A High Accuracy Integrated Bagging-Fuzzy-GBDT Prediction Algorithm for Heart Disease Diagnosis

Xiaoming Yuan[1], Xue Wang[1], Jianchao Han[1], Jiemin Liu[1], Haiyan Chen[1], Kuan Zhang[2], and Qiang Ye[3]
[1]Qinhuangdao Branch Campus, Northeastern University, Qinhuangdao 066004, China
[2]Department of Electrical and Computer Engineering, University of Nebraska-Lincoln, Omaha, NE 68182, USA
[3]Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada
E-mails: {yuanxiaoming, liujiemin, chenhaiyan}@neuq.edu.cn, {wangxue, jianchaohan}@stumail.neu.edu.cn,
kuan.zhang@unl.edu, q6ye@uwaterloo.ca

***Abstract*—Associated with high morbidity and mortality, heart disease has become a severe threat to peoples health throughout the world. The recent development of Internet of Things (IoT) and machine learning in e-healthcare have contributed to the monitoring, prediction and diagnosis of heart disease. Particularly, the heart disease prediction can effectively facilitate disease prevention, diagnosis and timely treatment. However, traditional prediction models are weak in accuracy and generalization. In this paper, we propose a high accuracy integrated prediction algorithm for heart disease diagnosis. The fuzzy logic and Bootstrap Aggregating (Bagging) algorithm based on Gradient Boosting Decision Tree (GBDT) algorithm are combined to process heart disease data and generate multiple weak classifiers. At first, we integrate the fuzzy logic with GBDT to reduce the complexity of data. Moreover, we develop the Fuzzy-GBDT model integrated Bagging algorithm to avoid the interference of sensitive points and achieve partial parallelism. The simulation results show the proposed Fuzzy-Bagging-GBDT algorithm improves the accuracy and recall of heart disease prediction compared with GBDT.**

I. Introduction

The increasing incidence of chronic disease severely affects people's life, costs huge economic loss, and seriously hinders the development of national medical and health services [1]. The high morbidity and mortality of heart disease must be attached with great importance in the field of chronic diseases. At present, this problem cannot be completely solved which can only be alleviated by early detection and prevention.

The development of Internet of Things (IoT) technology advances the convenient monitoring of patients vital signs no matter where they are and what they are doing. The IoT system consists of few sensors to collect data, a processor to process data and wireless links to transmit data [2]. Using IoT technology, Li et al. [3] proposed a heart disease monitoring system which can send patients' physical signs to remote medical applications in real time. In [4], patients can collect their vital signs and send them to doctors or hospitals. Currently, the explosion of disease-related data stimulates the need of high efficient data processing technologies for disease prediction and diagnosis. Machine learning has become an important technical supporting for IoT health monitoring owing to its high accuracy and efficiency in models training [5]. With the combination of machine learning technology and IoT in e-healthcare [6], the chronic diseases prevention, monitoring and management are getting more and more effective. Although most heart diseases cannot be cured completely, we can provide health guidance and intervention through disease prediction before attacking according to patients' real-time status.

But challenges in predicting heart disease should not be neglected. On one hand, heart disease prediction algorithms should ensure the reliability of the results. On the other hand, the generalization of the algorithms should be as high as possible due to constantly update of patients' vital signs. Thus the most challenging issues in heart disease prediction are accuracy and generalization ability of prediction algorithm.

There are numerous algorithms designed for heart disease prediction. Such as decision tree [7], [8]. Soni et al. [8] found that decision tree was the most accurate method in heart disease prediction among the comparison of neural network, naive bayes and decision tree algorithms. But the prediction model based on decision tree algorithm is difficult to process continuous variables. Moreover, authors [9], [10] developed a highly efficient Gradient Boosting Decision Tree (GBDT) as a classification prediction algorithm based on decision tree. GBDT can deal with various types of data including continuous and discrete values.

