Transactions on Materials, Biotechnology and Life Sciences
ISSN: 2960-1932, eISSN: 2960-2319 | Volume 5
MMTCS 2024

A Study on Prediction and Assessment of Diabetes Mellitus Based on BP Neural Network and Decision Tree Model

Yuyuan Pan *

School of Mathematics and Physics, Lanzhou Jiaotong University, Lanzhou, China

* Corresponding Author Email: 2817814388@qq.com

Abstract. Diabetes mellitus is a metabolic disease characterised by the patient's blood glucose being chronically higher than the standard value. In this paper, we first used multiple feature screening technique to gradually screen 17 main indicators from 34 test indicators, subsequently, we used BP neural network model for glucose value prediction and trained the model by back propagation algorithm, and then we used decision tree model combined with the new test data to assess the risk of diabetes, by constructing a tree structure to classify and predict the presence or absence of the feature based on the risk of diabetes.

Keywords: Correlation analysis; Lasso Regression; BP neural network model; Decision tree model.

1. Introduction

Diabetes mellitus is a common metabolic disease, which leads to many serious complications and greatly affects the health and quality of life of patients. Mathematical models can help people better understand the factors and mechanisms associated with diabetes, predict the development trend of the disease, assess the effectiveness of intervention and treatment, and improve the prevention and treatment of diabetes [1-4].

In this paper, firstly, the main variables were gradually screened out from 34 testing indicators and were used in order, including Correlation Analysis (CA), Gradient Boosting method (GB), and Lasso Regression (LR). Subsequently, a Back Propagation Neural Network (BPNN) model was used for glucose value prediction, and the model was trained by a back-propagation algorithm. Finally, factor analysis was used to select 9 indicators from 17 indicators into the decision tree algorithm, and the results of the features and classification methods learned based on the training data, as well as the results of selecting and evaluating the features using indicators such as information gain, were used as the basis for judging the risk of diabetes.

2. Analysis of Impact Factors

2.1. Correlation Analysis

Correlation analysis is a statistical method for assessing the strength of a linear relationship between two variables. It reflects the degree of linear correlation between two variables by calculating the correlation coefficient.

The steps are as follows:

(1) Define the variables: Let and be two random variables, each with one observation.

(2) Calculate the mean: First calculate the means of ad, which are denoted as and respectively.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \tag{1}$$

(3) Calculate the covariance: Next, calculate the covariance of and ($Cov(X,Y)$), which measures the degree of co-variation of the two variables.

$$Cov(X,Y)=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y}) \tag{2}$$

(4) Calculate the standard deviation: Then, calculate the standard deviation of and, denoted as σ_X and σ_Y respectively.

(5) Calculate the Pearson correlation coefficient: Finally, the covariance and standard deviation calculated above are used to calculate the Pearson correlation coefficient (r).

$$r=\frac{Cov(X,Y)}{\sigma_X\sigma_Y}=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \tag{3}$$

By calculating the correlation coefficients (Pearson coefficients) between each variable and the blood glucose values, a preliminary list of influencing factors can be obtained in this paper.

The sign of the correlation coefficient is considered by ranking the correlation coefficients according to the magnitude of their absolute values. Positive correlation indicates that the blood glucose value increases when the indicator value increases, and negative correlation indicates that the blood glucose value decreases when the indicator value increases. From these, variables with significant positive or negative correlation with blood glucose values were screened. The main variables with the most plausibility and relevance were further filtered by considering the research background and domain knowledge.

2.2. Lasso Regression

Lasso Regression is a feature selection technique that promotes sparsity of the coefficients by introducing an L1 regularisation term in a linear regression model to enable automatic feature selection [5]. In Lasso regression, the regularisation term is the sum of the absolute values of the coefficients, which causes some coefficients to be compressed to zero, thus removing unimportant features [6].

