Research on diabetes prediction method based on electronic medical record data analysis

Chen Hui[1], Wang Mingyuan[2], Tang Dingjun[3, *], Zhang Longwei[4], Guo Ziyan[5] and Zhao Jun[6]

[1]Zhejiang University, Hangzhou,310058, China
[2]Sichuan University, Chengdu, 610065, China
[3]Dongguan University of Technology, Dongguan, 523808, China
[4]Shandong University of Traditional Chinese Medicine, Jinan, 250355, China
[5]Shandong University of Traditional Chinese Medicine, Jinan, 250355, China
[6]Shandong University of Traditional Chinese Medicine, Jinan, 250355, China

Abstract: The continuous progress of computer science and technology has accelerated the pace of informatization construction of the medical system. Medical technology has developed rapidly in various research directions, and the construction of medical IT systems has been continuously improved. The popular application of electronic medical records has produced massive medical data in the medical process. At the same time, in medical behavior, more and more rely on data to make relevant judgments. The coverage of medical equipment is becoming more and more extensive, and the accuracy of data is constantly improving, and the clinical diagnosis is gradually shifting from qualitative judgment to quantitative analysis. Based on the analysis of electronic medical record data, this article studies and analyzes the risk factors leading to diabetes. By analyzing the characteristic variables, the risk factors significantly related to diabetes are obtained as the input variables of the BP neural network model. For complex problems, machine learning algorithms have higher accuracy and stronger generalization capabilities. Based on the BP artificial neural network model, this paper builds and builds a machine learning simulation to predict diabetes.

1 Introduction

Diabetes Mellitus (DM) is a common chronic disease. Hyperglycemia is its main feature, and the family genetic characteristics of DM are obvious. Pathological analysis results show that the pathological causes of diabetes include two types: one is that when the pancreas cannot produce sufficient insulin, it will cause type 1 diabetes (T1D); when the body cannot effectively use the insulin produced, it will be triggered Type 2 diabetes (T2D). Normally, T1D is called primary diabetes, and its pathogenesis is due to the lack of insulin in the body when insulin-secreting β cells in the pancreas are damaged, resulting in the blood sugar level cannot be reduced in time [1]. T2D is called non-insulin-dependent diabetes, mainly due to insulin resistance or defects in insulin secretion, etc., does not effectively use the insulin in the body, resulting in high blood sugar. Factors such as lifestyle, physical activity, eating habits and genetics are the main causes of type 2 diabetes [2].

Diabetes can cause many problems for patients. If there is no reasonable treatment and control, it will cause more diabetes-related complications and seriously affect the patient's life. Because it is difficult to cure itself and the related complications are more serious, people are paying more attention to the early prevention and mid-term treatment of diabetes. A common T1D control method is insulin administration, and insulin can also be provided in some T2D patients. The current drug goals are saving lives and reducing disease symptoms; preventing long-term diabetes complications, and eliminating important risk factors, thereby extending lifespan.

In the medical field, medical prediction is of great significance to medical institutions, patients and countries. Medical prediction is the focus of scholars in related research fields. With the continuous development of Internet technology and medical and health informatization, data warehouse, data mining, machine learning, etc. are widely used in the medical field, which makes it easier to obtain medical information data and provides a good opportunity for medical prediction and development. The electronic medical record system is a typical representative of the implementation of medical informatization in China [3]. Based on EMR, medical data can be obtained more quickly. By digging effective information prompts and early warning medical personnel, providing clinical decision support, it can monitor specific events or early disease prediction [4].

[*] Corresponding author: tangdingjun.dgut@qq.com

2 Electronic medical records and diabetes data mining

2.1. Electronic medical record

The connotation of electronic medical records is basically the same, and they contain two most basic attributes: complete record of long-term medical information of patients, including basic information, medical records, family medical history, test results, etc.; It can be organized, stored, and shared with the help of computer and network technology [5].

Compared with paper medical records, electronic medical records have more advantages. First of all, the electronic medical record is complete in content and uniform in format, which is convenient for storage and retrieval; Secondly, electronic medical records can store data for a long time, and the stored data includes text data, test results, medical images, etc.; Finally, through the implementation of a series of standards and the application of computer technology, electronic medical records can easily achieve data retrieval, data mining, and data sharing, thereby improving medical quality and medical work efficiency, as well as supporting remote consultations and medical consultations [6].

2.2. Diabetes data mining

Using data mining technology, this article mainly studies diabetes prediction research based on electronic medical record data analysis, mainly predicts type 2 diabetes, uses data mining technology to obtain feature information from electronic medical records, and uses artificial intelligence analysis models to make diabetes predictions [7]. On the basis of using data preprocessing methods to reasonably process a large number of raw diabetes data, we obtain useful data; through machine learning algorithm models, we analyze and predict the potential risk factors and the incidence rules of high-risk groups of diabetes. The process of diabetes data mining based on electronic medical records is shown in Figure 1.