However, most of current prediction algorithms or models are not excellent enough in over-fitting avoidance and low generalization problem. Bootstrap Aggregating (Bagging) algorithm can reduce the variance of base estimator by introducing randomness in the process of constructing the model. Thus over-fitting can be effectively solved by Bagging algorithm. Khosla et al. [11] considered that bagged algorithms outperform the simple model with a higher ranking quality (AUC). Moreover, fuzzy logic is usually used to deal with uncertain things, which is suitable for processing constantly updated data and further increase the generalization of prediction algorithms. In [12], authors developed a novel fuzzy-learning algorithm to achieve the optimal network throughput even in face of uncertain information. Khatibi et al. [13] input information's vagueness through fuzzy logic and provides more generalization disease prediction results. But the aforementioned works solve the over-fitting and low generalization problems in one aspect instead of addressing the problems simultaneously.

In this paper, we propose a Bagging-Fuzzy-GBDT algorithm heart disease prediction model for heart disease diagnosis, which improves the accuracy and generalization compared with traditional algorithms. Base on GBDT, we introduce fuzzy logic to process heart disease data into different intervals by the membership function, which improves the adaptability for data with large and different ranges of values. Due to the constantly updating and changing characteristics of data, we need on-line learning to update the model in time. Bagging algorithm can save learning time by processing data in parallel, which ensembles the weak learners of Fuzzy-GBDT to get the final results. Our contributions are concluded as follows:

- Firstly, we consider the characteristics of heart disease data and combine fuzzy logic approach with GBDT to convert one single attribute value into three levels. The Fuzzy-GBDT method improves the generalization ability of the GBDT prediction algorithm.
- Moreover, we generate additional data for training by using the original data, and integrate Bagging algorithm with Fuzzy-GBDT to reduce the variance of prediction model.
- Finally, we conduct performance evaluation and the results prove the proposed Bagging-Fuzzy-GBDT model is high accuracy and generalization for heart disease prediction.

The remainder of the paper is organized as follows: Section II introduces the GBDT algorithm and related heart disease data set. Section III presents the details of the proposed integrated Bagging-Fuzzy-GBDT prediction algorithm. We discuss and analyze the simulation results in section IV. Finally, section V concludes the paper.

II. Preliminary

In this section, we make a brief overview about GBDT algorithm to make preparation for the following improved algorithm. In addition, we describe the data set of heart disease.

A. GBDT Algorithm

GBDT is a member of ensemble learning Boosting family. It is an ensemble model of the decision tree, which are trained in sequence. In each iteration, GBDT learns the decision tree by fitting the negative gradient (also known as residual error). The expression is shown as Eq. 1. Where $a, b, c...$ are the coefficient of each weak classifier.

$$F\left(x\right) = aF_0\left(x\right) + bF_1\left(x\right) + cF_2\left(x\right) + \cdots + \rho F_\rho\left(x\right) \quad (1)$$

In the training process of GBDT algorithm, each iteration creates a new decision tree in the direction of reducing residuals. GBDT algorithm uses the negative gradient value of the loss function to approximate the residual. The loss function is a expression to measure loss and error, which reflects the credibility of the model. The smaller loss function is, the higher accuracy of model is. The training process of GBDT algorithm is shown in Fig. 1.

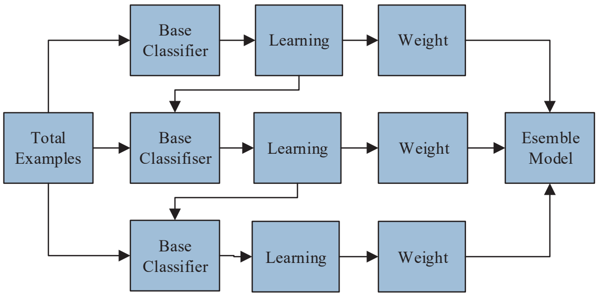

Fig. 1: GBDT training progress.

TABLE I: Data set description

Number	Attribute	Description
1	Age	Age in year
2	Sex	Sex of subject
3	Cp	Chest pain
4	Trestbps	Resting blood pressure
5	Chol	Serum cholesterol
6	Fbs	Fasting blooding sugar
7	Restecg	Resting electrocardiographic result
8	Thalach	Maximum heart rate achieved
9	Exang	Exercise induced angina
10	Oldpeak	ST depression induced by exercise relative to rest
11	Slope	Slope of peak exercise ST segment
12	Ca	Number major vessels colored by fluoroscopy
13	Thal	Defect type
14	Num	Heart Disease

B. Data Set

The data set in this paper is from UCI's open source heart disease data set, which from four different medical institutions. This database has a total of 303 datas, containing 14 important parameters such as age, heart rate, cholesterol and blood pressure were selected. The data set description is shown in Table I.