The objective function of Lasso Regression can be expressed as:

$$\min_{\beta}\frac{1}{2n}\|y-X\beta\|_2^2+\alpha\|\beta\|_1 \tag{4}$$

Where: y is the vector of the dependent variable (target variable); X is the matrix of independent variables (features); β is the vector of coefficients; $\|\cdot\|_1$ is the L1 paradigm number; $\|\cdot\|_2$ is the L2 parameter;$\propto$ is a regularisation parameter to control sparsity. Larger values result in more coefficients being compressed to zero, thus achieving feature selection.

The optimisation objective consists of two parts: the data fitting term (i.e., the sum of squares of the errors) and the regularisation term (the sum of the absolute values of the coefficients). By adjusting the value of the term, a balance between model complexity and fitting accuracy can be achieved.

The results of the above steps can be used as a reference for the initial screening of the main variables in this paper. Finally, this paper will select those variables that show importance in several steps as the main variables.

Based on the results of the correlation analysis and domain knowledge, the main variable indicators initially selected included: total cholesterol, triglycerides, *r-glutamyltransferase, *alanine aminotransferase, LDL cholesterol, urea, uric acid, age, *aspartate aminotransferase, *alkaline phosphatase, erythrocyte pressure volume, mean hemoglobin concentration of erythrocytes, erythrocyte count, creatinine, hemoglobin, gender, and high-density lipoprotein cholesterol. Based on the extent to which a range of variables were observed to be associated with blood glucose levels. The R^2 value of the model was 0.868, implying that these variables collectively explained about 86.8% of the variation in blood glucose levels, showing a strong predictive power. Blood glucose levels are influenced by a variety of factors, including gender, lipid levels, age, and a number of hematological and biochemical markers. These findings provide important reference information for blood glucose management and diabetes prevention, suggesting that multiple factors need to be considered when developing intervention strategies.

3. Glucose Forecasting

3.1. BP Neural Network Model

BP network is a multi-layer feed-forward neural network consisting of input, hidden, and output layers. Layers are fully interconnected with no interconnections between layers and there can be one or more hidden layers. Constructing a BP neural network requires determining the characteristics of its processing units, the neurons, and the topology of the network. Neuron (for input, for weight, for threshold, for output) [7]. The most basic processing unit of a neural network Neurons in the hidden layer use S-type transformation functions Neurons in the output layer can use S-type or linear transformation functions.

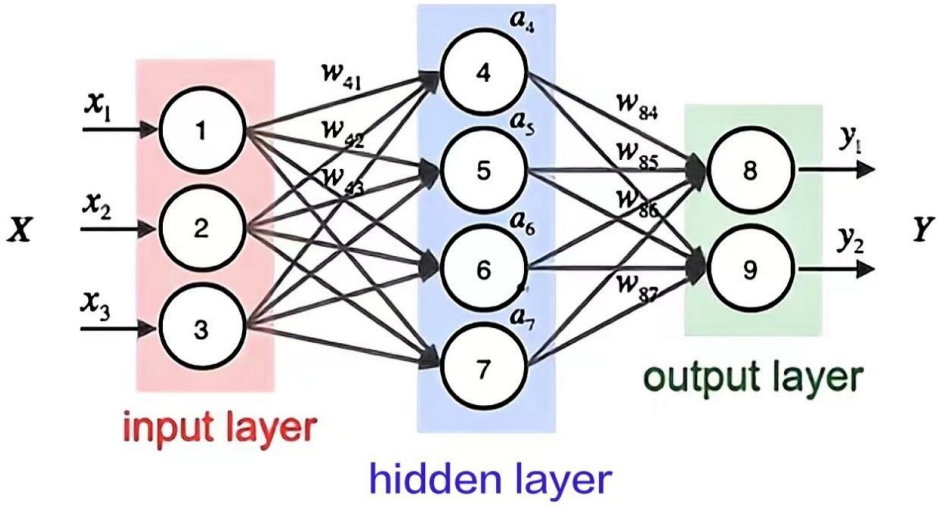

Fig. 1 Feed-forward neural network model

The input and output values of each neuron in the hidden and output layers are calculated as shown in the following equation.

