International Journal of Advanced Smart Convergence Vol.10 No.2 73-78 (2021)
http://dx.doi.org/10.7236/IJASC.2021.10.2.73

IJASC 21-2-11

A Research on Accuracy Improvement of Diabetes Recognition Factors Based on XGBoost

Yongsub Shin* , Dai Yeol Yun**, Seok-Jae Moon** , Chi-gon Hwang***

**Graduate School of Smart Convergence Kwangwoon University, Seoul, Korea*
***Professor, Department of Plasma Bioscience and Display, KwangWoon University, Seoul 01897, Korea*
*** Professor, Department of Computer Science, Kwangwoon University, Seoul, Korea*
**** Professor, Department of Computer Engineering, Institute of Information Technology,*
Kwangwoon University, Seoul, Korea
iceboy724@kw.ac.kr, hibig10@kw.ac.kr, msj80386@kw.ac.kr, duck1052@kw.ac.kr

Abstract

Recently, the number of people who visit the hospital due to diabetes is increasing. According to the Korean Diabetes Association, it is statistically indicated that one in seven adults aged 30 years or older in Korea suffers from diabetes, and it is expected to be more if the pre-diabetes, fasting blood sugar disorders, are combined. In the last study, the validity of Triglyceride and Cholesterol associated with diabetes was confirmed and analyzed using Random Forest. Random Forest has a disadvantage that as the amount of data increases, it uses more memory and slows down the speed. Therefore, in this paper, we compared and analyzed Random Forest and XGBoost, focusing on improvement of learning speed and prevention of memory waste, which are mainly dealt with in machine learning. Using XGBoost, the problem of slowing down and wasting memory was solved, and the accuracy of the diabetes recognition factor was further increased.

Keywords: *Random Forest, XGBoost, Supervised Learning, Diabetes*

1. Introduction

The number of obese people is gradually increasing due to the recent westernized diet and lack of exercise. Accordingly, the amount of insulin required by the body gradually increases. Also, the ability to secrete insulin gradually decreases, causing problems with the pancreas, leading to diabetes [1]. The last study suggested that blood sugar, sex, BMI, triglyceride, and cholesterol levels are valid as diabetic classification factors using Random Forest, a type of bagging. It also classified diabetes by combining five factors [2]. The Random Forest used in this paper generates N Bootstrap samples and creates a decision tree for each training set. Calculate the average value of the created tree and analyze it around the reduction of variance. Decision trees have a common problem of data over fitting, but Random Forest creates a larger amount of trees to reduce over fitting [3]. For

Manuscript Received: April. 22, 2021 / Revised: April. 27, 2021 / Accepted: May. 4, 2021
Corresponding Author: duck1052@kw.ac.kr
Tel: +82-10-3959-9490, Fax: +82-2-940-5289
Professor, Department of Computer Engineering, Institute of Information Technology,, Kwangwoon University, Seoul, Republic of Korea

this reason, the random forest takes a long time to predict and uses a large amount of memory. In this paper, using XGBoost, we solved the problem of reducing prediction time and using a lot of memory, which are important in recent machine learning [4].

2. Related Work

2.1 Bagging and Boosting

The origin of the name diabetes is given to the fact that blood sugar increases and the glucose detected in the urine increases. Diabetes occurs when cells in the body become unable to respond to insulin or when the pancreas has a problem with secreting insulin. There is types 1 and 2 diabetes, [5]. Type 2 accounts for the majority of diabetes in Korea. It mainly affects the food supply and oil quality. Lack of exercise, lack of diet, and stress are significantly related to diabetes [6]. Women have a slightly higher incidence than men due to hormonal changes during pregnancy. Currently, it is assumed that the glycated hemoglobin, which is the criterion for diabetes, is 6.5% or more [5,7]. In this paper, we study to improve the performance and speed of diabetes recognition factors.

2.2 Bagging and Boosting

Boosting is one of the methods of generating multiple classifiers by manipulating initial sampling data like Bagging, but the biggest difference is the sequential boosting method. In the case of bagging, when the classifier learns, it is a technique that ends the learning without correlation and synthesizes the results. Each model is trained independently and in parallel. The Figure1 is a picture of bagging learning.