According to the data of the heart disease training set, the methods of preprocessing the data include data normalization, data transformation, filling missing values, deleting outlier and null values. In this paper, the data set is divided into a testing set and a training set according to a ratio of 3 to 7.

III. The Proposed Bagging-Fuzzy-GBDT Prediction Algorithm

In order to reduce the complexity of heart disease data and strengthen generalization ability, we add fuzzy logic to GBDT algorithm. Moreover, we introduce Bagging algorithm into Fuzzy-GBDT algorithm to eliminate interference by multiple self-sampling, which can avoid over-fitting. We proposed the Bagging-Fuzzy-GBDT algorithm increases accuracy, recall and AUC for heart disease prediction.

A. Fuzzy-GBDT Algorithm

In reality, the patients with large differences in values under the same attribute always is diagnosed as the same result. The four attributes: age, tresthps, chol and thalach whose values with large fluctuations affect the classification accuracy when using the GBDT algorithm for heart disease prediction. Fuzzy logic uses a membership function to describe the difference between the data. Fuzzy sets are hierarchical structures that describe the degree to which data objects belong to attributes. The value assigned to each object ranges from 0 to 1. The closer the result to 1 is, the higher the degree of membership. So the complexity of data set is greatly reduced.

Fuzzy logic can theoretically reduce the complexity of the data, improving the accuracy of prediction model, increasing the generalization ability. The choice of membership function plays a crucial role in a certain kind of problem determined by fuzzy description. We choose to apply the triangular membership function. The rules are as follows:

Definition: Let the fuzzy number on the domain U be M. If the membership function of M is μ_M, it can make $\mu \rightarrow [0,1]$, then m is a triangular fuzzy number, which is a triangular membership function. ω can be converted into three membership degrees $(\omega_1, \omega_2, \omega_3)$, where max represents the maximum value of the data interval and min represents the minimum value of the data interval. Let $m' = \frac{\max - \min}{3-2\alpha}(0 < \alpha < 1)$, α represents the degree of coincidence of subintervals, $\varphi = \min + (1-\alpha)m'$, then

$$\omega_1 = \begin{cases} 0, \quad x > \min + m' \\ \frac{\min + m' - x}{m'}, \min <= x <= \min + m' \\ 1, x < \min \end{cases}$$
$$\omega_2 = \begin{cases} 0, \quad x < \varphi \quad OR \quad x > \min + (2-\alpha)m' \\ \frac{x-\varphi}{m'/2}, \quad \varphi \leq x \leq \varphi + m'/2 \\ \frac{\min + (2-\alpha)m' - x}{m'/2}, \quad \varphi + m'/2 < x < \varphi + m' \end{cases}$$
$$\omega_3 = \begin{cases} 0, \quad x < \min + (2-2\alpha)m' \\ \frac{x-m'}{m'}, \quad \min + (2-2\alpha)m' \leq x \leq \max \\ 1, \quad x > \max \end{cases} \tag{2}$$

B. Data fuzzification

According to the regularity of the data samples, the attribute age, tresthps, chol and thalach are fuzzed by above formula, the remaining nine values of attributes ranges from 0 to 10, it is not necessary to be fuzzed. Chol, tresthps, thalach and age are divided into three levels, so that a single attribute value ω is converted into three membership degrees $(\omega_1, \omega_2, \omega_3)$. The four attributes α values take 0.2, 0.3, 0.2, 0.4 due to different ranges. Chol is divided into 150-250 for poor-chol, 230-330 for average-chol, 310-410 for good-chol. Trestbps is divided into 90-140 for poor-trestbps, 125-175 for average-trestbps, 160-210 for good-trestbps. Thalach is divided into 80-130 for poor-thalach, 120-170 for average-thalach, 160-210 for good-thalach. Age is divided into 25-50 for poor-age, 40-65 for average-age, 55-80 for good-age. The corresponding membership function graph is shown as Fig. 2. We combine the fuzzified data with GBDT, and the process is as follows:

1) Let α_i, β_i, γ_i be the representative values of the three intervals of ith attribute A_i, and map them to the respective membership functions $(\omega_{i1}, \omega_{i2}, \omega_{i3})$, respectively. Let $B_i = \max(\omega_{i1}, \omega_{i2}, \omega_{i3}) \rightarrow (\alpha_i, \beta_i, \gamma_i)$, Let $A_i = B_i$;

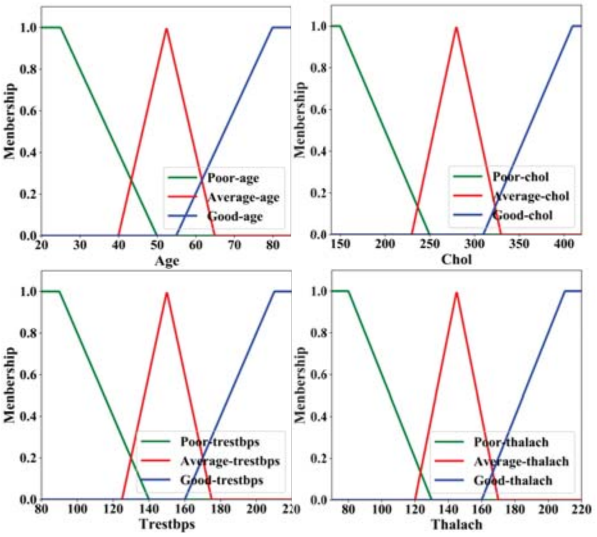

Fig. 2: The fuzzy membership of different attributes.

2) When the fuzzified data is combined with GBDT, the data input format of every weak Fuzzy-GBDT learner becomes T=$\{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$
(i) Initialize weak Fuzzy-GBDT learner

$$F_0(X) = 0.5 * \log\left(\frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n}(1-y_i)}\right) \tag{3}$$

(ii) From 1 To t do ($t = 1, 2, ..., T, T$ is the maximum number of iterations):

(a). Calculate negative gradient $\tilde{y_i}$, we adopt cross entropy loss function as L
$L(y_i, F_t(x)) = -\{y_i \ln p_i + (1-y_i)\ln(1-p_i)\}$
where, $p_i = \frac{1}{1+e^{-F_t(x_i)}}$

$$\begin{aligned} \tilde{y}_i &= -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{t-1}(x)} \\ &= y_i - \frac{1}{1+e^{(-F_{t-1}(x_i))}} \end{aligned} \tag{4}$$

(b). Use $(x_i, \tilde{y}_i)(i = 1, 2, ...n)$, fitting a CART regression tree and get $t-th$ regression tree. Its corresponding leaf node area is $R_{tj}, j = 1, 2, ..., J$, J is the number of leaf nodes of the tree t.

(c). Calculate the best fit c_{tj} for the leaf $j = 1, 2, 3, ..., J$,

$$c_{tj} = \underbrace{\arg min}_{c} \sum_{x_i \in R_{tj}} \log(1+\exp(-y_i(f_{t-1}(x_i)+c))) \tag{5}$$

(d). Update learner, I is indicator function, $I \in [0,1]$.

$$f_t(x) = f_{t-1}(x) + \sum_{j=1}^{J} c_{tj} I(x \in R_{tj}) \tag{6}$$

(iii). Get strong learner $f(x)$.

$$f(x) = f_T(x) = f_0(x) + \sum_{t=1}^{T}\sum_{j=1}^{J} c_{tj} I(x \in R_{tj}) \tag{7}$$