$$z_j^l = \sum_i w_{jk}^l a_i^{l-1} + b_j^l \tag{5}$$

$$a_j^l = \sigma\left(\sum_i w_{jk}^l a_i^{l-1} + b_j^l\right) \tag{6}$$

Where, w_{jk}^l: denotes the connection weight of the ith neuron in layer $(l-1)$to the jth neuron in layer l; b_j^l: denotes the bias top of the jth neuron in layer l; z_j^l: denotes the input value of the jth neuron in layer l; a_j^l: denotes the output value of the jth neuron in layer l; σ: denotes the activation function.

3.1.1. Activation Function

Commonly used activation functions are the sigmoid function, Tanh function, Relu function. Their graphs are as follows:

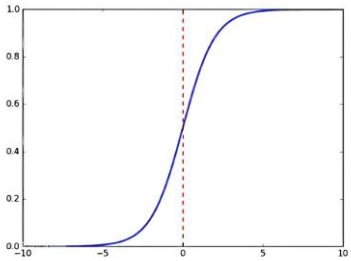

Fig. 2 Sigmoid function

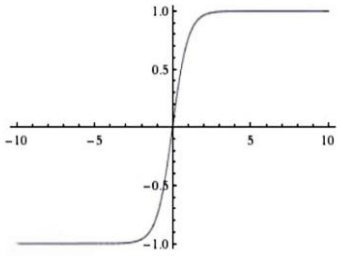

Fig. 3 Tanh function

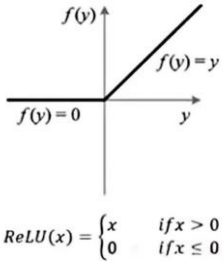

Fig. 4 Relu function

3.1.2. Loss Function

The loss function loss is a function of the network table y_{out} with respect to the true result y, with very small values. Then this paper will know that if the loss of y between a network's computed result y_{out} and the true result is always very small, then it can be shown that the network is very close to the true relationship. So the purpose of this paper, is to constantly adjust the weights (that is, the parameters of the network) to make the network calculated results y_{out} as close as possible to the real results y, which is equivalent to the loss function is as small as possible, here the use of gradient descent method to find the minimum value.

3.1.3. Forward and Backward Propagation

The process of getting y_k is also a forward propagation, whereas in a complete BP neural network, a prediction is obtained by randomly configuring the hyperparameters of the network. This is a forward propagation process. And then the gap between the predicted value and the real value is calculated, and the parameters are adjusted accordingly to this gap, which is a backward propagation process. As shown in the figure below, the blue arrow is the process of forward propagation and the yellow line is the process of back propagation.

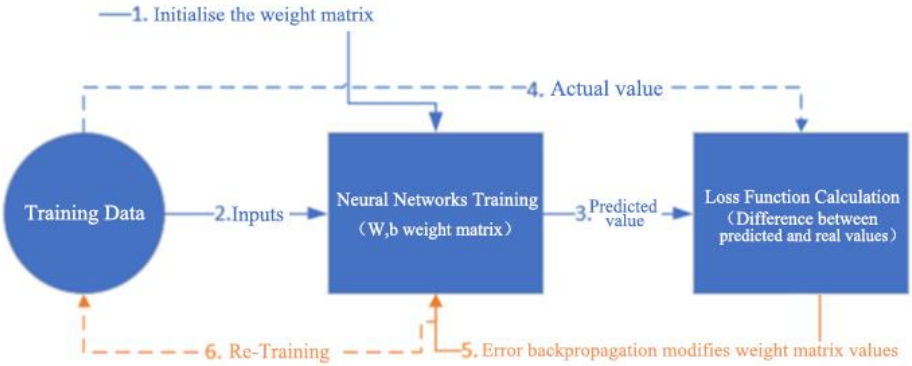

Fig. 5 Schematic diagram of BP neural network

Where the weight is w, the predicted value is o, the true value is t, and is the learning rate (which can be interpreted as the step size of the gradient descent method):

