Risk prediction of type II diabetes based on random forest model

Weifeng Xu, Jianxin Zhang, Qiang Zhang*, Xiaopeng Wei
Key Lab of Advanced Design and Intelligent Computing, Ministry of Education,
Dalian University, Dalian, P. R. China
*Corresponding author: zhangq26@126.com

***Abstract*—In recent years, type II diabetes has become a serious disease that threaten the health and mind of human. Efficient predictive modeling is required for medical researchers and practitioners. This study proposes a type II diabetes prediction model based on random forest which aims at analyzing some readily available indicators (age, weight, waist, hip, etc.) effects on diabetes and discovering some rules on given data. The method can significantly reduce the risk of disease through digging out a clear and understandable model for type II diabetes from a medical database. Random forest algorithm uses multiple decision trees to train the samples, and integrates weight of each tree to get the final results. The validation results at school of medicine, University of Virginia shows that the random forest algorithm can greatly reduce the problem of over-fitting of the single decision tree, and it can effectively predict the impact of these readily available indicators on the risk of diabetes. Additionally, we get a better prediction accuracy using random forest than using the naive Bayes algorithm, ID3 algorithm and AdaBoost algorithm.**

Keywords—data mining; prediction model; random forest; type II diabetes

I. INTRODUCTION

Diabetes is the most common disease nowadays in most age groups. It is a disease in which the body does not produce or properly use insulin [1]. Type II diabetes, which accounts for about 95% of diabetic patients, is one of the focuses of chronic disease prevention and health management. In order to find hidden relationships between type II diabetes and symptoms, we can use data mining technology to mining medical data from that covert and unknown knowledge and extract useful patterns. In recent years, using predictive classification in medical diagnosis has received a strong boost owing to earnest research activity in this field in recent times. And majority of papers published deal with the goal of improving accuracy. For example, Karol Grudzinski used the KNN algorithm (k=22) to obtain the highest accuracy of the model is 75.5% [2]. The neural network achieved an accuracy of 75.4% whereas the Bayesian approach achieved 79.5% accuracy in reference [3]. Allahverdi proposed a hybrid neural network (artificial neural networks (ANN) and fuzzy neural network (FNN) model), the precision of this method to get the value of 84.24% [4]. Humar Kahraman used mixed model (k-means algorithm and C4.5 algorithm) to obtain an accuracy of 84.5% [5]. Although the accuracy of the model has been improving, it is obvious that the research methods and the improved algorithms are the indisputable single classifier. However, the ensemble classifier is better than the single classifier in many cases. Additionally, the above models are established on the Pima Indian diabetes datasets which are obtained that pregnant women over the age of 21 and that dramatically reduces the expansibility of the model [6]. Also, these predictors in the model are not directly visible and need to be measured by certain medical equipment, which increases the cost of the patient's diagnosis. Therefore, we re-selected the datasets of the University of Virginia, because the ratio of male and female are even in this datasets [7]. Besides, the datasets contain a lot of readily available indicators (such as age, height, waist and hip), using these indicators to predict diabetes that can greatly reduce the cost of diabetes risk prediction, so we choose these external readily available indicators to judge the impact on diabetes and try to prevent it in the bud. Random forest is an ensemble classifier composed of multiple decision trees, which has the advantages of high accuracy and good robustness [8]. Therefore, the present study uses random forest as the basic classifier.

The rest of the paper is organized as follows: random forest algorithm is described in Section 2. Data pre-processing is given in Section 3 and experiment design is presented in Section 4. The results and model evaluation are discussed in Section 5. Finally, Section 6 presents the conclusions.

II. RANDOM FOREST

The random forest algorithm, proposed by Dr. Breiman in 2001, has been extremely successful as a general purpose classification and regression method. The approach, which combines several randomized decision trees and aggregates their predictions by averaging, has shown excellent performance in settings where the number of variables is much larger than the number of observations[9]. It is an algorithm based on statistical learning theory, which uses Bootstrap randomized re-sampling way to extract multiple versions of the sample sets from the original training datasets, then building a decision tree model for each sample set, the final combined all the results of the decision trees to predict the results of classification by the established voting mechanism. The detailed process is shown in Fig.1.