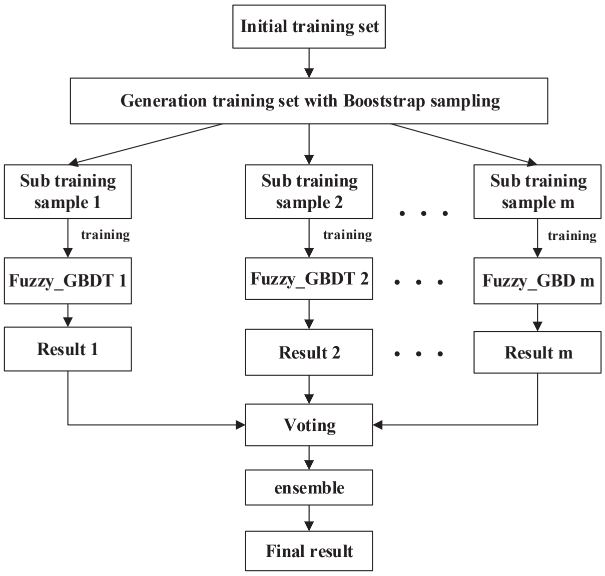

Fig. 3: Fuzzy-GBDT integrated Bagging algorithm.

C. Bagging-Fuzzy-GBDT Algorithm

Though the Fuzzy-GBDT algorithm strengthens the generalization ability, GBDT itself is sensitive to data and inefficient to process a large amount of medical data. In order to ensure the accuracy of the Fuzzy-GBDT model and save learning time, we introduce Bagging algorithm into Fuzzy-GBDT model. It can reduce the variance of the classifier results and avoid over-fitting. In addition, it improves the stability and accuracy of the model. The integrated algorithm not only enhances the stability and efficiency but also implements the parallelism in data processing. Bagging can eliminate interference from a single sensitive point by multiple self-sampling and avoid over-fitting.

The proposed integrated Bagging-Fuzzy-GBDT algorithm is shown in Fig. 3. The algorithm can also be operated in the following steps:

1) Suppose a training set D is given, the size is N;
2) The m subsets $D_i(i = 1, 2, ..., m)$ are with replacement from the training set and the size is $n'(n' < N)$;
3) Fuzzy-GBDT classifier is created and trained with above sampled subset D_i;
4) Repeat step 3) to create m Fuzzy-GBDT classifiers;
5) The strong classifier Bagging-Fuzzy-GBDT can be obtained from m weak learners by taking the voting methods.

The heart disease prediction model based on Bagging-Fuzzy-GBDT algorithm is composed of Bagging algorithm and fuzzy logic combined with GBDT algorithm. The algorithm utilizes the advantage of GBDT is more robust than Decision Tree in over-fitting. It also adds fuzzy logic to fuzzed heart disease data from UCI, which reduces the complexity of data, and makes the data more standard available. At the same time, the addition of Bagging method to generate Fuzzy-GBDT which can generate several weak classifiers in parallel. It improves the adaptability of the algorithm to data with small fluctuation. Using IoT and machine learning technologies to obtain patients' vital signs and predict heart disease, which is conducive to timely treatment of patients.

IV. Performance Evaluation

We utilize data of heart disease from UCI as training set and testing set to verify the performance of the algorithm. The data set description is shown as Table 1. In this paper, all parameters of Bagging-Fuzzy-GBDT algorithm are set to: the number of sampling for Bagging is 20, the learning rate is 0.1, the number of iteration is 7, the maximum depth of each estimator is 3, the minimum number of samples used to split the nodes is 2. Indicators such as accuracy, precision and recall are often used to evaluate the performance of classification algorithms. Accuracy is defined as the ratio of the correctly classified samples to the total given number of samples. The expression of accuracy can be shown as Eq. 8.

$$\text{Accuracy} = \frac{TS + TN}{TS + TN + FS + FN} \tag{8}$$

where TS represents the number of people who are predicted to get sick and actually sick; FS represents the number of people who are predicted in illness but actually not sick; FN means the number of people who are predicted to have no disease, but actually have heart disease; TN means the number of people in good health with no disease both in prediction and the actual results.

Precision is the ratio of correctly detected sick samples to all detected sick samples.

$$\text{Precision} = \frac{TS}{TS + FS} \tag{9}$$

Recall is the ratio of correctly detected sick samples to all actual sick samples.

$$\text{Recall} = \frac{TS}{TS + FN} \tag{10}$$

In order to fully verify the rationality and effectiveness of the prediction model using Bagging-Fuzzy-GBDT algorithm adopted in this paper, the GBDT model and the Bagging-GBDT model, the Fuzzy-GBDT model and the Decision Tree model are constructed. The training samples and testing samples are applied to the above five models respectively. The histogram is shown as Fig. 4. It can be clearly seen from the histogram that the Bagging-Fuzzy-GBDT model has a significant increase in each evaluation index. In particular, Bagging-Fuzzy-GBDT has an absolute advantage in accuracy and precision. Prediction model with high accuracy are used for real-time heart disease diagnosis with greater reliability.

