DOI: 10.23977/acss.2023.070804
ISSN 2371-8838 Vol. 7 Num. 8

Application of XGBoost Model in the Field of Diabetes Prediction

Yilin Wang[1,a], Wenhao Jiang[2,b]

[1]School of Mechanical and Electrical Engineering, Wuhan University of Technology, Wuhan, 430070, China
[2]College of Medical Information Engineering, Shandong First Medical University & Shandong Academy of Medical Sciences, Tai'an, 271016, China
[a]325536@whut.edu.cn, [b]jwhint@163.com

Keywords: XGBoost model; diabetes prediction; standardized screening

Abstract: Diabetes is a metabolic disorder that threatens people's health, and standardized screening is an important way to diagnose and treat it early. It is low cost and high efficiency to screen through data, therefore, to predict diabetes early has become crucial. Diabetics were taken as the research subject in this paper, and XGBoost algorithm was used to process the patient's data from physical examination, so a model for predicting diabetes was established to predict the blood glucose level of patients and to explore the application of XGBoost model in the field of diabetes prediction. The experimental results have been shown that the mean square error of the sample using this model has been just 0.0598, and it have been verified that the prediction error of the model is small and the accuracy is high, which will soon provide a good means for the pre-screening and clinical prediction of diabetes.

1. Introduction

Diabetes is a metabolic disorder disease, and its incidence rate is increasing rapidly around the world. About 9% people in the world suffer from diabetes [1], which threatens the health of patients and increases the burden of medical resources, so early prediction and accurate diagnosis are particularly important. With the rapid development of machine learning technology, more and more prediction models have been developed at home and abroad, and they have been used widely in the field of diabetes prediction [2,3,4].

As a machine learning algorithm based on gradient lifting tree, XGBoost algorithm has excellent prediction performance and efficient calculation speed. The purpose of this paper is to use XGBoost model to model the clinical data of diabetic patients, build an accurate and reliable diabetes prediction model, and provide efficient and accurate methods for early screening, diagnosis and treatment of diabetes, so as to improve the ability to identify the early risks of diabetes, thus reducing the incidence of complications effectively and improving the quality of life of patients.

In the data preprocessing stage, we will clean the original data of diabetes and divide it into data sets. By reducing the dimension of the data, we can reduce the feature dimension and improve the efficiency and generalization ability of the model. In the model training stage, we will use the training set to train the model, set the parameters of XGBoost model, and update the parameter optimization performance of the model constantly through iterative optimization algorithm to minimize the loss function. After completing the training of the model, we will use the test set for model validation and evaluation to understand the predictive ability and accuracy of the model.

2. Data Pretreatment

2.1 Data Introduction

Based on the data set of diabetes provided by Tianchi, this paper processed and analyzed it, introduced the clinical data and physical examination data of patients, analyzed the main characteristics affecting the blood sugar value of diabetic patients with blood sugar concentration as an index, and tried to design a framework based on machine learning model to predict the blood sugar value of diabetic patients, so as to achieve an accurate and efficient prediction of diabetes risk. People were used as the research subjects in the medical data, and the patient data had the commonness of privacy, polymorphism, lack and timeliness. Because of the complexity of people and the diversity of tests, the types and structures of medical data were diverse, we studied structured medical data principally.

2.2 Data Cleaning

In order to ensure the quality and consistency of the data and improve the accuracy and stability of the model, these data need to be cleaned, including the processing of missing values, outlier, data normalization and standardization:

(1) For missing values, we can choose to delete, fill or forecast. If the missing values account for a small proportion, use interpolation methods to fill the missing data, such as average value, fixed value, nearest neighbor interpolation, function model interpolation or interpolation based on regression model. For data with a missing ratio of more than 70%, we can delete this field, such as hepatitis B core antibody, and at the same time delete the residual "blood sugar" characteristic title in the test set, define this field as the characteristic title to be deleted, and delete the corresponding columns in turn.

(2) For abnormal values, judge according to the distribution characteristics of the data, so as to correct or eliminate them.

(3) Data normalization converts features by scaling each feature to a fixed interval, which can eliminate scale differences between features and is more suitable for situations where numerical values are relatively concentrated. The conversion function is:

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

Among them: $x_{max}$is the maximum value of the sample data, and $x_{min}$is the minimum value of the sample data.

