Diabetes Prediction Based on XGBoost Algorithm

Mingqi Li¹, Xiaoyang Fu¹* and Dongdong Li²

¹School of Computer Science, Zhuhai College of Jilin University, Zhuhai, China
²Lilian Information Technology Company, Beijing, China

*Corresponding author e-mail: 93964281@qq.com

Abstract. Exploring important features of diabetes through analytical methods of data mining is able to predict and prevent diabetes. This paper proposes a diabetes prediction algorithm based on XGBoost algorithm with the numerical features being separated while some important features are extracted from the text features of experiment data. Experiment results show that accuracy of diabetes prediction based the improved XGBoost algorithm with features combination is 80.2%, which is feasible and effective method for diabetes prediction.

1. Introduction

Chronic diseases, which has become the biggest obstacle to the development of medical and health care in China, are a kind of lifelong diseases with high disability rate, high mortality rate. The studies show that disease prevention is the most effective measure in the management of chronic diseases. With development of the Internet technology, the form and quantity of medical data are increasing exponentially. People find out gradually that machine learning methods can achieve better prediction accuracy or probe the inherent law [12,13] while dealing with complex data [1,2,3].

In recent years, Su Tianpei has used the XGBoost algorithm and random forest for the prediction of diabetes [4] as well as Liu Yu et al. combined clustering and XGBoost algorithm for heart disease prediction [6]. Some machine learning algorithms, such as Support Vector Machine, Decision Tree, Naive Bayes, Logistic Regression, BP neural network have also been used by the industry and academia for disease prediction [7], but the existing methods still have major limitations. GBDT (The Gradient Boosting Decision Tree) algorithm is improved by Chen Tianqi, and an optimal distributed decision gradient lifting library XGBoost (Extreme Gradient Boosting) is proposed. The principle is implemented by iterative calculation of weak classifier. In order to get accurate classification results, this paper proposes to make use of XGBoost algorithm with data preprocessing in the prediction of diabetes with the experiment data from National Institute of Diabetes and Digestive and Kidney Diseases[14], establishing a classification model to predict the results. The rest of the paper is organized as follows: section 2 briefly concludes the XGBoost algorithm and model training mechanism, the experiments and analysis in section 3, and finally, conclusion in section 4.
2. Methodology

2.1 XGBoost algorithm
XGBoost is an integrated model based on decision tree [6,7].

Target optimization function:

\[
\text{Obj}^t = \sum_{i=1}^{n} [l(y_i, \hat{y}_i^t) + f_t(x_i)) + \Omega(f_t) + \text{const} \tan t] \\
\]

Taylor expansion:

\[
f(x + \Delta x) \approx f(x) + f'(x)\Delta x + \frac{1}{2} f''(x)\Delta x^2
\]

Model complexity calculation formula:

\[
\Omega(f_t) = \gamma T + \lambda \frac{1}{2} \sum_{j=1}^{T} w_j^2
\]

At the same time:

\[
g_i = \frac{\partial L(y_i, \hat{y}_i)}{\partial \hat{y}_i}, h_i = \frac{\partial^2 L(y_i, \hat{y}_i)}{\partial \hat{y}_i^2}
\]

Substituting equations (3), (4), and (5) into equation (2), the final result is:

\[
\text{Obj}^t = \sum_{j=1}^{T} [(\sum g_i)w_j + \frac{1}{2}(\sum h_i + \lambda)w_j^2] + rT
\]

Let \( G_j = \sum g_i, H_j = \sum h_i \):

\[
\text{Obj}^t = \sum_{j=1}^{T} [G_j w_j + \frac{1}{2}(H_j + \lambda)w_j^2] + rT
\]

Solve the equation (6) and find the best \( w_j^* \) and the corresponding objective function optimal value.

The two results correspond to the following:

\[
w_j^* = \frac{G_j}{H_j + \lambda}
\]

Target result:

\[
\text{Obj} = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + rT
\]

It is usually impossible to enumerate all possible tree structures to add the model by solving the objective function to find a tree of optimal structure, so we use the greedy algorithm to find the optimal segmentation scheme in the end.

2.2 Model training mechanism
Firstly, the input raw data is preprocessed into text data and numerical data after denoising and removing abnormal value. After the separated processing, the important features affecting diabetes are extracted by RFE and chi-square test [4], training a prediction classification model with the XGBoost algorithm in the end. shown in Fig 1:
3. Experimental results and analysis

3.1. Data preprocessing

The experimental data in this paper are derived from the physical examination data of more than 768 people from National Institute of Diabetes and Digestive and Kidney Diseases. We make use of some data preprocessing methods, such as null value filling, denoising, one-hot encoding, removing some abnormal features, keyword extraction, outlier processing etc. to handle the raw diabetes data [11]. Data after feature extraction is shown in Table 1.

