The Risk Prediction of Type 2 Diabetes based on XGBoost

Wei Ji[a], Shaofu Lin[b]

Faculty of Information Technology, Beijing University of Technology, Beijing, China

[a]15611635876@163.com, [b]linshaofu@bjut.edu.cn

Keywords: XGBoost; type 2 diabetes; risk prediction

Abstract: This paper applies the XGBoost method to construct a predictive model for the risk of type 2 diabetes which based on the physical examination data. The paper takes the real physical examination records of the same batch of people in a health check-up center from 2010 to 2015 as the data source, and evaluates the feature importance. Finally, 28 characteristic variables are selected as the model input, and a phase is obtained. Compared with other common classification algorithms, the prediction model with higher prediction accuracy and stronger generalization ability has certain clinical reference value for the risk prediction of type 2 diabetes.

1. Introduction

Diabetes is a kind of lifelong metabolic diseases characterized by chronic hyperglycemia caused by multiple causes. Long-term blood sugar increases damage to large blood vessels and microvessels and endangers the heart, brain, kidney, peripheral nerves, eyes, feet, etc. The complications are more than 100, which is the most common complication of diseases [1].

According to the 2017 Global Diabetes Map (8th edition) report published by the International Diabetes Federation (IDF), approximately 425 million adults worldwide with diabetes in 2017, and it is estimated that by 2045, diabetes patients may reach 629 million. The data shows that there are about 114.4 million people with diabetes in China, and about 84,939 patients die of diabetes, of which 33.8% are younger than 60 years old [2]. At present, there is no cure for diabetes, so early intervention and treatment of diabetes is of great significance.

Based on the continuous annual batch data collection of the same batch of people, this paper applies XGBoost algorithm to establish a high-predictive risk prediction model for type 2 diabetes. After the model is established, the risk of diabetes in the next year can be predicted according to the current physical examination index which provides a credible basis for predictive diagnosis of diabetes and early intervention.

2. Overview of diabetes prediction models and introduction to xgboost

In recent years, the rapid development of a new generation information technology has provided new ideas for the prediction of diabetes in China and abroad. For example, Jiang Lin used the support vector machine to establish a type 2 diabetes prediction model, Qian Ling's predictive model of diabetes and impaired glucose tolerance based on the decision tree method, Jin Park's risk prediction model based on neural network, and association rules for the use of Simon GJ to develop diabetes-related factors and people susceptible to diabetes [3,4,5,6]. The applied algorithms have their own advantages and disadvantages, and the predicted response values have a long interval. The XGBoost used in this paper is an integrated learning algorithm based on gradient boosting. It makes full use of multi-core CPU for parallel computing and improves accuracy. At the same time, according to the current physical examination data indicators, it can timely and accurately predict the risk of the next year.

XGBoost is called eXtreme Gradient Boosting, which means extreme gradient lifting tree. It is an integrated learning algorithm based on gradient boosting. Its principle is to achieve accurate classification effect through iterative calculation of weak classifier [7]. The XGBoost model can be

expressed as

$$\hat{y}_i = \Sigma_{k=1}^{K} f_k(x_i), f_k \in F \quad (1)$$

i is the number of samples, F is the corresponding the set of all regression trees, is a function in F. When building a f_k model, you need to find the optimal parameters, usually choose the parameters that make the value of the objective function the smallest. The objective function usually includes the loss function L(θ) (related to the task) and the regular term Ω(θ) (related to the complexity of the model), which can be expressed as

$$O_{bj}(\theta) = L(\theta) + \Omega(\theta) \quad (2)$$

$$L(\theta) = l(y_i, \hat{y}_i) = \Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (3)$$

$$\Omega(\theta) = \Sigma_{k=1}^{K}\Omega(f_k) \quad (4)$$

The purpose of the loss function is to select the best model to match the training data. The regular term makes the model the simplest. The model has the characteristics of high accuracy, not easy to over-fitting and strong scalability [7].

3. Data and methods

3.1 Data Set Description

The model expects to adopt the data set prerequisites: the same batch of people to check the data set for two consecutive years. The time dimension is not limited. For example, the physical examination data of Zhang from 2010 to 2011 and the physical examination data of Li from 2012 to 2013 are both valid data. The data comes from the physical examination records of a health check-up center from 2010 to 2015. Each sample contains two consecutive physical examination data. The first year data includes basic conditions, color ultrasound, blood routine, urine routine, and biochemistry. The characteristics are 5 quantitative features and 44 qualitative features. The second year data is blood glucose level.