According to the results of the five models, the corresponding Receiver Operation Characteristic (ROC) curve and the Area Under The Curve (AUC) indicator evaluation table are drawn as Fig. 5 and Fig. 6. ROC is a receiver operating characteristic curve, false positive ratio is horizontal axes, true positive ratio is vertical axes. AUC is the area under the ROC curve and the encirclement of the coordinate axis. It is known from the ROC curve and the AUC area that the Bagging-Fuzzy-GBDT algorithm has the best performance. In this experiment, the Decision Tree and GBDT have the comparable performance. The AUC value was promoted after we combine the fuzzy logic and Bagging algorithm. The

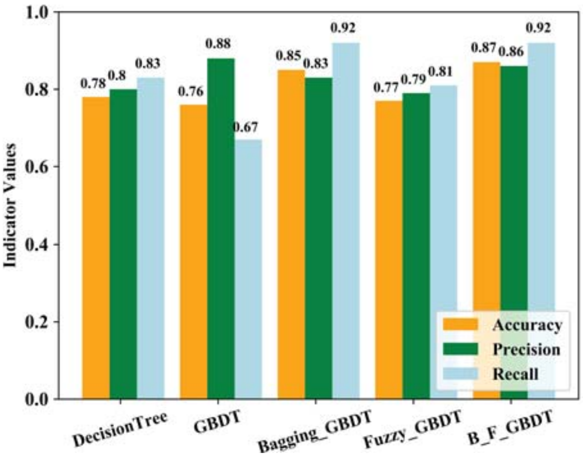

Fig. 4: The indicator values of different models.

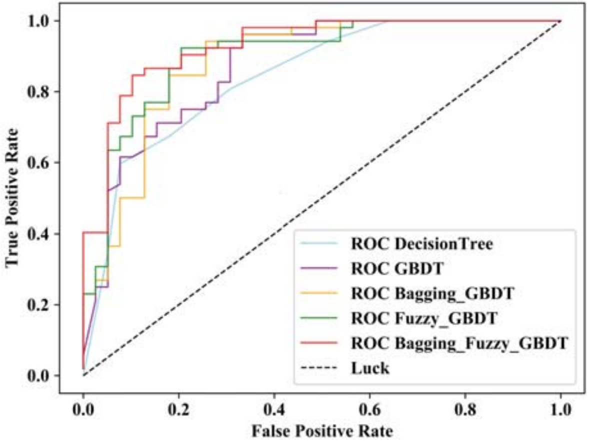

Fig. 5: The ROC curves of different models.

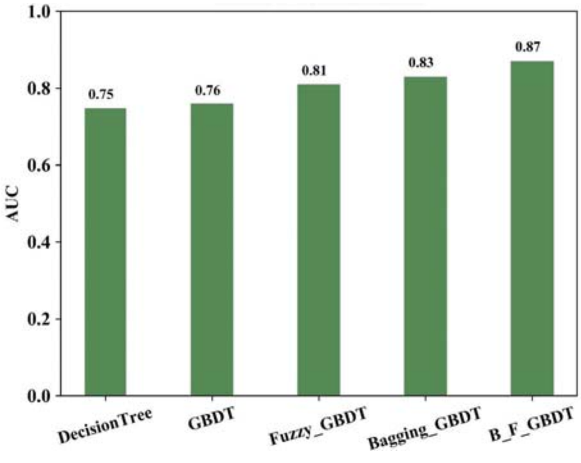

Fig. 6: The AUC of different models.

Bagging-GBDT was a little effective than the Fuzzy-GBDT 0.2 on AUC value. It indicates that the fuzzy logic and Bagging algorithm are effective on the accuracy of model. We combine both fuzzy logic and Bagging algorithm with GBDT, the performance improved dramatically which has the AUC value of 0.87. From above, we can conclude that the fuzzy logic and Bagging algorithm are work on the accuracy and generalization ability.

