Fasting Blood Glucose Change Prediction Model Based on Medical Examination Data and Data Mining Techniques

Wenxiang Xiao
College of Information Engineering, Qingdao University
Qingdao China
Email: xiao_304324035@qq.com

Fengjing Shao*
College of Information Engineering, Qingdao University
Qingdao China
Email:sfj@qdu.edu.cn

Jun Ji
College of Information Engineering, Qingdao University
Qingdao China
Research Lab, Qingdao Rebind Information Technology
Inc., Qingdao, China
Email: junji@qdu.edu.cn

Rencheng Sun
College of Information Engineering, Qingdao University
Qingdao China

Chunxiao Xing
College of Information Engineering, Qingdao University
Qingdao China

Abstract*—Fasting blood glucose (FBG) is an important indicator for human's health. Prediction for FBG is meaningful for finding and healing diseases, especially for diabetes mellitus. Based on four years' historical medical examination data, a prediction model of coming year's FBG is presented using traditional data mining techniques with a novel algorithm to estimate the FBG change probability and a proposed feature selection algorithm, which combines the feature importance scores of ensemble learning and Sequential Backward Selection (SBS) algorithm to select an optimal feature subset. Experimental data are collected from a medical examination database containing 108,386 users, in which 7,136 people have four years' records. Compared with traditional support vector machine (SVM) and random forest, experimental results demonstrate that the feature selection algorithm can improve the performance of both SVM and random forest. Also the proposed method to estimate the probability of the FBG change works promisingly for giving an intuitive description of predictive result.

***Keywords*—fasting blood glucose; medical examination data; data mining; feature selection; probability estimate**

I. INTRODUCTION

Fasting blood glucose (FBG) is an important factor for human's health. The study for FBG is helpful for finding and healing some diseases, especially for diabetes mellitus. Diabetes mellitus is a growing epidemic, and many people are hard to know they have the disease [1]. Diabetes leads to many serious diseases. Appropriate management of patients at risk with lifestyle changes and medications can decrease the risk of developing diabetes by 30% to 60% [2]. Therefore, by predicting the FBG to predict diabetes mellitus is feasible and meaningful.

Medical examination data is a form of electronic medical data, similar to other electronic medical data, for example: electronic health records (EHR), electronic medical record (EMR) and the like. Researches using data mining methods to explore medical data have a long period. During past decades, Jin Park presented a method by modeling a neural network based on health risk assessment data to predict the risk of diabetes [3]. Jionglin Wu predicted the risk of heart disease in the following six months, and they compare different machine learning algorithms and different feature selection algorithms [4]. Also, some researches to analyze factors related to the development of diseases were performed. Simon G.J used the expanded association rules to explore factors associated with diabetes and populations likely to develop diabetes from electronic medical data [5]. Meric F showed medical data to predict disease-specific survival based on medical data [6]. Applications using data mining techniques are widely designed, such as the specific disease predicting system for hospital database. Karabatak designed an expert system based for detection of the breast cancer based on association rules and neural network, the system was applied to the Wisconsin breast cancer database [7]. Vijayanv V analyzed and compared some novel data mining algorithms for prediction and diagnose of diabetes mellitus, such as, EM, KNN, K-means, ANFIS and so on, and experimented on Pima Indian Diabetic Set from University of California, Irvine (UCI) Repository of Machine Learning databases [8]. Razvan Bunescu propose a generic physiological model of blood glucose dynamics to generate informative features for a Support Vector Regression model that is trained on patient specific data [9]. Sellappan Palaniappan develops a prototype heart disease prediction system using tree data mining classification techniques [10]. So far, aforementioned researches mainly focused on mining knowledge in hospital data to make corresponding predictions, while the widely existing examination data have not been leveraged to make any predictive model. In this paper, in order to effectively analyze the examination data, a FBG prediction model is proposed with a novel algorithm to estimate the FBG change probability and a new feature selection algorithm to model.