Error function:

$$E_d = \frac{1}{2} \sum_{k \in outputs} (t_k - o_k)^2 \tag{7}$$

Gradient descent:

$$\nabla w_{ji} = -\eta \frac{\partial E_d}{\partial w_{ji}} \tag{8}$$

Chain rule:

$$\frac{\partial E_d}{\partial w_{ji}} = \frac{\partial E_d}{\partial y_j} \cdot \frac{\partial y_j}{\partial w_{ji}} = \frac{\partial E_d}{\partial y_{ji}} \cdot x_{ji} \tag{9}$$

Bringing into the gradient descent formula can be obtained:

$$\nabla w_{ji} = -\eta \frac{\partial E_d}{\partial w_{ji}} = \eta (t_j - o_j) o_j (1 - o_j) x_{ji} \tag{10}$$

Adjusting the weights w accordingly to the obtained results completes a backpropagation. Then start the next iteration, and the cycle repeats until the convergence condition is reached and the cycle is jumped out. Neural network learning using the improved BP algorithm learning process consists of a forward calculation process and an error backpropagation process. In the forward computation process, the input information is computed layer by layer from the input layer through the hidden layer and transmitted to the output layer, and the state of neurons in each layer only affects the state of neurons in the next layer. If the output layer fails to obtain the desired output, it is transferred to the error backpropagation process where the error signal is returned along the original connecting pathway by modifying the weights of the neurons in each layer to minimize the error of the network system. Finally the actual output of the network is approximated to the respective desired output.

Based on the provided metrics of the predicted results of the BP neural network, it is possible to derive:

Table 1. Results of model evaluation

	MSE	RMSE	MAE	MAPE	R^2
Training set	1.775	1.332	0.759	1.87	0.88
Test set	2.351	1.533	0.777	1.76	0.89

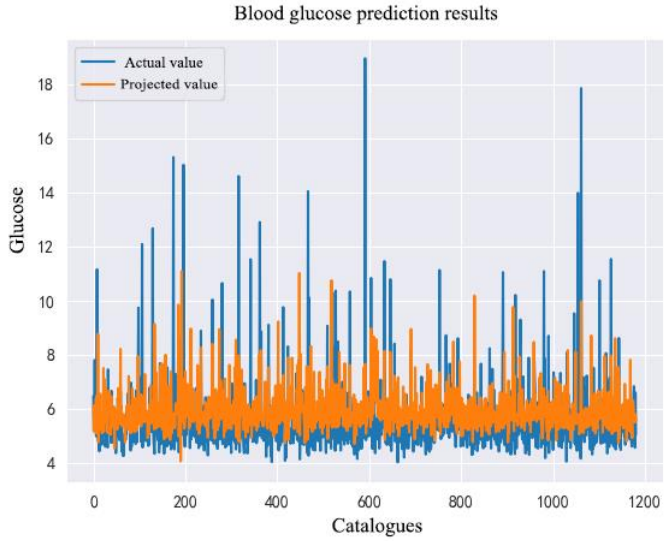

Fig. 6 Model prediction results

MSE (Mean Square Error) and RMSE (Root Mean Square Error) are commonly used measures of prediction error, with smaller values indicating lower prediction error of the model. The RMSE is relatively low on both the training and test sets, indicating that the BP neural network model has a certain accuracy in predicting blood glucose values.

MAE (Mean Absolute Error) and MAPE (Mean Absolute Percentage Error) are used to measure the average degree of difference between the predicted and actual values. Lower values of MAE and MAPE indicate that the BP neural network model is relatively accurate with less average prediction error of blood glucose values on both training and test sets.