(4) Standardization deals with data according to the columns of the characteristic matrix. The most common is Z-Core standardization. The processed data conforms to the standard Normal distribution, and its conversion function is:

$$x_{new} = \frac{x - \mu}{\sigma} \tag{2}$$

Among them: μ is the mean of the sample data and σ is the standard deviation of the sample data.

Due to the wide range and diverse types of medical data such as diabetes, and the data contains many types of characteristics such as numerical type and classification type, it needs to be properly processed to make it acceptable to the XGBoost model. For classified features, methods such as unique hot encoding and label encoding can be used to convert them into numerical features. In addition, through data consolidation, independent variables can be concatenated into a data matrix in a column manner.

2.3 Partition of Data Sets

Before the prediction model is established, the data set is divided into training set, test set and label set in proportion, for example, 70% of the data is used for training and 30% for testing. The training set is used for model training and parameter tuning, the test set is used to evaluate the performance of the model, and the label set contains the prediction label of each sample, that is, whether diabetes occurs or not. The first behavior field name in the training set and test set file, the first column is the individual ID, and the data contains 41 fields such as numerical type and character type, such as the patient's age, gender, blood sugar level and other information. The last column of the training set file is the label column, that is, the target blood glucose value to be predicted; The last column of the test set file is null, and the actual blood glucose value is stored in the test data label file.

3. Build Model

3.1 Model Introduction

XGBoost algorithm is an iterative decision tree algorithm, which was proposed by Chen Tianqi of Washington University in 2014. It is essentially a gradient boosting decision tree (GBDT), which contains a large number of CART regression trees, which can improve the problem of over-fitting of a single tree model effectively. Because of its excellent learning effect and efficient training speed, it has gained extensive attention since its advent.

The XGBoost model adopts gradient lifting algorithm, and improves the prediction results by training a series of decision tree models iteratively, and improves the performance of the overall model gradually. It supports parallel optimization and column sampling, reduces the time cost of model training, and improves the efficiency and practicability of prediction greatly. It introduces distribution and transforms the loss function into second-order Taylor expansion. According to the first derivative and the second derivative, the basis learner is generated iteratively, which makes the loss more accurate, can customize the loss, and its speed and efficiency are maximized. In addition, it can estimate the importance of each feature to the model prediction and identify the features that play a key role in the prediction results, so as to carry out more targeted feature engineering and model improvement.

In this experiment, the structured single-point data is used to predict the blood sugar value of patients, which belongs to the regression prediction problem. It is also possible to change the labels into high and low categories and turn them into classified prediction problems by setting the diabetes risk threshold. The XGBoost model is helpful to deal with the problems of nonlinear relationship, interaction and sparseness of features in data. Combining the known factors such as patient's detection information, the model can be used to make a preliminary diagnosis and prediction quickly for diabetic patients. What’s more, it can also predict the future development of patients with chronic diseases, and the accuracy of the current model can reach a high level.

3.2 Build XGBoost Model

3.2.1 Dimension Reduction and Model Training

In order to accelerate the training speed of the model and reduce the risk of overfitting, we used the Principal Component Analysis (PCA) to reduce the dimensions of the sample data. It can not only compress the data, but also make the data features independent of each other after dimensionality reduction. If we want to minimize the projection error, we need to find k vectors to project the original data to reduce the dimensions from n to k, the formula is as follows:

$$\frac{\frac{1}{m}\sum_{i=1}^{m}||x^{(i)}-x_{approx}^{(i)}||^2}{\frac{1}{m}\sum_{i=1}^{m}||x^{(i)}||^2} \leq 0.01 \tag{3}$$

In the formula above, m represents the number of features, and the numerator above represents the sum of projected distances. The smaller the error, the more complete the data before the dimensionality reduction is. In practical application, it is not necessary to select k here, but set the PCA method parameter to 0.01, so that 99% of the information is retained after dimensionality reduction. The feature dimension is reduced by dimensionality reduction to reduce the computation cost, and the feature with the most information is retained.