Supported by National Natural Science Foundation of China (61503208)

Supported by Natural Science Foundation of Shandong Province, China (ZR2015PF002) Foundation of Shandong Province, China (ZR2015PF002)

* Corresponding author: Fengjing Shao
E-mail address: sfj@qdu.edu.cn

978-1-5090-1893-2/15 $31.00 © 2015 IEEE
DOI 10.1109/SmartCity.2015.157

In the second section of this paper, the techniques in the process of developing the model are introduced. In the third section, the data preprocessing is introduced, including data cleaning and feature creation. In the fourth section, the process of creating the FBG prediction model is described in detail. In the fourth, the performance of the feature selection method and the method to predict the probability of FBG change are shown. In addition, the the results of feature selection and analysis on age and gender are also discussed.

II. Technology Overview

Random forests is an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees [11]. Random forests can be used to rank the importance of variables in a regression or classification problem in a natural way [12]. The random forest algorithm estimates the importance of a variable by looking at how much prediction error increases when out of bag (OOB) data for that variable values are permuted while all others are left unchanged. The necessary calculations are carried out tree by tree as the random forest is constructed [13].

Support vector machine (SVM, also support vector networks) is supervised learning model with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis [14]. Given a set of training examples, each marked for belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier [15]. In addition to performing linear classification, SVM can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces [16].

Feature selection is the process of selecting a subset of relevant features (variables, predictors) for use in model construction. The objective of feature selection is three-fold: improving the prediction performance of the predictors,providing faster and more cost-effective predictors, and providing a better understanding of the underlying process that generated the data.[18] The main methods of feature selection have three categories: wrapper methods, filter methods and embedded methods. Wrapper methods use a predictive model to score feature subsets. Sequential Backward Selection (SBS) algorithm is always utilized in wrapper method. Filter methods use a proxy measure instead of the error to score a feature subset. Embedded methods are a catch-all group of techniques which perform feature selection as part of the model construction process. The feature importance scores estimated by random forest could be taken as embedded method for feature selection.

III. Data preprocessing

The data in the research come from a medical examination center in Beijing, which contain 108,386 users. In order to keep the data integrity and standard, the records from January 2011 to December 2014 are selected from the medical examination database. And the following records are excluded: 1) users whose records last less than 4 years; 2) users who take part in medical examination in one year more than once; 3) records of which medical examination items are less than these our chosen.

Data preprocessing steps include data cleaning, data normalization, medical examination item selection and feature creation.

Data cleaning: there are some records in which the ID can't match with the medical examination items, in order to ensure data accurate, these records are deleted. For records containing missing values, in order to ensure data quality, such records are deleted too.

Data normalization: the expression of some features is not standardized. There are features represent the meaning of them with symbols, such as '+', '++', '+++' and other symbols (the meaning of the symbol is how much sugar exist in urine); if one gives up his medical items, 'give up', 'abandon examination' and other terms will indicate that he had not taken a part in the items; for some features are continuous values, feature values contain words or non-standard symbols. To solve the aforementioned cases, special symbols are replaced by numbers; similar terms are replaced by same values; non-standard terms and symbols in features are removed.

Medical examination item Selection: medical examination information includes the users' basic information (such as age, gender, etc.) and medical examination items. Both are considered as medical examination items. Medical examination items are used to created features, and the sensitive identity information not included. The medical examination item selection criterion is as follows: on one hand, the items closely related with blood sugar, such as blood lipids four, urine, urine ketene bodies, fatty liver, blood pressure, etc. are selected; on the other hand, the items a lot users participating are chosen, such as other items in blood routine, urine routine.

Feature creation: features are classified into two classes: one is global feature that almost does not vary with time or just one year's item is meaningful, such as height and age, the other is local feature whose medical examination results probably vary with year, such as weight, FBG, blood pressure and so on. The global feature is unique in data set. The local feature exists in every year, but in one year it's unique.