R^2 (coefficient of determination) is used to measure the ability of the model to explain the dependent variable (blood glucose value) and takes a value between 0 and 1. The closer to 1 means that the model explains the dependent variable better. The high R^2on both the training and test sets indicate that the BP neural network model has a strong predictive ability for the blood glucose value and is able to explain most of the variance in the target variable.

4. Diabetes Risk Assessment

4.1. Model Building

As the blood glucose value in the human body fluctuates strongly with the food structure of the diet, which leads to the instability of the blood glucose value, it is easy to have a large bias in the judgment process resulting in the consequences of misjudgment [8]. In order to further improve the accuracy of the judgment, this section will create an auxiliary model for assessing the risk of diabetes, which requires the use of the decision tree model in machine learning [9].

The steps are as follows:

(1) The 17 data screened using multiple feature selection are first based on the hemoglobin status of 120-160g/L as normal, divided into 1 (normal), 0 (abnormal) a new column.

(2) In order to further examine whether the 17 indicators screened and blood glucose values are closely related, these data were subjected to factor analysis, and the results were ['*r-glutamyltransferase', '*alanine aminotransferase', '*aspartate aminotransferase', '*alkaline phosphatase', 'LDL cholesterol', 'urea', 'uric acid', 'age', 'gender'] for these nine indicators.

(3) Then build a decision tree model.