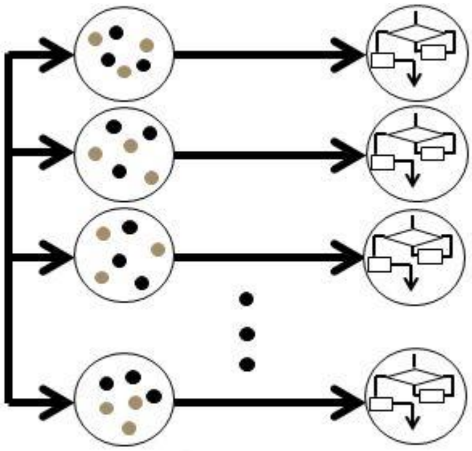

Figure 1. Bagging learning process

On the other hand, boosting is a method of assigning weights to incorrectly predicted data based on the learning result of the previous classifier and adjusting the sample weight to proceed with learning [8]. In general, there are fewer errors and better performance than bagging.

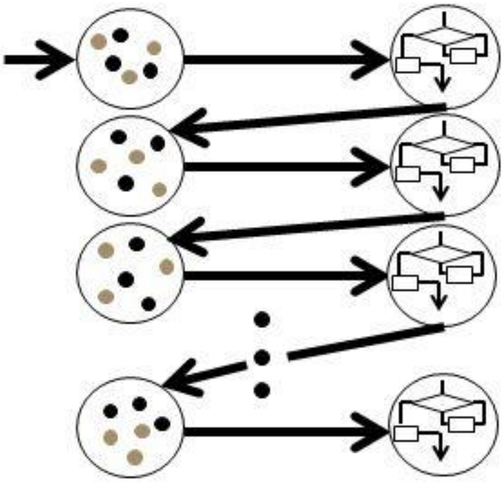

Figure 2. Boosting learning process

Unlike Bagging, as shown in the Figure2 above, Boosting does not discard the initially generated result values continuously, but gives weights sequentially. The big differences between Bagging and Boosting are as follows.

Table 1. Difference between bagging and boosting

	Bagging	Boosting
Principle of performance	Combined by sampling	Iteration of weight rebalancing
Purpose of performance	Reduce model variability	Improve model accuracy
Applied operation	Majority vote, average	Weighted linear combination
Early model	Bootstrap Model	Weak Classification
Final model	Bagging Model	Strong Classification
Classification performance	Excellent in the presence of missing values	Excellent for multiple data

Let's talk about table1. The big difference between bagging and boosting is that in the case of Bagging, several trees are created through sampling and then combined, whereas boosting is performed by adjusting the weights of missing values. In addition, bagging is calculated through majority vote after average values of various models, but boosting is performed through linear combination of weights. Finally, Bagging is excellent when there are missing values, and Boosting is excellent when there are many data. In the case of the boosting technique, the performance is excellent, but there is a problem that it is easily over-fitting. To solve this problem, XGBoost, which adds a regularization term to GBM, which is the basis of boosting, was used [9].

2.3 XGboost

XGBoost is one of the Boosting models. . It is based on GBM (Gradient Boosting Machine). However, in the case of GBM, it is slow and there is an issue of over fitting. Therefore, when GBM uses the powerful Ensemble

Boosting Algorithm, there is an issue to be considered. The model created to solve such a problem is XGboost. XGBoost is used to prevent over-fitting of the model, and contains regulations that can prevent over-fitting. It is also easier to visualize and understand compared to neural networks. Because the learning system is flexible, you can freely create an optimal model. The more resources (CPU, Memory) you have, the faster you can learn and predict. In the data set, more than 90% shows higher performance than GMB, and in fact, most of the results of using XGBoost in Kaggle occupy the top spot, so the performance is excellent. As it is based on CART (Classification and Regression Tree), it is excellent for both classification and regression. Early Stopping is provided, and the maximum value is assigned to the Gradient Descent, a characteristic of Ensemble [9].

3. Experiment and evaluation

3.1 Accuracy trend according to variable importance and number of data

Variable importance is a number used to determine which variable has the strongest predictive power among the many variables used after learning the data. It is important to identify and classify variable importance as it can affect the performance of the model.