There are 107,854 users with 9,073,312 medical examination records in the database. The records are from 2011 to 2014. Data set consists of 8,788 users, which contains 4 years' medical examination records. 1,679 users are excluded because the records lack some feature values. Finally, records of 7,109 validated users are adopted. In 7,109 users, 4,095 are males, and 3,501 are females. Compared to the third year's FBG, 4,738 users' FBG are up in the fourth year, 2,371 users' FBG are down in the fourth year. Specially, if one's FBG unchanged, his FBG change is considered down. The processing of 4 years' medical examination data is demonstrated in Fig. 1. Data set has 139 features, including local features for three years, 45 medical examination items per year, items mainly in Blood routine, urine routine, blood biochemistry, department of internal medicine, electrocardiogram. Data set also includes four global features, the fourth year's FBG, height, age, sex. The fourth year's FBG is the response variable; the height is the mean of four years' heights; the age is the third year's age; the sex is the third year's too. A demographics is demonstrated as TABLE Ⅰ. In the database, the youngest user is 22 and the oldest is 94. Only two age groups are considered, young (20-50) and old (>50).

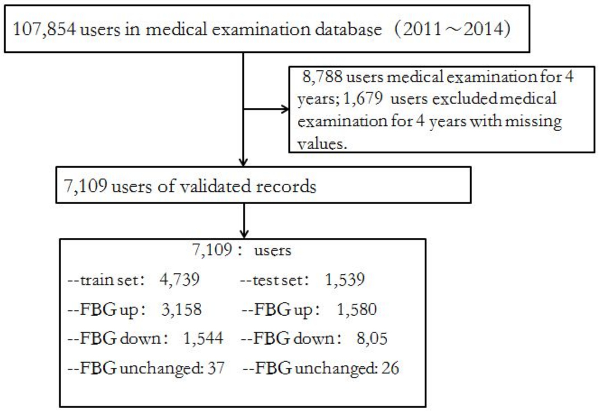
107,854 users in medical examination database（2011～2014）
8,788 users medical examination for 4 years; 1,679 users excluded medical examination for 4 years with missing values.
7,109 users of validated records
7,109： users
--train set： 4,739
--test set： 1,539
--FBG up： 3,158
--FBG up： 1,580
--FBG down： 1,544
--FBG down： 8,05
--FBG unchanged: 37
--FBG unchanged: 26

Fig. 1. Processing of 4 years medical examination data

TABLE I. DEMOGRAPHICS

		Train set		Test set	
Number		N=4,739		N=2,370	
Gender	Male	2,543		1,304	
	Female	2,196		1,066	
Age	20-50	(M)1408	(F)1305	(M)721	(F)638
	50-	(M)1135	(F)891	(M)583	(F)428

IV. MODELING

In this section, the FBG change prediction model will be introduced in detail, including the process of the feature selection algorithm, modeling with random forest, the method estimating the probability of the change of the FBG. The flow chart of the model is as shown in Fig. 2.

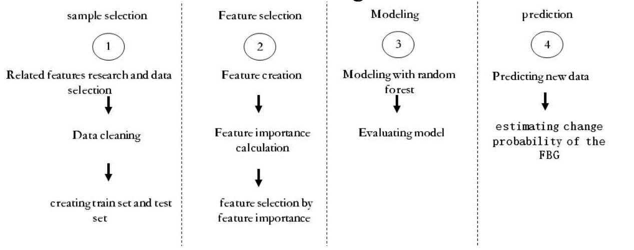
sample selection
1
Related features research and data selection
Data cleaning
creating train set and test set
Feature selection
2
Feature creation
Feature importance calculation
feature selection by feature importance
Modeling
3
Modeling with random forest
Evaluating model
prediction
4
Predicting new data
estimating change probability of the FBG

Fig. 2. The flow chart of the FBG change prediction model

A. FBG change prediction model design

The data set for modeling the FBG change prediction is created from 4 years medical examination data. A model to predict the change of the coming year's FBG based three years medical examination data are presented, and the results are translated into probability of the change of the FBG. The process of modeling is as follows:

1)Create the data set containing four years' records as the descriptions in section Ⅲ.