**Table 1. Processed data sets**

| No. | Data Name               | Data type       |
|-----|-------------------------|-----------------|
| 1   | Outcome                 | 0/1             |
| 2   | Pregnancies             | Numerical type  |
| 3   | Glucose                 | Vector type     |
| 4   | Blood Pressure          | Numerical type  |
| 5   | Skin Thickness          | Numerical type  |
| 6   | Insulin                 | 0/1             |
| 7   | BMI                     | Numerical type  |
| 8   | Diabetes Pedigree Function | Numerical type |
| 9   | Age                     | Numerical type  |

**Fig 1.** Diabetes prediction model trained based on XGBoost algorithm
3.2. Data Correlation

After data preprocessing, we have to learn whether or not the data is processed properly and how much the correlation between the data is. The correlation coefficient can be used to reflect the close relationship between the variables. The correlation coefficient is calculated by the difference method. It is also based on the dispersion of the two variables and their respective averages. The two differences are multiplied to reflect the degree of correlation between the two variables. The linear single correlation coefficient is studied as shown in Fig 2:

From correlation coefficient matrix, we can figure out whether or not there is a linear correlation between diabetes and various parameters, such as the number of pregnancies is 0.22, the glucose content is 0.47, the blood pressure concentration is 0.065, the cortical thickness is 0.075, the insulin content is 0.13, and the body mass index is 0.29. Diabetes lineage the function is 0.17 and the age is 0.24. We can conclude that glucose, body weight and age are much more relative to the diabetes than that of several other factors. As shown in Fig 3:

3.3. Result analysis

By the iterated training and adjustment parameters of the training set, the comparison results is showed between XGBoost and the traditional algorithm models in Table 2:
Table 2. Model comparison results

| No. | Algorithm | Accuracy | Operation hours |
|-----|-----------|----------|-----------------|
| 1   | SVM       | 0.655    | 1.269           |
| 2   | KNN       | 0.719    | 1.253           |
| 3   | NB        | 0.767    | 1.024           |
| 4   | DT        | 0.703    | 1.254           |
| 5   | LR        | 0.767    | 1.121           |
| 6   | XGBoost   | 0.812    | 0.203           |

From Table 2, we compare the XGboost algorithm with other traditional algorithm on training prediction model. The experimental results show the best accuracy of all six algorithm with the least training time is XGboost. It is extremely important to reduce the complexity in the face of vast amounts of cumbersome data.

In order to find out optimal solution, we have further improved the XGBoost algorithm compared to that combining with other algorithm. The principle of XGBoost + LR combination is to train an XGBoost model with data, then give the instance of the training data to the XGBoost model to get the leaf node of the instance, and then use the leaf node as a feature training. Cross-combining different features use iterative methods to find out the optimal subset in order to generate new derived variables, which may affect the accuracy of prediction and more likely to have clinical significance meaning to extract valuable information for clinicians. As shown in Table 3:

Table 3. Model comparison results

| No. | Algorithm model | Accuracy | Recall | F1  |
|-----|----------------|----------|--------|-----|
| 1   | XGBoost        | 0.796    | 0.660  | 0.722|
| 2   | XGBoost+LR     | 0.779    | 0.674  | 0.636|
| 3   | Data feature stitching +XGBoost | 0.802 | 0.706 | 0.750 |

The XGBoost algorithm uses sigmoid as the activation function, the final predicted value of XGBoost is equal to the sum of the weights of the leaf nodes.

From the above results, XGBoost + LR is close to or weaker than XGBoost because XGBoost can learn appropriate weights of some leaf nodes while LR do not have much help for learning better weights for leaf nodes. Although the accuracy of improved XGBoost model (Data feature stitching +XGBoost) is similar with the XGBoost algorithm, the recall rate shows the significantly better (0.706) than the XGBoost algorithm by itself(0.660). That means the data feature combination is help for improving the diabetes prediction. We divide the glucose content into low [0-90], medium [90,180], and high [180,250] and divide the family genetic coefficients related to diabetes into yes and no, and combine these two features into 6 new labels. That is because what has a diabetes family history and a high glucose content has the highest probability of illness while if there is no diabetes family history and the glucose content is very low, the probability of illness is very small.