4.2. Assessment and Analysis

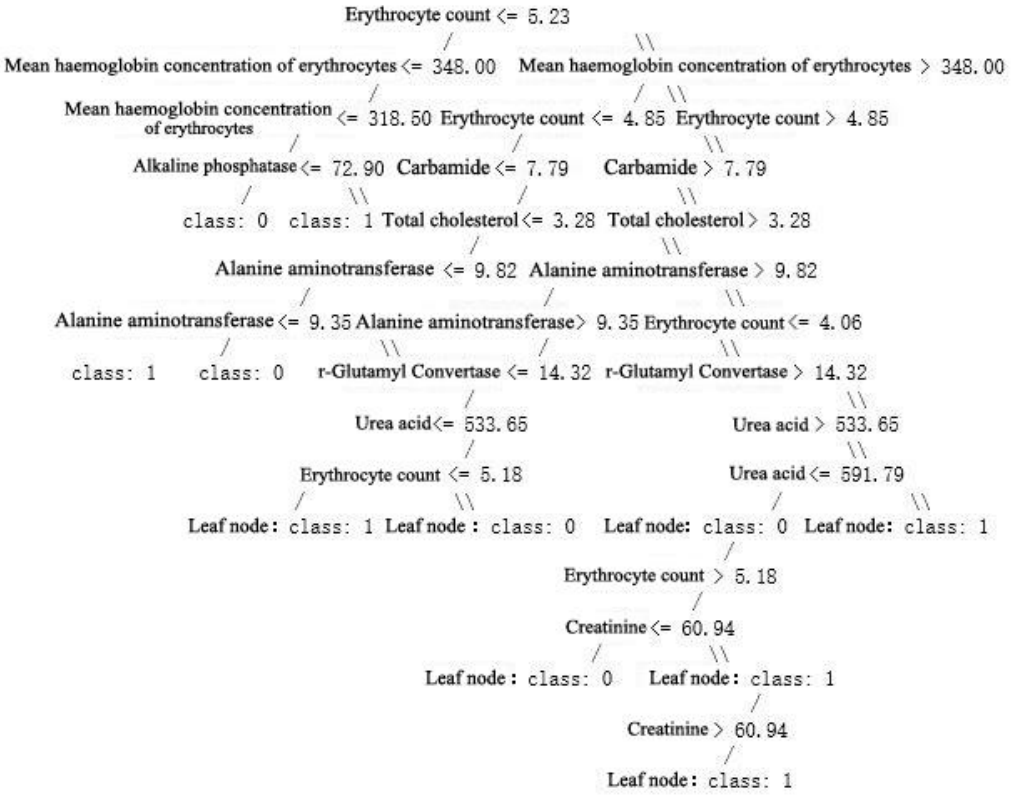

Fig. 7 Decision tree

As in Figure 7, red blood cell count $\leq$ 5.23, mean red blood cell hemoglobin concentration $\leq$ 348.00, mean red blood cell hemoglobin concentration $\leq$ 18.50, *alkaline phosphatase > 72.90 then class:1 (normal), mean red blood cell hemoglobin concentration > 318.5, urea $\leq$ 7.79, total cholesterol > 3.28, *alanine aminotransferase $\leq$9.82, *alanine aminotransferase $\leq$ 9.35 then class:1 (normal), *alanine aminotransferase > 9.82

Erythrocyte count $\leq$ 4.06,*r-glutamyl converting enzyme $\leq$ 14.32 then class:0 (abnormal), Erythrocyte count >4.06, Uric acid $\leq$ 533.65, Erythrocyte count $\leq$ 5.18 then class:1 (normal), Erythrocyte count >5.18, Creatinine >60.94 then class:1 (normal), Creatinine >60.94 then class:1 (normal).It was concluded that there is no risk of diabetes mellitus.

5. Summary

In this study, this paper firstly used various feature screening techniques such as correlation analysis, gradient boosting method and Lasso Regression to accurately screen the main variable indicators from 34 testing indicators. The effective application of these methods not only helps this paper identify the features with the strongest correlation with the target variable (blood glucose value) but also further removes irrelevant features, thus avoiding the overfitting of the model and improving the accuracy of the prediction. The BP neural network is trained by the backpropagation algorithm, whose training process includes two steps of forward propagation and backpropagation, and adjusts the weights by continuously optimizing the loss function, thus achieving the accuracy of the prediction of blood glucose value. Thus, the accurate prediction of blood glucose value was achieved. In this paper, several indicators are used as the basis for judgment using the decision tree model, and one column is considered the target column (1: no risk, 0: diabetes risk), which means that the individual is considered to be at risk of diabetes. This method of assessment is accurate and effective and provides a powerful tool for clinicians and patients.

References

[1] Wang Yiwen, Li Jianjun, Qu Zepeng. Research on short-term blood glucose prediction method based on Transformer[J]. Journal of China University of Weights and Measures,2023,34(03):372-378.

[2] Xiao Zhengbang. Research on the assessment method of glycolipid metabolism based on respiratory quotient detection[D]. Supervisor: Jianming Zhu. Guilin University of Electronic Science and Technology, 2023.

[3] Chen Yanzhang. Research on non-invasive blood glucose detection based on electrochemical detection technology [D]. Supervisor: Chuanpei Xu. Guilin University of Electronic Science and Technology, 2023.

[4] Cao Ke, Tan Chong, Liu Hong, Zheng Min. Data fusion algorithm for wireless sensor networks based on improved grey wolf algorithm optimized BP neural network[J]. Journal of Chinese Academy of Sciences University,2022,39(02):232-239.

[5] Cai Yuguo, Zheng Yongli, He Min, Pu Yu. Construction and validation of an early diagnostic model for knee tuberculosis based on LASSO regression[J]. Chinese Journal of Anti-Tuberculosis,2023,45(03):297-304.

[6] Zhu Hailong, Li Pingping. Analysis of factors affecting fiscal revenue in Anhui Province based on ridge regression and LASSO regression[J] Journal of Jiangxi University of Science and Technology,2022,43(01):59-65.

[7] Xiaochen Zhang, Xuejun Chen, Yu Song, Peixin Ma. Particle swarm algorithm to optimize BP neural network in slope stability[J]. Mining Research and Development,2022,42(01): 71-76.DOI:10.1382k.2022.01.024.

[8] Zhou Jiachen. Research on personalized short-term blood glucose prediction model based on deep learning [D]. Shanghai Normal University, 2023.

[9] Fan Bin. Research on diabetes risk prediction model based on convolutional neural network [D]. Nanjing University of Posts and Telecommunications, 2022.