2)Use the data set to select the optimal medical examination items subset and optimal feature subset.

3)Divide the data set into two parts, the 2/3 of the original data set is the train set, and the rest of the data set is the test set. Generate the FBG prediction model based on the training set and the random forest algorithm.

4)Based the model above, input the testing set, get the prediction for the change of the coming year's FBG. Subtract the third year's FBG values form the predicted values, the differences are the change of FBG prediction change values. If the change values greater than 0, the users will be labeled with '1'. Otherwise, the users will be labeled with '0'. The differences are considered as the scores of the FBG change. As a result, the prediction for value of the coming year's FBG is transformed into a two-classification problem.

5)Calculate the probability of the change of the fasting blood glucose.

After the modeling, the population of probability more than 80% will be analyzed. By achieving the ROC of the model, area under curve (AUC)[19] is used to evaluate the performance of the model predicting the change of the FBG. In addition, mean absolute error (MAE) and root mean square error (RMSE)[20] are used to evaluate the accuracy of the model predicting the values of the FBG.

B. Importance of Medical Examination Item

Medical examination data contains rich information, but there is some information that is not related to the changes of FBG. This paper designs a feature selection method based on random forest, which can effectively give the importance of medical examination item, and achieve the goal of increasing model's performance and dimension reduction.

When random forest method fit the dataset, they can calculate the feature importance score. The more important the feature is, the higher the feature importance score is. After calculation of the feature importance, the medical examination item importance scores will be calculated out easily.

The medical examination item importance score calculation steps as follows:

1)Fit random forest to the whole dataset, and then feature importance scores are calculated.

2)calculate the average value of features corresponding same medical examination item, the medical examination item importance scores that related to local features can be calculated easily. For global features, the medical examination items corresponding global feature has the same importance score with the global feature.

In the feature selection stage, the SBS algorithm is used. The SBS algorithm removes one feature at one time based on the classifier performance until a feature subset of the desired size *k* is satisfied. The procedure of algorithm is as follows:

1)Sort the medical examination items by the scores of their importance, the most important is at the top.

2)Fit the train set of the corresponding medical examination items set with the random forest, and calculate the AUC value on test set.

3)Remove the medical examination item whose importance score is the lowest, fit the train set corresponding to the medical examination items set with the random forest, and calculate the AUC value on test set.

4)Repeat (3) until the medical examination item set only contains $k = 2$ items.

5)Based the AUC values, choose the medical examination item subset corresponding to the biggest AUC value as the optimal subset.

According to the above algorithm, the subset of medical examination items corresponding to the biggest AUC value is the optimal medical examination item subset. The model performs best when selecting the top 17 medical examination items. And results are shown in Fig. 3.

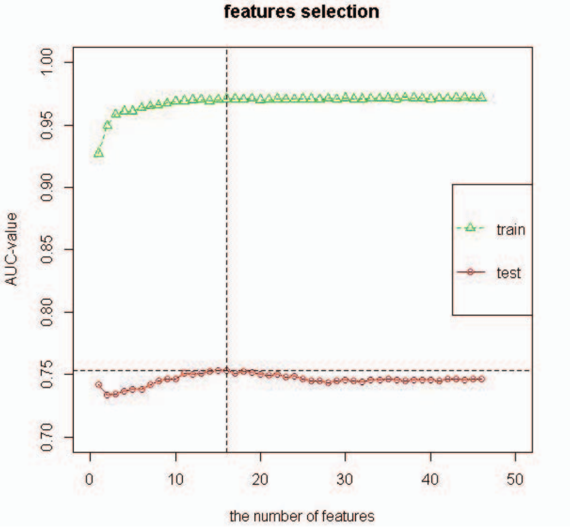

Fig. 3. The result of feature selection

By fitting the train set corresponding to the optimal medical examination items with random forest, the coming year's FBG values could be predicted. The third year's FBG values are subtracted from the predictive values, and the differences are considered as the score for the change of one's FBG in future. Based the scores, the probability distribution for the coming year's change scores of FBG is estimated. The 80% probability is taken as the threshold to distinguish whether one's FBG will be up. If one gets the probability more than 80%, his FBG will be up; otherwise, his FBG will be down. And the work adopts the 80% probability to distinguish whether one's FBG will be down is carried out.