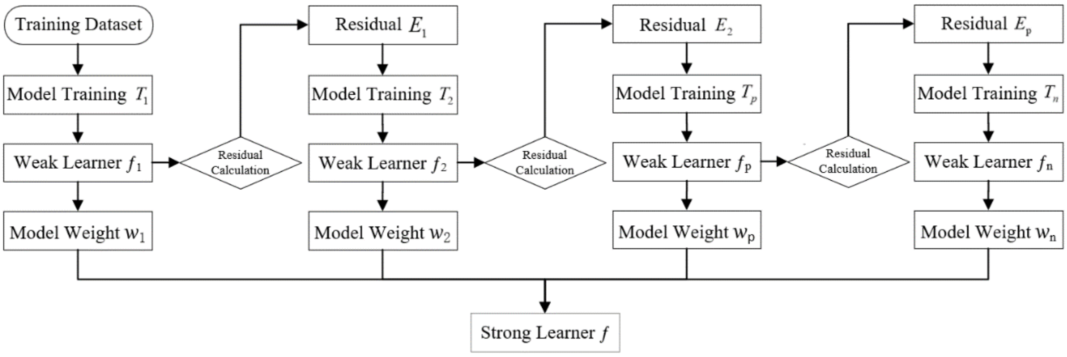

Figure 1: The Process of Model XGBoost Building

This experiment was implemented in Python language, and Pycharm was used as the simulation platform. As shown in Fig.1, the XGBoost model was trained using training sets [5]. The training process involved generating the decision tree model iteratively and optimizing it according to the optimization criteria of loss function.

3.2.2 Determine Parameter of Model

Before establishing the XGBoost model, the parameters of the model need to be set, which are divided into general parameters, Booster parameters and learning target parameters (As shown in Table 1). General parameters include the number of threads, etc. Booster parameters include learning rate parameters, number of CART trees, depth, and model complexity, etc. Learning target parameters include classification, regression and other parameters.

To prevent overfitting, XGBoost introduces the idea of Shrinkage, not fully trusting the residual learned by each weak learner. Therefore, the residual value fitted by each weak learner needs to be multiplied by the step size of each iteration. If the step size is too large, the running accuracy will be reduced, while the step size is small, several weak learners can be learned to make up for the insufficient residual, but the running speed will be slowed down.

The maximum depth of the tree is used to control overfitting. The deeper the tree is, the more information about the data can be captured, the model learning is more specific at the moment. The subsample parameter controls the proportion of the subsamples used in each decision tree to the total samples. Setting it as 0.8 means that 80% samples are selected for GBDT decision tree fitting, which can reduce variance and prevent overfitting, but the deviation of sample fitting will be increased.

Table 1: Parameters of the XGBoost Model

Type	**Parameters**	**Meaning**	**Value**
General	booster	Type of weak learner.	gbtree
Booster	gamma	Minimum loss function decline required for node splitting.	0.1
	learning_rate	The step length of each iteration.	0.1
	subsample	The proportion of random samples for each tree.	0.8
	colsample_bytree	The proportion of column numbers randomly sampled per tree.	0.8
	max_depth	The maximum depth of the tree.	10
	min_child_weight	The sum of the smallest sample weights in the child nodes.	12
Learning Target	objective	Learning tasks and corresponding learning objectives.	reg:linear
	eval_metric	Evaluation index.	rmse

The setting of these parameters affects the performance and generalization ability of the model directly. For XGBoost models, we use methods such as cross-validation to find the best combination of parameters for better model performance.

3.2.3 Model Export

The trained XGBoost model was used to test the test set. The performance of the model was evaluated by comparing the predicted value of blood glucose with the real value and calculating the accuracy rate, accuracy rate and recall rate. Finally, based on the results of the model performance evaluation, the diabetes predictive output results can be derived. These results can be either a prediction label for a binary classification or a prediction probability. We predicted the blood glucose level in this experiment.

4. Model Solution

4.1 Evaluation Index

Area Under Curve (AUC) and accuracy is usually used as an evaluation index for prediction problems, while mean square error is commonly used as an evaluation index for regression problems. Different indicators evaluate the performance of machine learning models from different perspectives.