III. DATA PRE-PROCESSING

The data should be carefully collected, integrated and prepared for analysis. In this study, we applied the techniques of data pre-processing to improve the quality of the mining

978-1-5090-5434-3

results and the efficiency of the mining process. The raw datasets is provided by School of medicine, University of Virginia which has 403 testers, and each tester consists of 19 features, including age, sex, cholesterol, hemoglobin, waist, hip, etc. In this datasets, we can judge whether or not the tester has diabetes according to the Glycosylated Hemoglobin value. It is normal when the Glycosylated Hemoglobin value less than 7.0. However, when it exceeds 7.0, we think it is abnormal [10]. So we can convert the Hemoglobin Glycosylated value to a Boolean type and set it as the class label. Additionally, there are some features of missing values in different degrees, including Total Cholesterol, High Density Lipoprotein, Cholesterol/HDL Ratio, height, weight, frame, waist, hip, etc. However, our study are mainly used some readily available indicators to predict the risk of diabetes, we can conduct dimensionality reduction first. The key benefit of dimensionality reduction is to improve the performance of the algorithm, because the dimensionality reduction can remove irrelevant features and reduce the noise [11]. So all unrelated features are removed, including Total Cholesterol, High Density Lipoprotein, Stabilized Glucose, High Density Lipoprotein, the Cholesterol/HDL Ratio and the First Systolic Blood Pressure. Also, the Weight loss 1 value, the Frame loss 11 values, the Waist and the Hip, respectively loss 2 values, so we have eliminated the missing values. After removing all the above features and values, only 373 instances and 10 features remain from the data in our study. Statistical results are shown in TABLE I.

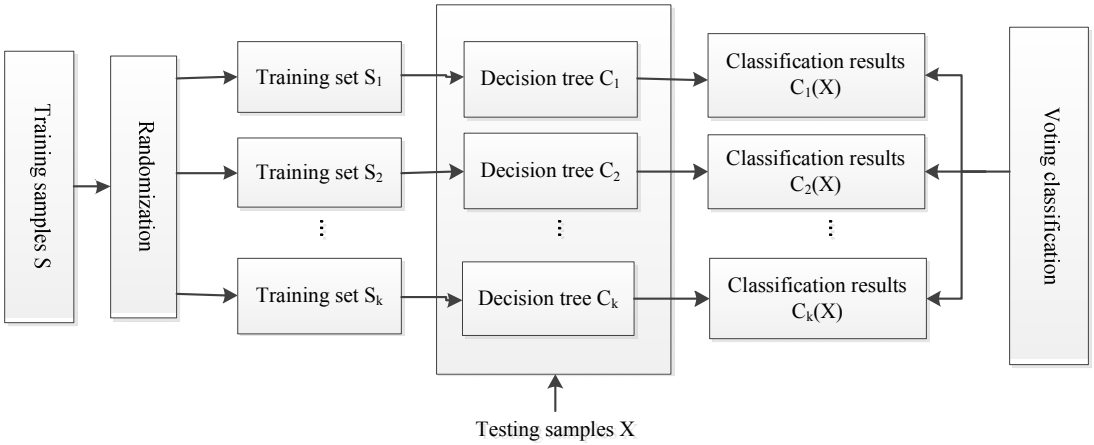

Fig. 1. Flow chart of the random forest algorithm

TABLE I. STATISTICS DATA AFTER PRELIMINARY ANALYSIS

Name	Type	Miss	Min	Max	Average
ID	Polynomial	0			
Location	Binomial	0	Buckingham(190)	Louisa(201)	
Age (years)	Integer	0	19.0	92	46.662
Gender	Binomial	0	Male(156)	Female(217)	
Height (inches)	Integer	0	52	76	65.979
Weight(pounds)	Integer	0	99	325	177.928
Frame	Polynomial	0	Large(96)	Medium(177)	
Waist (inches)	Integer	0	26	56	37.925
Hip (inches)	Integer	0	30	64	43.046
diabetes	Binomial	0	Yes(56)	No(317)	