The data set is divided into three parts: train set, test set, and validation set. The train set is used for generating the model, and the test set is used to calculate the threshold value of 80%. The validation set is used for testing the results of the threshold of 80% probability. The process of calculating the probability of the FBG change is as follows:

1)Fit the train set with random forest and predict the coming year's FBG values

2)Based the coming year's FBG values, calculate the FBG up (down) scores

3)Sort the FBG up (down) scores in an ascending order, set the step size w, from the smallest score value p, every move a step, $p=p+w$, then calculate the ratio of the population whose FBG up (down) in people whose score bigger than p.

4)Repeat moving the step until the ratio of the blood glucose up (down) is greater than 80%. At this moment, we get a FBG up (down) 80% probability and the corresponding threshold p.

C. Results Analysis and Comparison

The Fig. 4 shows the top 17 medical examination items and their importance scores, and the top 7 are effective items. They are FBG, age, waist, waist height ratio, weight, urine sugar, BMI. Not surprisingly, FBG plays a most important role in the model. Besides FBG, age is the second important role, researches have proven that age is an independent risk factor for FBG rising. Waist, waist height ratio, weight and BMI reveal the level of one's obesity. Urine sugar also indicate higher FBG in body.

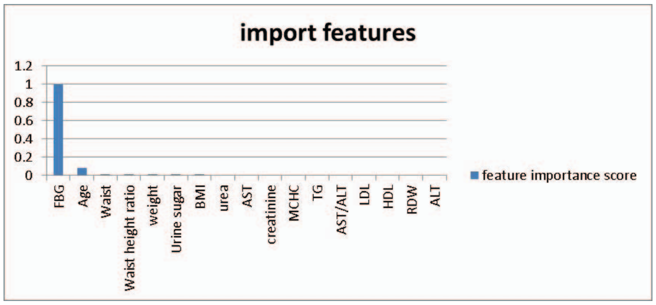

Fig. 4. Importance scores of important medical examination items

Different groups of age and gender were experimented with random forest, the models were made on train set and validated on test set with AUC. The performance of these models are demonstrated in TABLE Ⅱ. The results shown that gender is not an obviously important feature for distinguishing changes of FBG. Models on the young(20-50) population performed better than the old(50-).

TABLE II. THE PERFORMANCES OF DIFFERENT GROUPS

Age	Male	Female
20-50	74.17%	73.55%
>=50	71.26%	71.70%

The optimal medical examination item subset consists of the top 17 medical examination items. And the feature subset for the top 17 medical examination items is the optimal feature subset. In the experiment for feature selection, the data set is divided into two parts: train set and test set, the ratio is 2:1 and train set contains 4,739, test set contains 1,539. The train set is used for modeling, and the test set is used to evaluate the model results. To validate the feature selection algorithm, we model on the original data set and on the data set with feature selection individually.

Experiments are done with the R programming language 3.2.1 (https://www.r-project.org/), the main packages: RandomForest 4.6-10 (https://cran.rproject.org/web/packages/randomForest/) and e1071 1.6-7 (https://cran.r-project.org/web/packages/e1071/) are to implement random forest and SVM. The algorithm runs on a CPU E3-1225 workstation using Windows 2012 Server system, the CPU frequency is 3.20GHZ.

According to the feature subset of the optimal medical examination items, the performances of the random forest and SVM on the data set and on the data set with feature selection are compared. To compare the performances of the method, AUC, RMSE and MAE are calculated individually on the original data and the data dealt with by the feature selection method. Among all those models, FS-random Forest and FS-SVM generate the model on the data with feature selection, and random Forest and SVM generate the model on the original data. The algorithm results are shown in TABLE III.