To predict the blood glucose detection value of patients, it is necessary to compare the predicted value with the actual blood glucose. The smaller the error, the more accurate the prediction. Used as a statistical measure and loss function in regression models, Mean Square Error (MSE) measures the deviation between the predicted value and the true value, the calculation formula is as follows:

$$MSE = \frac{1}{\mathrm{n}}\sum_{i=1}^{n}(\hat{y}_i - \mathrm{y_i})^2 \tag{4}$$

In the formula above, n represents the total number of patients, it is the predictive value of blood glucose in patient i. And $\mathrm{y_i}$ represents the actual blood sugar of patient i. Root Mean Square Error (RMSE) is more sensitive to the overall error size because it amplifies the larger error squares, the calculation formula is as follows:

$$RMSE = \sqrt{\frac{1}{\mathrm{n}}\sum_{i=1}^{n}(\hat{y}_i - \mathrm{y_i}\,)^2} \tag{5}$$

Mean Absolute Error (MAE) is more sensitive to outliers because it is considered that the absolute value of the error only, the calculation formula is as follows:

$$MAE = \frac{1}{\mathrm{n}}\sum_{i=1}^{n}|\hat{y}_i - \mathrm{y_i}| \tag{6}$$

The smaller the value of MSE, RMAE and MAE, the more accurate the prediction. When the value is 0, the predicted value is consistent with the true value exactly.

4.2 Model Export

The model was trained by the test set and the blood glucose values predicted by the test set were saved to a CSV file. As shown in Fig. 2, the predicted value were compared with the actual value, and we drew a scatter plot to visualize the predicted result.

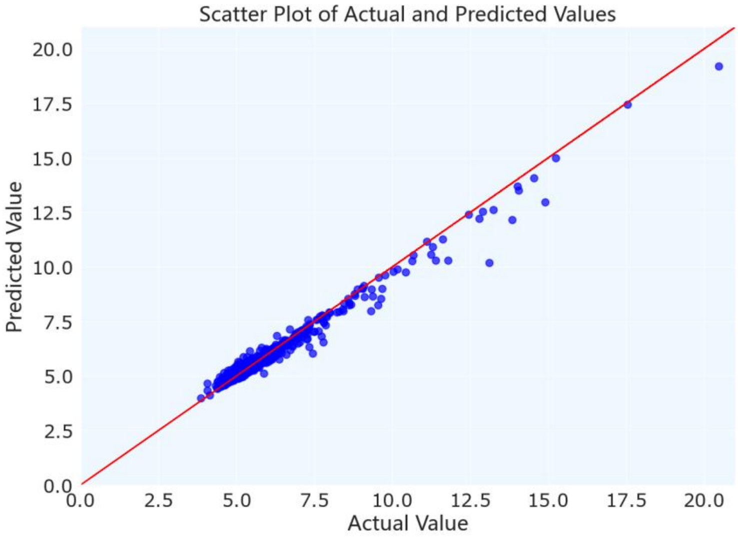

Figure 2: Scatter Plot of Actual and Predicted Values

The values of MSE, RMSE and MAE were calculated after predicting the results. It will have a great impact on the performance of the model for the difference in data processing, including the selection of features, the generation of features and so on. It is also a project to adjust the parameters of the model itself, therefore, it's also important to choose the optimal parameter to achieve the best predicted results.

After obtaining a relatively good predictive performance, further research can be divided into two broad categories. One is to increase the interpretation of data, which is particularly important in medical data. The other is to improve the performance of model sequentially. Both of these types require data exploration, that is, analyzing some data one by one.

We explored whether each column of data conforms to a normal distribution, exemplified by Statistical analysis of triglycerides, we calculated the mean and standard deviation. As shown in Fig. 3, it was tested for the probability distribution of the sample data.

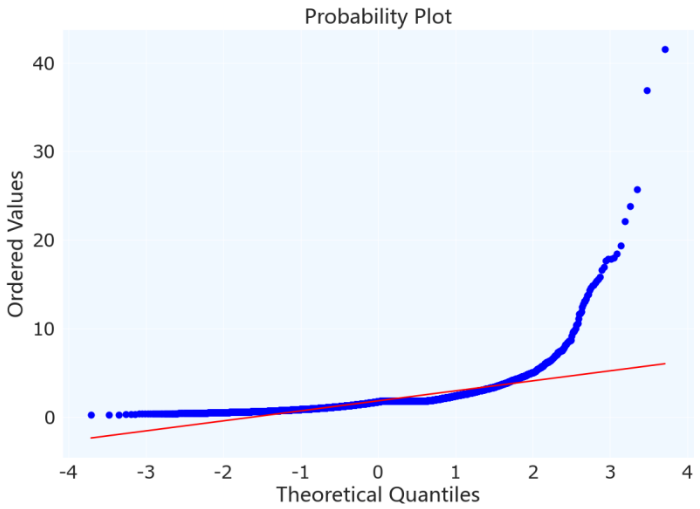

Figure 3: Probability Distribution of Sample Data

In the picture above, the red line represented the normal distribution, while the blue represented the sample data. The closer the blue is to the red reference line, the more consistent the expected distribution is.