In order to improve the accuracy of the model, the continuous features are often needed to be discretized[12]. Discretization involves two tasks: First, to determine the number of classification that we need; Second, to determine how to map continuous features values to these classification values. For the first sub-tasks, we handle like this: after the continuous features values are sorted, divide them into n intervals by specifying the n-1 points. As for the second sub-tasks, we will map all the values in an interval to the corresponding classification value. Therefore, the discretization is to choose the number of split points and determine the point location problem. In order to facilitate the processing of the data, we will divide each feature into three parts and using low, medium and high represent these feature values, respectively. The next step is to determine the split point; there are three kinds of methods to determine the split point, namely: width discretization, frequency discretization and k-means discretization. After the experiment, we know that the performance of k-means discretization is the best [13]. According to the center point of each feature obtained by k-means discretization, the features of discretization are shown in TABLE II.

978-1-5090-5434-3©2017 IEEE

TABLE II. DISCRETIZED FEATURES

Name	Low	Medium	High
Age	≤38	39~57	≥58
Height	≤62	63~65	≥66
Weight	≤170	171~227	≥228
Waist	≤36	37~44	≥45
Hip	≤41	42~50	≥51

IV. EXPERIMENT DESIGN

After the data pre-processing, the next goal is to dig out the relationships between the various features and extract some useful patterns. Now, we begin to develop a risk of type II diabetes model to predict whether a person will develop diabetes. The construction steps of the random forest mainly include generating a training set, choosing the splitting point, repeating construct the classification and regression tree and the voting. Detailed procedure is as follows:

Step1: using Bootstrap re-sampling techniques to generate k (In this paper, the k is 10) samples. Theoretically k samples cover 2/3 of the original datasets, and the rest of the data is called Out-Of-Bag (OOB), OOB can be used as test data [14].

Step2: using the k samples to generate k decision trees. At each node of each tree, we are randomly selected m features (m<M) in the M features, it is suggested starting with $m = \sqrt{M}$ and then decreasing or increasing m until the minimum error for the OOB data set is obtained. Finally choose the best split according to the Gini criterion [15]. Gini criterion and prediction class labels are shown in the Eq. (1) and Eq. (2).

$$\mathrm{Gini}(A_t) = 1 - \sum_{i=1}^{n} p_i^2 \tag{1}$$

Where p_i represents the probability of the i-th class instance; n is the number of classes; A_i represents the i-th feature.

$$c_D = \arg\max_c \left(\frac{1}{k}\sum_{i=1}^{k} I\left(\frac{n_{h_i}, c}{n_{h_i}}\right)\right) \tag{2}$$

Where c_D represents prediction class labels; $\arg\max_c$ represents a parameter to find the maximum score c; k represents the number of decision trees in a random forest; I(*) represents indicator function; n_{h_i}, c represents the classification results of the decision tree for the c class; n_{h_i} represents the number of leaf nodes in the decision tree h_i.

Step3: according to the previous two steps to predict the test samples, and combined with the test results of each tree and determines the final result in accordance with majority rule voting mechanism.

In order to validate the effectiveness of the proposed methods, we utilized another three algorithm, namely ID3 algorithm, Naive Bayes algorithm and AdaBoost algorithm[16]. Additionally, in order to further demonstrate the effectiveness of the method used in this study. We designed a different set of contrast experiments. First, the data set was divided into four subsets (20%, 40%, 60%, and 80% of the total data set, respectively), and each model was compared in each subset. The overall framework of model building is shown in Fig.2.