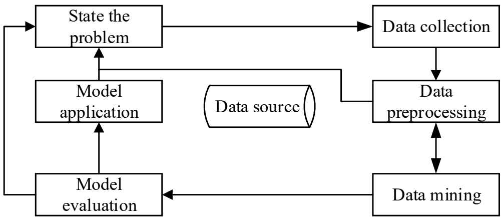

Fig. 1. Diabetes data mining process based on electronic medical records

The original data has problems such as vacancies, inconsistent data, and redundant and redundant data. It is necessary to preprocess the data and clean the relevant data. The processing methods of vacancy values include ignoring records, removing attributes, using default values, using attributes, or sample averages. The method used in this study is to manually delete obviously unreasonable recorded values; for vacant values, remove the irrelevant variable attributes, delete the vacant values or use the average value to fill the relevant vacant variables [8].

3 Feature selection of diabetes prediction model

With the advancement of science and technology, disease prediction models have been continuously developed, and began to use machine learning algorithms such as neural networks for modeling research, which can achieve good disease prediction results.

The data used in this research comes from a hospital's medical examination data set in 2019, which records the physical signs and clinical characteristics of the hospital's population in 2019. The data includes 4632 samples and 305 feature variables. The feature variables of the sample contain a large number of irrelevant feature variables, and related variables need to be selected from them. With reference to the medical literature and relevant physical examination data, the variables related to diabetes include age, gender, race, body mass index, blood pressure, family history of diabetes, SBP, TC, TG, LDL-C, HDL-C, WC, etc. In this study, 21 characteristic variables that may have influence on diabetes were selected as the input and variables of the model, including the patient's age, gender, body mass index, family history of diabetes, family history of hypertension, systolic blood pressure, diastolic blood pressure, total cholesterol and other variables [9].

The statistical description of the quantitative variables is shown in Table 1. The minimum age of the medical examination population is 22, and the maximum is 93, and the average value is 54.4. Body mass index, minimum 14.8, maximum 39, average 24.69, which is within the normal range of 18.5-24.99, but close to the upper limit, indicating that the health of the medical examination population is low. The statistical description of the qualitative variables is shown in Table 1.

Table 1. Statistical variable description

	Min	Max	Avg
Age	22	93	54.4
LDL-C	0.32	8.06	4.19
HDL-C	0.77	4.42	2.60
Triglyceride	0.27	12.16	6.22
Total cholesterol	2.7	9.8	6.25
DBP	47	128	87.5
SBP	82	196	139
BMI	14.8	39	26.9

The qualitative variables such as gender, history of hypertension, and coronary heart disease are all binary variables. There are three situations of daily tastes and daily eating habits, so the statistics of the third information are none of the former.

4 Diabetes prediction model construction

4.1. Choice of diabetes prediction method

Based on the analysis of electronic medical record data, this paper uses a diabetes prediction model based on BP neural network. Machine learning algorithms analyze existing data, obtain laws from them, and use these laws to predict unknown data. Machine learning algorithms have been used to process, analyze, and predict data in multiple fields. Machine learning algorithms mainly include neural networks and machine learning algorithms.

4.2. Data processing before artificial intelligence modeling

After stepwise regression and test on the original data of diabetes electronic medical records, 7 characteristic variables were obtained, and the number of samples was 2618.

In the process of machine learning, the samples are continuously trained to gradually understand the hidden rules of the sample data, and the known rules are used to judge the unknown sample data. In order to improve the promotion ability of the model, the prediction accuracy of the training samples cannot be considered, and the untrained data needs to be added to test the prediction performance of the model. By joining the test set to verify, in order to obtain a better model.

In the study, we divided 2618 samples according to 70%, 20%, and 10%, of which 70% were 1833 training data, and the report included 136 diabetic patients, and 20% were 524 test data, including 42 diabetic patients. 10% are 261 independent sample sets, including 19 diabetic patients. The training set is used to train samples, and the test set is used to verify the prediction effect of the model, and select models with better generalization ability. The independent sample set is used to independently test the prediction of the model and test the model's ability to generalize.

In order to unify the data unit and order of magnitude, after the sample is divided, the data is normalized to narrow the difference in the variable range. For categorical variables, no normalization is required. In this study, when normalizing the data, the centralized variables were converted into variables with mean 0 and variance 1.

4.3. Learning and screening process of BP neural network

In this study, the BP neural network model is implemented through the nnet function in the Rstudio software package. We use test samples to evaluate the generalization ability of the model; The area under the ROC curve (AUC) obtained by the ROC curve test is used as the test standard with the predicted value and the original value of the test set.

When choosing the optimal model, this study uses a cyclic debugging method to combine different parameters in different ways to calculate the test set AUC; find different ranges of variation. The parameter combination with the largest AUC of the test set is used as the parameter of this study. The model definition and training process are shown in Figure 2.