Table2. Accuracy value according to data size by type

	BMI	Triglyceride	FBS(Fasting Blood Sugar)	Cholesterol	Gender
Attribute Importance	0.5907	0.1624	0.1017	0.0977	0.0473

First, the variables used in the experiment are the five variables that were verified in the previous paper, and BMI, Triglyceride, FBS, Cholesterol, and Gender are used [2]. Boosting adjusts weights and classifies them, so the importance of each variable is important. The importance of the variables of the above five variables is as follows. The importance of 5 variables is shown in table2. A graph showing it is shown in Figure 3.

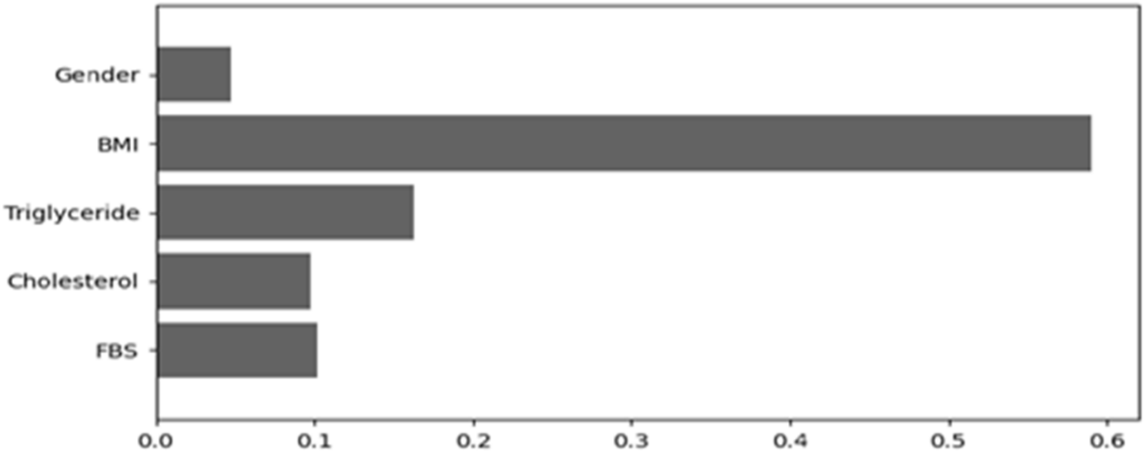

Figure 3. Attribute importance

First of all, the most important factor was 0.5907 points, BMI. BMI stands for body mass index. The next important thing was the triglyceride, which is 0.1624. Subsequently, FBS, Cholesterol, and Gender were respectively 0.1017, 0.0.0977 and 0.0473.In addition, when classification using Random Forest and classification using XGBoost, the accuracy trend according to the number of data is as follows.

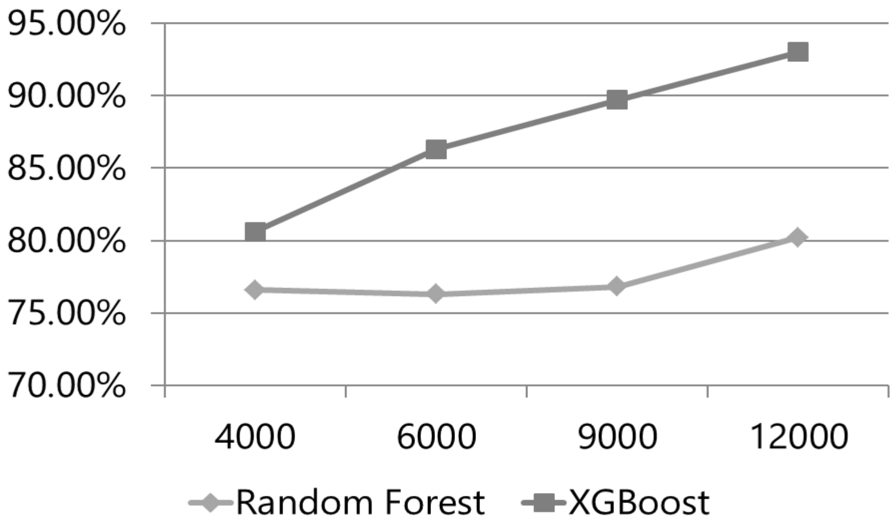

Figure 4. Accuracy according to the number of random forest and XGBoost data

Table3. Accuracy value according to data size by type

Size Algorithm	4000	6000	9000	12000
Random Forest	76.6%	76.3%	76.8%	80.2%
XGBoost	80.6%	86.3%	89.7%	93.0%