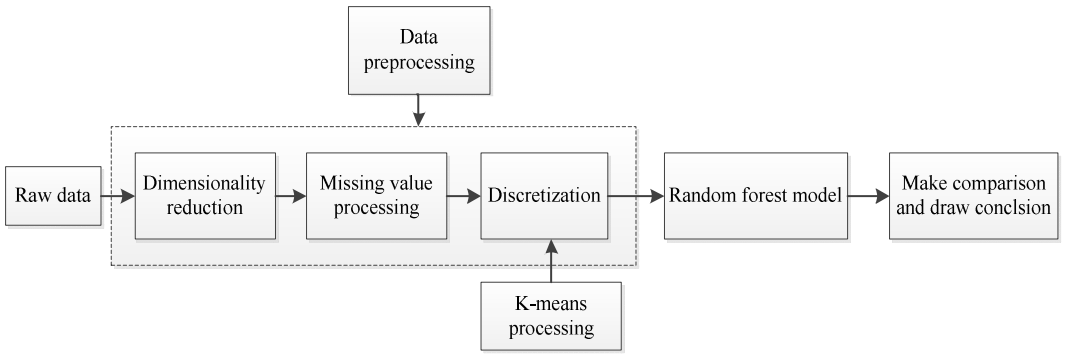

Fig. 2. The overall framework of model building

V. RESULTS ANALYSIS AND MODEL EVALUATION

According to the above experiment, some decision trees are shown in Fig.3. and Fig.4. It is easy to see that the root node of each tree in a random forest, including Waist circumference, Hip circumference, Weight and Age. It indicates that these are all important external indicators to determine whether they are suffering from diabetes. Therefore, we can extracted some rules from Fig.3 and Fig.4. The rules are as follows:

1) If Waist ≥45 inches and Hip ≥51 inches, then diabetic.

2) If Hip ≥51 inches and Height ≤ 62 inches, then diabetic.

If we consider the first rule, the rule is supported by a previous study [17], which showed that Waist Circumference (WC) or Waist-to-Hip Ratio (WHR) discriminate better the cases with diabetes from those without, as compared with BMI. Considering the second rule, recent study have shown a similar study which showed that Waist-to-Height Ratio

978-1-5090-5434-3©2017 IEEE

(WHtR) can be used to identify Type II diabetes [18]. According to these rules, when our weight, waist circumference, hip circumference is growing, we should adjust the diet structure, balanced diet and increase exercise to reduce the risk of illness.

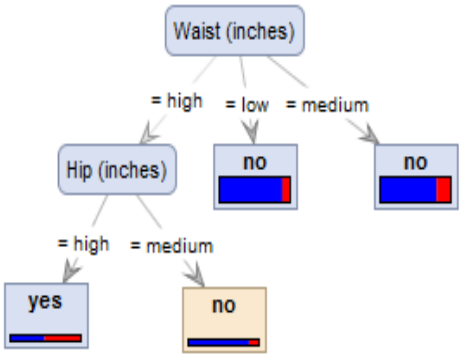

Fig. 3. Waist is the root of the decision tree

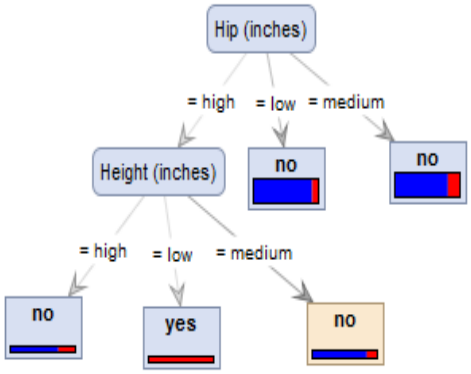

Fig. 4. Hip is the root node of the decision tree

In medical diagnosis accuracy, sensitivity and specificity are the common measures of performance metrics. Accuracy determines ability of the classifier to produce accurate disease diagnosis. Sensitivity measures the ability of the model to identify the occurrence of target class accurately. Specificity measures the ability of the model to separate the target class. The Accuracy, Sensitivity and Specificity are measured as follows [19].

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (4)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (5)$$

Where the True Positives (TP) and True Negatives (TN) are correct classifications. A False Positive (FP) occurs when the outcome is incorrectly predicted as yes (or positive) when it is actually no (negative). A False Negative (FN) occurs when the outcome is incorrectly predicted as no when it is actually yes.