V. Conclusions

In this paper, we have proposed a high accuracy integrated Bagging-Fuzzy-GBDT prediction algorithm for heart disease diagnosis. Specifically, we introduced fuzzy logic into GBDT to increase the generalization ability. Furthermore, the Bagging algorithm was integrated with the Fuzzy-GBDT algorithm to avoid over-fitting and achieve the data parallelization in the training process to save training time. Finally, the simulation results proved the proposed Bagging-Fuzzy-GBDT algorithm has an obvious improvement in accuracy, precision, AUC and other indicators compared with traditional algorithms. The Bagging-Fuzzy-GBDT algorithm combined with IoT can better monitor the health of patients and promote the development of IoT and machine learning technologies in medical field. In the future, we will optimize the complexity and training time of the proposed algorithm and improve the performance of heart disease prediction.

Acknowledgment

This work was supported by the Fundamental Research Funds for the Central Universities (N182303032).

References

[1] Bragg F., Holmes M. V., Iona A., et al. Association between Diabetes and Cause-Specific Mortality in Rural and Urban Areas of China. *JAMA-Journal of the American Medical Association*, 2017, 317(3):280-289.

[2] Hallfors N. G., Alhawari M., Jaoude M. Abi., et al. Nylon ECG Sensors for Wearable IoT Healthcare-nanomaterial and SoC Interface. *Analog Integrated Circuits and Signal Processing*, 2018, 96(2):253-260.

[3] Li C., Hu X., Zhang L. The IoT-based Heart Disease Monitoring System for Pervasive Healthcare Service. *Procedia computer science*, 2017, (112):2328-2334.

[4] Krishna C. S., Sasikala T. Healthcare Monitoring System Based on IoT Using AMQP Protocol. *Proceeding of International Conference on Computer Networks and Communication Technologies*, India, Apr. 26-27 2018:305-319

[5] Sun P., Li J., Bhuiyan M. Z. A., et al. Modeling and Clustering Attacker Activities in IoT through Machine Learning Techniques. *Information Sciences*, 2019, 479:456-471.

[6] Das A., Rad P., Choo KKR., et al. Distributed Machine Learning Cloud Teleophthalmology IoT for Predicting AMD Disease Progression. *Future Generation Computer Systems*, 2019, 93:486-498.

[7] Tayefi M., Tajfard M., Saffar S., et al. Hs-CRP is Strongly Associated with Coronary Heart Disease (CHD): A Data Mining Approach Using Decision Tree Algorithm. *Computer Methods and Programs in Biomedicine*, 2017, 141:105-109.

[8] Soni J., Ansari U., Sharma D., et al. Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. *International Journal of Computer Applications*, 2011, 17(8):43-48.

[9] Ke G., Meng Q., Finley T., et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Proceedings of 31st Advances in Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, Dec. 4-9, 2017:3146-3154.

[10] Zhao L., Ni L., Hu S., et al. InPrivate Digging: Enabling Tree-based Distributed Data Mining with Differential Privacy. *Proceedings of 2018 IEEE International Conference on Computer Communications (INFOCOM)*, Honolulu, HI, Apr. 16-19, 2018:2087-2095.

[11] Khosla R., Fahmy S., Hu Y. C., et al. Predicting Prefix Availability in the Internet. *Proceedings of 2010 IEEE International Conference on Computer Communications (INFOCOM)*, San Diego, CA, Mar. 14-19, 2010:1-5.

[12] Fan C., Li B., Zhang Y., et al. Robust Dynamic Spectrum Access in Uncertain Channels: A Fuzzy Payoffs Game Approach. in *Proceedings of 2017 IEEE Global Communications Conference (GLOBECOM)*, Singapore, Singapore, Dec. 4-8, 2017:1-5.

[13] Khatibi V., Montazer G. A. A Fuzzy-evidential Hybrid Inference Engine for Coronary Heart Disease Risk Assessment. *Expert Systems with Applications*, 2010, 37(12):8536-8542.