TABLE III. PERFORMANCES OF DIFFERENT CLASSIFIERS

Classifier	AUC	RMSE	MAE
FS-random Forest	74.92%	0.5706	0.3200
Random Forest	71.96%	0.5996	0.3347
FS-SVM	72.19%	0.6672	0.3533
SVM	71.67%	0.6909	0.3685

From the experimental results of TABLE Ⅱ, compared with randomForest classifier, the performance of FS-random Forest classifier is better: AUC rises 2.96%, RMSE and MAE decreases 0.0290 and 0.0147. Compared with SVM classifier, the FS-SVM's AUC rises 0.52%, RMSE and MAE decreases 0.0237 and 0.0152. These results show that the classifiers with feature selection improve in AUC, RMSE, and MAE, which indicates that the feature selection algorithm is useful for improving the performances in abilities of classification and the accuracy of prediction. The above results proof that there is feature subset performing better than the whole feature set. This indicates some features are not helpful for prediction of FBG.

In the experiment of FBG change probability, the data set is divided into three parts: train set, test set, and validation set. The train set is 4,739, the test set is 1,185 and the validation set is 1,185. The random forest containing 1,500 decision trees is used to carry on the regression analysis on train set. The test set is used to calculate the FBG change 80% probability with corresponding threshold p on test set, the validation set is used to validate the performance of the model. Sensitivity, specialty and Positive Predictive value (PPV) are used for evaluating the performance of the model. Results are as TABLE IV:

TABLE IV. RESULTS OF FBG CHANGE PROBABILITY >= 80%

The population with FBG change probability >=80%	**FBG up probability >= 80% N=572 455 FBG up**	**FBG down probability >= 80% N=21 13 FBG down**
Sensitivity	59.32%	3.17%
Specialty	72.00%	98.96%
PPV	79.54%	61.90%

* The users for testing=1185，767:FBG up，410:FBG down

In the testing set, there are 1185 users, in which 767 users' FBG will be up, 410 users' FBG will be down. The results in TABLE Ⅲ illustrate that in 80% probability population of FBG up, most of user's FBG will be up, but for FBG down, its result is not as well as the former. We aim at finding the population with huge FBG changes. Changes of FBG up always are huger and more than that of FBG down. So the results are reasonable.

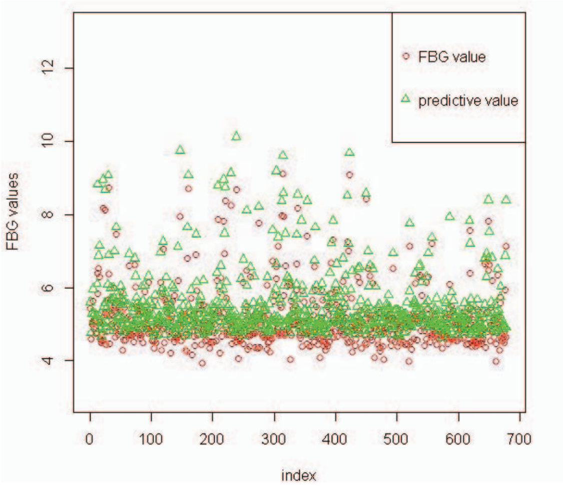

Fig. 5. Fitting of the FBG values and the predictive FBG values whose FBG up probability above 80%