The k-fold cross validation is a best measure for classifier performance [20]. Therefore, in our study, we use the 10-fold cross validation method to evaluate the reliability of the model. According to the Eq. (3), the accuracy rate of the random forest model is 85.00%, which is calculated by the 10-fold cross validation method. In addition, we also carried out ID3 algorithm, Naive Bayes algorithm and AdaBoost algorithms to obtain accuracy of the model are as follows: 78.57%, 79.89% and 84.19%. Fig.5 presents the bar graph of accuracy for 4 models.

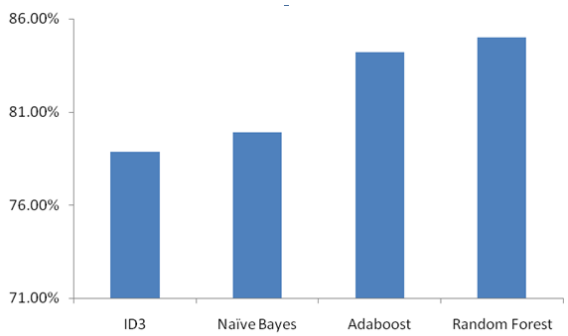

Fig. 5. Accuracy for different models

According to the Eq. (4), Eq. (5), we found the sensitivity reached a comparative high level, they achieved a sensitivity of 91.17%, 92.11%, 99.05% and 100% respectively. While the specificity achieved a quite low level. By observing the data and searching the literature, we find that the data imbalance leads to this problem. And this is a direction for the next research. However, compared to ID3, Naive Bayes and AdaBoost, the accuracy and sensitivity of random forest model are more satisfactory. It has a certain guiding significance for the early warning of diabetes. Taking into account the amount of data used in the experiments is relatively small, we made 4 groups of comparative experiment to strengthen the persuasiveness of experiments. According to the characteristics of the data set, we set up four subsets of different sizes, accounting for 20%, 40%, 60%, 80% of the total data set, respectively. With the above methods, the accuracy of each experimental group were shown in TABLE III.

TABLE III. COMPARISON WITH DIFFERENT SCALE DATA

Groups / Algorithms	Group 1	Group 2	Group 3	Group 4
Random Forest	60.00%	67.78%	80.07%	84.13%
Naive Bayes	66.65%	69.93%	78.64%	80.50%
ID3	62.25%	66.73%	66.36%	72.94%
AdaBoost	66.70%	70.10%	79.06%	81.61%

In order to better observe the effect of random forest algorithm in each experimental group, and make its line graph. It is shown in Fig. 6.

978-1-5090-5434-3©2017 IEEE

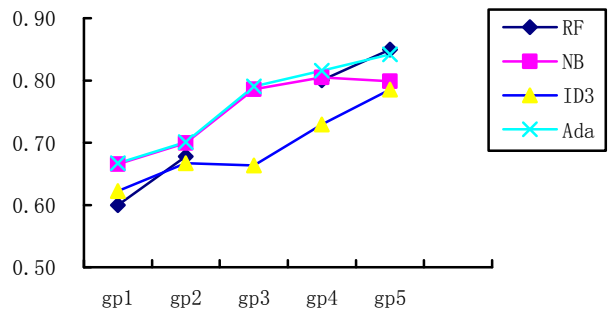

Fig. 6. The line graph of comparison with different scale data

As shown in Fig. 6, it can be concluded that with the expansion of the data, the accuracy of the random forest model is constantly improved. Additionally, the accuracy of the random forest model is also constantly improving while the same amount of increased data in the similar proportion of cases. So the random forest model can effectively predict the risk of diabetes in the case of a sufficient amount of data.

VI. Conclusions

Obviously, data mining has played a very important and decisive role in the medical industry. In this paper, we obtain some simple decision rules by establishing the random forest model, and we can make a simple prediction of whether or not we have diabetes by these simple and readily available indicators. Additionally, these indicators are relatively easy to obtain (they can be measured at home), so we can greatly reduce the cost of diagnosis. By using these indicators to predict diabetes will have a certain practical significance.

