重合的, 说明拟合效果非常好。网络的验证集的 $R=0.99128$, 测试集的 $R=0.96358$, 所有样本集的 $R=0.99077$ 。

下面是 GA-BP 训练集和测试集预测结果对比图(图 4)。

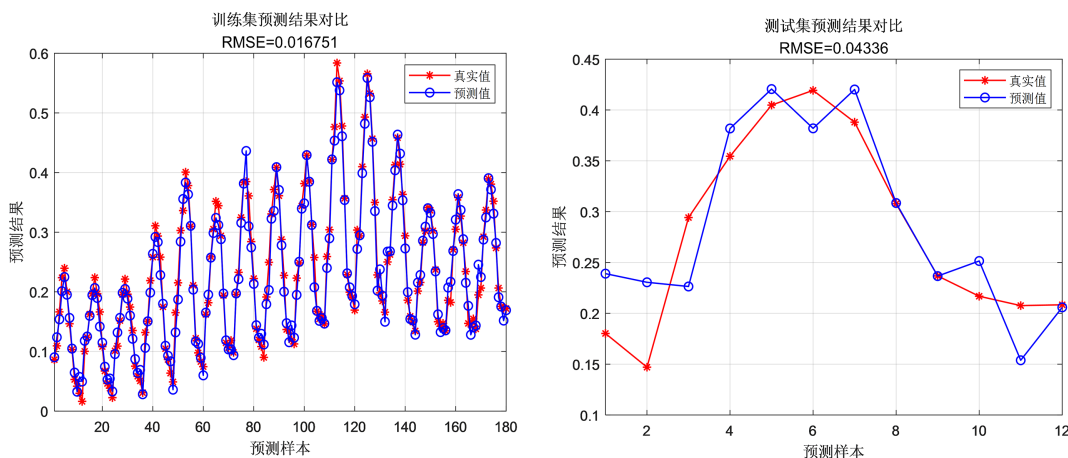


Figure 4. Comparison of GA-BP training set and test set prediction results

图 4. GA-BP 训练集和测试集预测结果对比图

由上图可以看出，测试集和训练集的真实值和预测值都非常接近，并且训练集的均方误差 RMSE 为 0.016751，测试集的均方误差 RMSE 为 0.04336，误差都比较小，说明 GA-BP 神经网络模型的预测效果很好。

4.3. 模型预测效果

用 MAE、MSE、RMSE、MAPE 四个指标来评价模型的预测效果，BP、GA-BP 的预测效果见下表 7。

Table 7. Comparison of prediction effect of each model
表 7. 各模型预测效果对比

模型	MAE	MSE	RMSE	MAPE
BP	0.0397	0.0022	0.0464	17.5937%
GA-BP	0.0346	0.0019	0.0434	1.1373%

由上表可以看出，GA-BP 的 MAE、MSE、RMSE、MAPE 值都小于 BP 神经网络的值，所以 GA-BP 模型的预测效果更好，可以用 GA-BP 神经网络模型对我国的布鲁氏菌病发病率进行预测。

5. 结论

基于 2004 年~2020 年我国布鲁氏菌病月发病率数据，将 2004~2019 年的数据作为训练集，2020 年的数据作为测试集。先建立 BP 神经网络模型，对我国布鲁氏菌病月发病率进行预测。由于 BP 神经网络极易陷入局部最优解，而遗传算法是一种全局搜索最优解的算法，所以用遗传算法来优化 BP 神经网络，对我国的布鲁氏菌病发病率进行预测。用遗传算法改进的 BP 神经网络的 MAE、MSE、RMSE、MAPE 都小于 BP 神经网络的 MAE、MSE、RMSE、MAPE，说明用遗传算法改进的 BP 神经网络的预测效果比 BP 神经网络的预测效果好，可以用遗传算法改进的 BP 神经网络模型对我国的布鲁氏菌病进行预测，从而为布鲁氏菌病的预防和控制提供科学依据。

参考文献

[1] 温福东, 赵彬宇, 苏月, 等. 基于 Prophet 模型预测中国布鲁氏菌病发病人数[J]. 中国医院统计, 2024, 31(1): 7-10.

[2] 张睿, 王晓风, 张业武, 等. 随机森林和 SARIMA 模型预测我国布鲁氏菌病发病率效果研究[J]. 公共卫生与预防医学, 2022, 33(1): 1-5.

[3] 田德红, 于国伟, 丁国武, 等. ARIMA-DES 混合模型在中国布鲁菌病分析和预测中的应用[J]. 中国卫生统计, 2016, 33(2): 245-248.

[4] 郑洪淼. 甘肃省 2015-2019 年布鲁氏菌病时空聚集性及与气象因素的关系及预测[D]: [硕士学位论文]. 兰州: 兰州大学, 2024.

[5] 张璐. 滨州市 2005-2015 年人间布鲁氏菌病流行特征及预测分析[D]: [硕士学位论文]. 济南: 山东大学, 2017.

[6] 米尔夏提江·阿力木. 使用 XGBoost 机器学习方法预测辽宁省人间布鲁氏菌病发病的研究[D]: [硕士学位论文]. 沈阳: 中国医科大学, 2022.

[7] 张翀. 长短期记忆网络深度学习模型对人间布鲁氏菌病预测研究[D]: [硕士学位论文]. 沈阳: 中国医科大学, 2022.

[8] 袁磊, 石锦浩, 刘丽娟, 等. GM(1, 1)残差修正模型在布鲁氏菌病发病率预测中的应用[J]. 医学动物防制, 2020, 36(5): 409-413.

[9] 操龙挺. 基于机器学习方法的布鲁氏菌病预测研究[D]: [硕士学位论文]. 合肥: 安徽医科大学, 2021.

[10] 黄璐, 孙娜, 许小珊, 等. 广义回归神经网络在布鲁氏菌病预测中的应用[J]. 郑州大学学报(医学版), 2018, 53(6): 751-754.