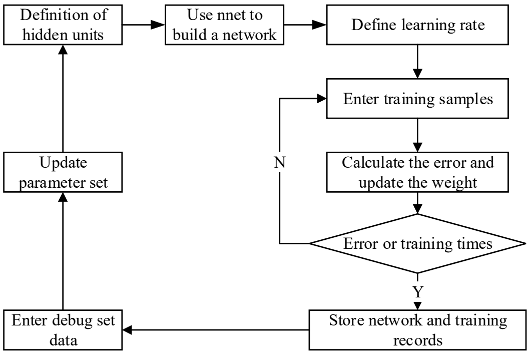
Definition of hidden units
Use nnet to build a network
Define learning rate
Enter training samples
Update parameter set
N
Calculate the error and update the weight
Error or training times
Y
Enter debug set data
Store network and training records

Fig. 2. Artificial neural network modeling flow chart

4.4. Analysis of results

The number of hidden layer units affects the complexity of the model, and the learning rate affects the prediction effect. These are two important parameters of the artificial neural network. When adjusting the parameters, the maximum learning rate of the test set AUC for each number of hidden layer units is required to select the model with the best prediction effect.

The relationship between parameter selection and prediction set fitting effect. With the change of learning efficiency, the change trend of AUC is to decrease first and then increase, and then start to decrease after reaching a certain level; AUC values of the test set are basically distributed in the range of 0.61 ~ 0.81.

Influence of the number of hidden layer units on the prediction effect. Set multiple different hidden layer units to ensure the maximum AUC of each test set. Comparing these artificial neural network models, the final BP artificial neural network model is the largest AUC of the test set as in table 2. According to the maximum principle of the test set AUC, the final model selects a network with 1 hidden layer units, and its topology is 7-1-1.

Table 2. AUC training and test set

Unit Number	Training AUC	Test AUC
1	0.8277	0.8214
2	0.8303	0.8208
3	0.8377	0.8154
4	0.8423	0.7981
5	0.8487	0.8083
6	0.8423	0.8138
7	0.8445	0.8049
8	0.8423	0.8136
9	0.8426	0.8047
10	0.8422	0.8045

5 Conclusion

With the advancement of medical level and the update of medical equipment, clinically, there are many cases that rely on data to judge. How to extract diagnosis and treatment knowledge and apply it in clinical practice as much as possible becomes an important research topic. Based on electronic medical records, extracting clinical diagnosis and treatment information can effectively improve the doctor's clinical diagnosis and treatment decision-making ability, improve disease diagnosis and treatment effect and hospital efficiency. This study is based on the analysis of electronic medical record data, extracting relevant data, as the input parameters of artificial intelligence prediction model, using artificial neural network model to predict diabetes. Using independent sample sets to build a diabetes prediction model based on BP artificial neural network algorithm has achieved good results in many aspects such as effectiveness and accuracy.

References

1. B. Song, L.J. Zhang, Y.X. Feng, Research on electronic medical records and related technologies, China Digital Medicine, 2017, vol.12, pp.10-12.
2. Y.M. Ding, Y.D. Chen, Refined consultation management based on electronic medical record system, Jilin Medical Journal, 2017, vol.9, pp.1796-1797.
3. P. Han, Y.Z. Liu, X.Y. Li, Chinese electronic medical record entity recognition research based on deep learning and multi-feature fusion, Journal of Nanjing University (Natural Science), 2019, vol.6, pp.942-951.
4. M.J. Yang, X.C. Xiong, Construction of Diabetes Knowledge Atlas Based on Reptile Technology and Electronic Medical Records, China Digital Medicine, 2020, vol.2, pp.6-8.
5. G.Y. He, X.C. Xiao, X.N. Xie, et al., Status and prediction of hypoglycemia in diabetic patients, Chinese Journal of Diabetes, 2019, vol.11, pp.877-880.
6. W. Qin, M. Gao, Y. Shen, etc., 3-month blood glucose prediction for type 2 diabetes patients based on machine learning algorithms, Chinese Journal of Disease Control & Prevention, 2019, vol.11, pp.1313-1317.
7. X.H. Wu, Y.P. Zhou, H.H. Xing, etc., Application Research of Machine Learning Classification Algorithm in Diabetes Diagnosis, Computer Knowledge and Technology, 2018, vol.35, pp.177-178 + 195.
8. M.J. Yang, K.X. Pu, J. Li, Data Preprocessing of Diabetes Electronic Medical Records, Journal of Medical Informatics, 2016, vol.5, pp.59-62 + 84.
9. X. Liu, C. Qu, S.T. Wang, et al. Analysis of the rules of Chinese medicine treatment of type 2 diabetes based on data mining, Chinese Archives of Traditional Chinese Medicine, 2019, vol.2, pp.1-17.