3.4. Feature importance assessment

The advantage of using the gradient boost algorithm is that importance score for each attribute can be obtained relatively directly after the promotion tree is created. In general, the importance score measures the value of the feature in the construction of the elevated decision tree in the model. Attribute importance is calculated by sorting and sorting each attribute in the dataset. The importance of the attribute is calculated by improving the amount of performance metrics by each attribute split point in a single decision tree, and the node is responsible for the weighting and number of records.
The performance metric can be either the Gini purity of the split node or other metric functions. Finally, the results of an attribute in all the promotion trees are weighted and summed, and then averaged to obtain the importance score. As shown in Fig 6:

![Feature importance of the XGBoost algorithm](image)

**Fig4.** Feature importance of the XGBoost algorithm

While carrying out diabetes prediction, the model can also give the order of importance feature for improving the accuracy of the prediction model through the tree model mechanism. These important features can also be used as some clinical reference value for doctors.

4. Conclusion

Data preprocessing is an important prerequisite for model accuracy guarantee, making the XGBoost modeling process more efficient. The comparison of the algorithms makes it clear that XGBoost is more efficient among some traditional algorithms. Through the comparative analysis with the integrated algorithm, we proposed the improved feature combination algorithm based XGBoost. The experiment results show that it is high prediction accuracy, good stability and high running speed to build diabetes prediction model based on the improved algorithm (Data feature stitching+XGBoost).

5. Acknowledgments

This work was financially supported by the Premier Key-Discipline Enhancement Scheme Supported by Guangdong Government Funds under Grant No. 2016GDYSZDK036. And Zhuhai Laboratory of Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education.

References

[1] Zhao Yi'an. Solving the problem of store product sales forecast based on machine learning XGboost model [J]. Communication World, 2018 (11): 250-252.
[2] Huang Puwen. Diabetes analysis and prediction based on integrated learning [J]. Electronic Production, 2018 (22): 73-75.
[3] Huang Wei, Zheng Yinger, Deng Yuqiao. Research on Traffic Prediction Based on XGBoost Holiday Road Network[J]. Highway, 2018, 63 (12)
[4] Wu Xinghui, Zhou Yuping, Xing Haihua, Long Haixia. Application of Machine Learning Classification Algorithm in Diagnosis of Diabetes[J]. Computer Knowledge and Technology, 2018, 14(35): 177-178+195.
[5] Lin Lingling. Research on classification of hypertension classification rules based on C4.5 algorithm [D]. Taiyuan University of Technology, 2012.
[6] Liu Yu, Qiao Mu. Heart disease prediction based on clustering and XGboost algorithm [J]. Computer System Application, 2019, 28 (01): 228-232.
[7] Su Tianpei. Diabetes risk prediction based on XGBoost [J]. Science and Technology Vision, 2019 (02): 155-156.
[8] Han Qidi, Zhang Xiaotong, Shen Wei. Lithology identification technology based on gradient lifting decision tree (GBDT) algorithm[J]. Bulletin of Mineralogy, Geochemistry, 2018, 37(06):
[9] Li Yuchu. Text Classification Based on Support Vector Machine[J]. China New Technology and New Products, 2019(01): 23-24.

[10] Masconi Katya L, Matsha Tandi E, Erasmus Rajiv T, Kengne Andre P. PLoS one. Effect of model updating strategies on the performance of prevalent diabetes risk prediction models in a mixed-ancestry population of South Africa. [J]. 2019(2)

[11] Tova M. Bergsten, Robert Donnino, Binhuan Wang, Andrew Nicholson, Yixin Fang, Sundar Natarajan. Predicting adults likely to develop heart failure using readily available clinical information[J]. Preventive Medicine, 2019.

[12] Abd El-Wahab Ekram W, Shatat Hanan Z, Charl Fahmy. Adapting a Prediction Rule for Metabolic Syndrome Risk Assessment Suitable for Developing Countries. [J]. Journal of primary care & community health, 2019, 10.

[13] Kang Xinyi, Liang Yuanyuan, Wang Shiyu, Hua Tianqi, Cui Jiwen, Zhang Mingjin, Ding Yunjunyu, Chen Liping, Xiao Jing. Prediction model comparison for gestational diabetes mellitus with macrosomia based on risk factor investigation [J]. The journal of maternal-fetal & neonatal medicine: the official journal of the European Association of Perinatal Medicine, the Federation of Asia and Oceania Perinatal Societies, the International Society of Perinatal Obstetricians, 2019.

[14] Zhang Tao. Research on classification of imbalanced data and its application in disease diagnosis [J]. Journal of Yellow River Institute of Science and Technology, 2019, 21(05): 15-22.