In this paper, we just use some readily available indicators to predict the risk of diabetes and there is no further study the impact of other indicators of illness, also not taken into account the impact of the tester itself suffering from other diseases on the prediction of diabetes. Expand other indicators to predict the risk of disease and update the perspective of data mining are the future direction of the prediction of the risk of type II diabetes.

Acknowledgment

We acknowledge the support from the National Natural Science Foundation of China (No. 91546123), Program for Changjiang Scholars and Innovative Research Team in University(No.IRT_15R07), and from the Program for Liaoning Innovative Research Team in University (No.LT2015002).

References

[1] C. H. Jen and C. C. Wang, “Application of classification techniques on development an early-warning system for chronic illnesses,” Expert Systems with Applications, vol. 39, pp. 8852-8858, 2012.

[2] K. Grudzinski, “Towards heterogeneous similarity function learning for the k-nearest neighbors’ classification,” Artificial Intelligence and Soft Computing, vol. 5097, pp. 578-587, 2008.

[3] J. C. Bioch and O. van der Meer, “Classification using Bayesian neural nets,” Neural Networks, vol. 3, pp. 1488-1493, 1996.

[4] H. Kahramanli, and N. Allahverdi, “Design of a hybrid system for the diabetes and heard diseases,” Expert Systems with Applications, vol. 35, pp. 82-89, 2008.

[5] B. M. Patil, R. C. Joshi and D. Toshniwal, “Hybrid prediction model for type-2 diabetic patients,” Expert Systems with Applications, vol. 37, pp. 8102-8108, 2010.

[6] http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes

[7] J. B. Schorling and J. Roach, “A trial of church-based smoking cessation interventions for rural African Americans,” Preventive Medicine, vol. 26, pp. 92-101, 1997.

[8] K. Fawagreh and M. M. Gaber, “Random forests: from early developments to recent advancements,” System science \& control Engineering, vol. 2, pp. 602-609, 2014.

[9] L.Breiman, "Random Forests," Machine Learning, vol.45, pp. 5-32, 2001.

[10] H. S. Kim, A. M. Shin, M. K. Kim, and Y. N. Kim, “Comorbidity study on type 2 diabetes mellitus using data mining,” Korean Journal of Internal Medicine, vol. 27, pp. 197–202, 2012.

[11] M. G. Ahamad, A. Aljumah, and M. K. Siddiqui, “Application of data mining: Diabetes health care in young and old patients,” Computer and Information Sciences, vol. 25, pp. 127-136, 2013.

[12] J. C. Han, J. C. Rodriguze, and M. Beheshti, “Diabetes data analysis and prediction model discovery using RapidMiner,” 2008 Second International Conference on Future Generation Communication and Networking, vol. 3, pp. 96-99, 2008.

[13] P.N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Pearson Addison-Wesley, 2006.

[14] M. Seyedhosseini and T. Tasdizen, “Disjunctive normal random forests,” Pattern Recognition, vol. 48, pp. 976-983, 2015.

[15] L. Breiman, J.H. Friedman, R.A. Olshen and C.J. Stone, Classification and Regression Trees, Chapman \& Hall, 1993.

[16] https://rapidminer.com/

[17] Q. Qiao and R. Nyamdorj, “Is the association of type II diabetes with waist circumference or waist-to-hip ratio stronger than that with body mass index?,” European Journal of Clinical Nutrition, vol. 64, pp. 30–34, 2010.

[18] Z. Xu, X. Qi1, A. K. Dahl and W. Xu, “Research: Epidemiology Waist-to-height ratio is the best indicator for undiagnosed Type 2 diabetes,” Diabet. Med. vol. 30, pp. 201-207, 2013.

[19] Z. H. Zhou, Machine Learning, 1st ed., Tsinghua University Press, 2016,pp.28-36.

[20] N. Esfandiari, M. R. Babavalian and V. K. Tabar, “Knowledge discovery in medicine: Current issue and future trend,” Expert Systems with Applications, vol. 41, pp. 4434-4463, 2014.

978-1-5090-5434-3