Table 1 Physical Examination Data Details

Project	Index	Remarks
Basic Situation	number	ID code
	gender	0: male, 1: female
	age	numerical value
	systolic blood pressure	numerical value
	diastolic blood pressure	numerical value
Color Ultrasound	fatty liver	0:yes, 1: no
Blood Routine	hemoglobin concentration	numerical value
	red blood cell count	numerical value
	hematocrit	numerical value
	average red blood cell volume	numerical value
	white blood cell count	numerical value
	percentage of lymphocytes	numerical value
	number of neutrophils	numerical value
	neutrophil percentage	numerical value
	platelet count	numerical value
	mean platelet volume	numerical value
	platelet pressure	numerical value

	platelet distribution width	numerical value
Urine Routine	urobilinogen	numerical value
	colour	0: light yellow, 1: yellow
	leukocyte	numerical value
	sharpness	0: clear, 1: unclear
	nitrite	0: negative, 1: positive
	occult blood	numerical value
	protein	numerical value
	PH	numerical value
	proportion	numerical value
	ketone body	numerical value
	bilirubin	numerical value
	glucose	numerical value
Biochemistry	alanine aminotransferase	numerical value
	aspartate aminotransferase	numerical value
	alkaline phosphatase	numerical value
	glutamyltranspeptidase	numerical value
	total bilirubin	numerical value
	direct bilirubin	numerical value
	total protein	numerical value
	albumin	numerical value
	globulin	numerical value
	white ball ratio	numerical value
	urea nitrogen	numerical value
	uric acid	numerical value
	creatinine	numerical value
	indirect bilirubin	numerical value
	total cholesterol	numerical value
	triglyceride	numerical value
	high density cholesterol	numerical value
	low density cholesterol	numerical value
	glucose	numerical value

The goal of this study is the risk prediction of type 2 diabetes. There is a certain chance of a single blood glucose test, and the blood glucose level alone cannot be used as a standard for the diagnosis of type 2 diabetes. According to the medical setting of the normal value of fasting blood glucose: (1) When the fasting whole blood glucose is above 5.6 mmol / liter (100 mg / dl), plasma blood glucose is above 6.4 mmol / liter (115 mg / dl), the glucose tolerance test should be done. (2) Fasting blood glucose≥6.7 mmol / liter (120 mg/ dl), plasma glucose≥7.8 mmol / liter (140 mg/dl), 2 repeated measurements can be diagnosed as diabetes. (3) When the fasting whole blood glucose exceeds 11.1 mmol/L (200 mg/dl), it means that there is little or no insulin secretion [8]. Therefore, when the fasting blood glucose is significantly increased, it is diagnosed as diabetes without further examination. The blood glucose values of the second year are divided into 5.6, 6.7, and 11.1 to establish four numerical intervals, which are defined as low, medium, medium high, and high, respectively, to represent the risk of type 2 diabetes, instead of the second year's blood glucose value in the original data set as a response feature.

Table 2 Risk Classification

Numerical Interval	x≤5.6	5.6＜x＜6.7	6.7≤x≤11.1	x＞11.1
Degree of Risk	low	medium	medium high	high

3.2 Data Preprocessing

Data preprocessing plays a crucial role in model construction and directly affects the final output of the model. The original data set has the problems of vacancy value, data inconsistency, redundant repetition, high noise, high dimension and so on. The data needs to be cleaned. The data problem of the original data set is mainly due to some data noise anomalies and the existence of vacancy values. The methods for dealing with this kind of problem are to ignore the record, delete the attribute, manually fill in the vacancy value, use the default value, use the attribute average, use the same sample average, subjective prediction possible value, etc [9]. The method used in this paper is obviously abnormal for noise. The value is manually deleted. For the vacancy value, the attribute is removed for the obviously unrelated variable, and the relevant variable is filled with -1. After the data processing work, the available samples were determined to be 2063 cases. The basic data was built and randomly divided into training sets and test sets, 80% of the samples were training sets and the rest were test sets.

3.3 Feature Extraction

There are many dimensions and information in the original data, but there are some information that have little correlation with the response characteristics of the second year. If all of them are modeled as important features, the modeling process will be inefficient and the model stability will be poor. XGBoost can filter the feature importance to achieve the goal of improving the performance of the model and the dimensionality reduction of the original data. The basic idea is to calculate which segment of the feature is selected according to the gain of the structure score. The importance of a feature is the sum of the occurrences in all trees [7,10]. Enter the preprocessed data into XGBoost and call plot_importance to see the importance of each feature.

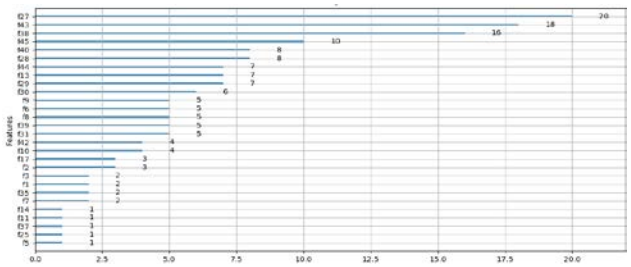

Figure 1. Feature Importance

The figure shows that there are 28 characteristics of the original data that have an effect on the response variables, namely urine sugar, triglyceride, urea nitrogen, low density cholesterol, creatinine, alanine aminotransferase, high density cholesterol, mean platelet volume, aspartate aminotransferase, alkaline phosphatase, percentage of lymphocytes, hematocrit, white blood cell count, uric acid, glutamyltranspeptidase, total cholesterol, number of neutrophils, color, diastolic blood pressure, fatty liver, systolic blood pressure, albumin, mean red blood cells volume, platelet count, neutrophil percentage, white ball ratio, ketone body, and red blood cell count were used as variables for the model.