As can be seen from Table 3, XGBoost showed better overall performance than Random Forest. When the number of data was 4000, they were 76.6% and 80.06%, respectively. As the number of data increased, it showed excellent performance, and even for 12,000 cases, XGBoost showed excellent performance at 80.2% and 93.0%. In conclusion, regardless of the amount of data, XGBoost performed better than Random Forest in all cases.

3.2 Learning Speed and Accuracy by Algorithm

The previously used bagging method, Random Forest, has better performance than Decision Tree and can reduce over fitting. However, there was a problem that it takes a long time to prevent over fitting of a random forest because it is a method of creating many trees. Therefore, the boosting method, which is a method that is generally faster and has superior performance compared to bagging, was used. In addition, XGBoost was used to reduce over fitting that may occur in boosting. The results of comparing Decision Tree, Random Forest, and XGboost were as follows.

Table4. Accuracy value according to data size by type

Algorithm	Learning Speed	Accuracy
Decision Tree	0.356988754465655	75.2
Random Forest	1.0761089324951172	80.2
XGBoost	0.6014680862426758	93.07

The contents of table4 are as follows. In the case of Decision Tree, the speed was the fastest among the three. However, the accuracy was much lower than the other two techniques. On the other hand, the accuracy of the Random Forest using the Bagging method is higher than that of the Decision Tree. However, it can be seen that the speed of the random forest is slow because of the over fitting regulation technique and the need to create a large number of trees. Lastly, it can be seen that the accuracy and learning speed increase because XGBoost continuously learns by weighting the missing values.

4. Conclusion

In order to improve the accuracy and learning speed, which are becoming increasingly important today, the XGBoost algorithm, which is an improved model than the existing Random Forest, was used. The prediction speed was improved and the use of a lot of memory was prevented, and the prediction performance was improved compared to the Random Forest. Since the speed has been greatly improved compared to the existing learning, even when using more data, it can be classified quickly. In addition, if in the future, not only diabetes, but also various other diseases, and the elements are learned, new factors affecting the disease can be identified and verified in advance through the learned data. Through such new elements, elements that have been easily thought of and overlooked in the past can be carefully considered. Through the factors found in this way, it can be controlled by good behavior and intake of the factors. In addition, it will be useful as it is of great help in preventing the disease by managing the elements of the disease after grasping the relationship between each factor through the verified data.

References

[1] Krishnasamy, S, Abell, T. L. (2018). Diabetic gastroparesis: principles and current trends in management. Diabetes Therapy, 9(1), 1-42., https://doi.org/10.6084/m9.figshare.6391592.

[2] sub Shin, Yong, Namju Lee, and Chigon Hwang. "A research on the key factors for classification of diabetes based on random forest." International Journal of Internet, Broadcasting and Communication 12.3: 102-107, https://doi.org/10.7236/IJIBC.2020.12.3.102

[3] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32. , https://doi.org/10.1023/A:1010933404324

[4] Sasaki, Yutaka. "The truth of the f-measure. 2007." (2007): 16.

[5] The Institute of Internet, Broadcasting and Communication, Submission of manuscript. *http://www.iibc.kr.*

[6] Minjin Lee, & Sang soo Kim. (2017). Obesity management in diabetics. *Journal of Korean Diabetes, 18*(4)., https://doi.org/10.1053/beem.1999.0017

[7] Mitchell, T. M. (1999). Machine learning and data mining. *Communications of the ACM*, *42*(11), 30-36.

[8] R, SAS, Data mining using MS-SQL / freeaca / Jeong jin Lee

[9] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016, https://doi.org/10.1145/2939672.2939785