We found that the generated distribution was more consistent with a skewed distribution than a normal distribution. In order to improve the performance of the model, the skewness data can be transformed into normally distributed data to improve the data. The heat map of the correlation matrix was drawn in our experiments, it could help us analyze the data comprehensively and find highly correlated variable combinations to generate new features.

4.3 Model Analysis

After the model was trained, the predicted value should be compared with the actual value. As shown in Table 2, the values of MSE, RMAE and MAE should be calculated to evaluate the model. Through the parameters for model evaluation, it was not difficult to find that the XGBoost model has a small error in the prediction results.

Table 2: Parameters for Model Evaluation

Model	MSE	RMSE	MAE
XGBoost	0.0598	0.2445	0.1369

4.4 Model Application

The XGBoost model deals with diversified and complex medical data flexibly because of strong expandability, so it can mine the data for information better and improve the accuracy and interpretability of the prediction. This model has a wide range of application scenarios and important significance.

Early diabetes risk identification and intervention can be carried out through this model, and doctors combine clinical data and physical examination results to identify the risk level of patients accurately in the early stage of diabetes development and develop a personalized preventive intervention programs. For large populations, the model can identify those at high risk through diabetes risk screening. When integrated into a clinical decision support system, based on the characteristic data of patients, personalized advice and guidance can be generated automatically to help doctors develop more effective treatment plans.

This model reduce the risk of diabetes effectively, delay the progression of the disease and reduce the occurrence of related complications. It is helpful to optimize the utilization of resources and improve the efficiency of screening. What's more, it improves the decision-making effect of doctors and the treatment results of patients, so it has important significance in clinical decision support.

4.5 Experimental Conclusion

In this study, XGBoost model was used to predict the blood glucose of diabetic. We calculated that MSE was 0.0598, RMSE was 0.2445 and MAE was 0.1369, thus, the overall error is small, which verified that the advantages of XGBoost model prediction accuracy and high practicability. We analyzed the data through thermal maps to explore the importance of features and find a combination of variables with strong correlations.

5. Conclusion

Diabetics were taken as the research subject in this paper, we aimed to study the effect of a large number of characteristics on blood glucose through data from patient. After the XGBoost model was built and trained, we used MAE, MSE and RMSE as the measurement indicators, the blood glucose of patients could be predicted more accurately and reasonably through this model. Effective information was extracted through data merging and dimension reduction to reduce errors caused by redundant information, and the accuracy and convergence speed of the model were improved significantly through data normalization. The experimental results have been shown that the mean square error of the sample using this model has been just 0.0598, which proved that XGBoost model had strong applicability, high prediction accuracy and generalization ability when applied to the field of diabetes prediction. This model, meanwhile, will soon provide a good means for the pre-screening and clinical prediction of diabetes.

References

[1] Wenlong Qu , Yiyi Li, Lei Z .Application of XGBoost algorithm in diabetic blood glucose prediction[J].Jilin Normal University Journal(Natural Science Edition), 2019.

[2] Cahn A, Shoshan A, Sagiv T, et al. Prediction of progression from pre-diabetes to diabetes. development and validation of a machine learning model[J]. Diabetes/Metabolism Research and Reviews, 2020, 36(2):e3252.

[3] Moreno LM. Vergara J, Alacon R. Predictive risk model for the diagnosis of diabetes mellitus type 2 in a follow-up study 15 years om. PROD12 Study [J]. European Journal of Pubic Heath, 2019, 29(1):178-182.

[4] Zuo D, Zhao XL, Dai XL. Construction and verification of hypoglycemia risk prediction model in patients with type 2 diabetes [J]. Journal of nursing science, 2021, 36(1):30-33.

[5] Su X, Guo CR, Li YJ. Research on precision milling quality prediction based on XGBoost algorithm [J]. Machine building and automation, 2023, 52(02):72-76. DOI:10. 19344/j. cnki. issn1671-5276. 2023. 02. 020.