3.4 Determine Parameters

The XGBoost model has many parameters, which can be roughly divided into three categories: general parameters, booster parameters, and learning target parameters. The maximum depth of the tree in the booster parameter, the learning rate, the minimum function drop required for node splitting, the minimum sample weight required for the leaf node, the sample sampling ratio for

constructing each tree, the proportion of features used to construct each tree. The proportion of the features used in each split of the tree and the L2 regular penalty coefficient have a great influence on the performance of the constructed model.

In this paper, the following methods are used to determine the parameters. First, set a higher learning rate and select the number of ideal decision trees corresponding to this learning rate. Then use the grid search method to adjust the specific parameters such as the decision tree depth. Next, adjust the L2 regular penalty. The coefficients reduce the complexity of the model, and finally reduce the learning rate and the number of decision trees and determine the ideal parameter values. The parameters are shown in the following table.

Table 3 Parameter Setting

Parameter Name	Numerical value
max_depth	9
gamma	0
subsample	0.8
colsample_bylevel	1
learning_rate	0.2
min_child_weight	0.9
colsample_bytree	0.7
reg_lambda	1

4. Result analysis

In the XGBoost model with 28 characteristics input parameters that have an influence on the response variables, the optimal state model is trained, and then the trained model is used to predict the test set results. The classification accuracy is 87.16%. In order to better consider the effect of the model, the current commonly used classification algorithm KNN, Random Forest, and Logistic Regression were used to conduct comparison experiments on this data set. The accuracy rate was used as the evaluation function of the prediction model. The results are shown in the following

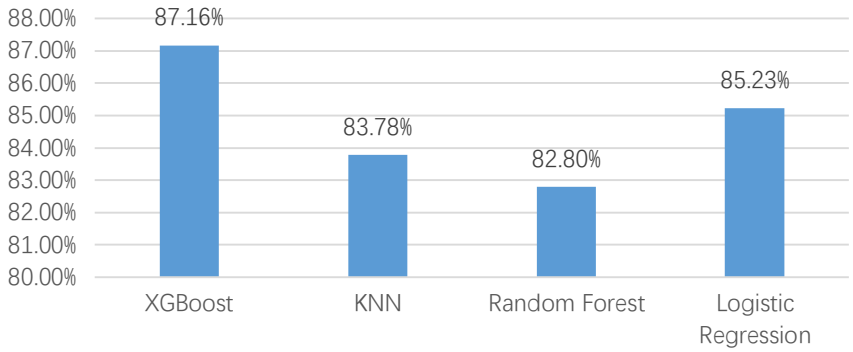

Figure 2. Comparison Test Results

From the results of the control experiment, it can be seen that the XGBoost model predicts the risk of type 2 diabetes under the data set, and the best fit is better for other commonly used classification algorithms.

5. Conclusion

From the aspect of model prediction effect, the XGBoost-based prediction model for type 2 diabetes risk based on this paper is successful, and its prediction accuracy of 87.16% is obviously superior to other common classification algorithms such as KNN, Random Forest and Logistic

Regression. Good stability and high prediction accuracy can play an active role in early warning and effective intervention of high-risk populations with diabetes, and have strong operability and generalization.

References

[1] Hayes A, Arima H, Woodward M, et al. Changes in Quality of Life Associated with Complications of Diabetes: Results from the ADVANCE Study[J]. Value in Health the Journal of the International Society for Pharmacoeconomics & Outcomes Research, 2016, 19(1):36-41.

[2] International Diabetes Federation (IDF) (2017) IDF Diabetes Atlas. 8th Edition, International Diabetes Federation, Brussels. http://www.diabetesatlas.org/resources/2017-atlas.html

[3] Jiang Lin, Peng Li. Discrimination and Feature Screening of Type II Diabetes Based on Support Vector Machine[J]. Science Technology and Engineering, 2007, 7(5): 721-726.(in Chinese)

[4] QIAN Ling, SHI Lu-yuan, CHENG Mao-jin. Artificial neural network applied to predicting the incidence of diabetes and impaired glucose tolerance in individuals[J]. Chinese Journal of Prevention and Control of Chronic Diseases, 2005, 13(6): 277-280.(in Chinese)

[5] Jin P, Edington D W. A sequential neural network model for diabetes prediction[J]. Artificial Intelligence in Medicine, 2001, 23(3):277-293.

[6] Simon G J , Caraballo P J , Therneau T M , et al. Extending Association Rule Summarization Techniques to Assess Risk of Diabetes Mellitus[J]. IEEE Transactions on Knowledge and Data Engineering, 2015, 27(1):130-141.

[7] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System[C]// Acm Sigkdd International Conference on Knowledge Discovery & Data Mining. 2016.

[8] Kang Jian. The Ins and Outs of Blood Sugar[J]. Life and Health, 2008(7):28-29. (in Chinese)

[9] García S, Luengo J, Herrera F. Data Preprocessing in Data Mining[J]. Computer Science, 2000, 72.

[10] Chen W, Fu K, Zuo J, et al. Radar emitter classification for large data set based on weighted-xgboost[J]. Iet Radar Sonar & Navigation, 2017, 11(8):1203-1207.