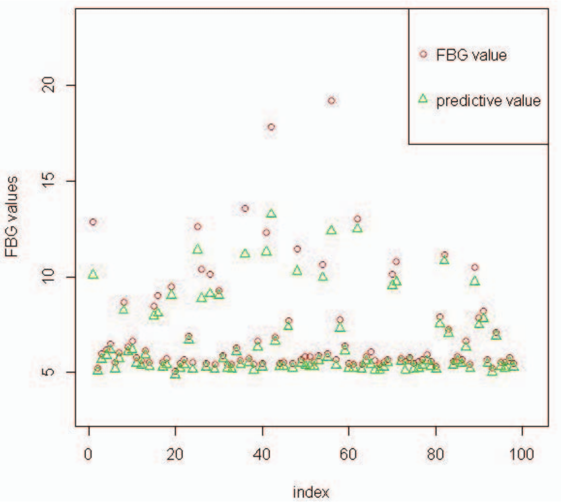

Fig. 6. Fitting of the FBG values and the predictive FBG values whose FBG down probability above 80%

As Fig. 5 and Fig. 6 shown, the predictive FBG values fit the practical FBG values generally. But in Fig. 5, almost all predictive FBG values are bigger than the practical FBG values; in Fig. 6, the fitting between them shows better. These also can explain that Changes of FBG up always are huger and more than that of FBG down.

V. CONCLUSIONS

In this paper, we propose a model based on four years' medical examination data to predict the change probability of the coming year's FBG. The model contains a feature selection method to improve the performance of the model and a novel method estimating the probability of the change of the FBG. By comparing the experimental results with random forest and SVM, the feature selection method can effectively improve the model performance. Although this method discards some weak medical examination items, some other medical examination items related to FBG may also be discarded. And in the feature subset we choose, there may exist some features whose correlations are strong. In our future research, we will explore more effective methods to select features. The approach we propose can also be used for other disease, and the related experiments will be conducted. As there are a lot of data less than 4 years, we will try semi-supervised learning algorithms to improve the performance of the model with more examination items data adopted.

References

[1] Jin P, Edington D W. A sequential neural network model for diabetes prediction[J]. Artificial Intelligence in Medicine, 2001, 23(3):277-293(17).

[2] Wu J, Roy J, Stewart W F. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches.[J]. Medical Care, 2010, 48(6):S106-S113.

[3] 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\s&\sdata Engineering, 2015, 27(1):130-141.

[4] Meric F, Mirza NQ, Vlastos G,te al: Positive surgical margins and ipsilateral breast tumor recurrence predict disease-specific survival after breast-conserving therapy. Cancer 97:926-933, 2003.

[5] Karabatak M, Ince M C. An Expert System For Detection Of Breast Cancer Based On Association Rules And Neural Network[J]. Expert Systems with Applications, 2009, 36(2):3465-3469.

[6] Palaniappan S, Awang R. Intelligent heart disease prediction system using data mining techniques[C]// Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on IEEE, 2008:108-115.

[7] Vijayanv V, Ravikumar A. Study of Data Mining Algorithms for Prediction and Diagnosis of Diabetes Mellitus[J]. International Journal of Computer Applications, 2014, 95(95):12-16.

[8] Bunescu R, Struble N, Marling C, et al. Blood Glucose Level Prediction Using Physiological Models and Support Vector Regression[C]// Machine Learning and Applications, International Conference on. IEEE, 2013:135 - 140.

[9] Centers for Disease Control and Prevention (CDC). National diabetes fact sheet: national estimates and general information on diabetes and prediabetes in the United States, 2011[J]. Atlanta, GA: US Department of Health and Human Services, Centers for Disease Control and Prevention, 2011, 201.

[10] Knowler W C, Elizabeth B C, Fowler S E, et al. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin[J]. New England Journal of Medicine, 2002, 346(3):: 393–403.

[11] Andy Liaw, Matthew Wiener. Classification and Regression by randomForest[J]. R News, 2002, 23(23).

[12] Breiman L. Random Forests[J]. Machine Learning, 2001, 45(1):5--32.

[13] Andy Liaw, Matthew Wiener. Classification and Regression by randomForest[J]. R News, 2002, 23(23).

[14] Cossock D, Zhang T. Statistical Analysis of Bayes Optimal Subset Ranking[J]. Information Theory IEEE Transactions on, 2008, 54(11):5140-5154.

[15] H Drucker, Burges C, Kaufman L, et al. Vapnik V.: Support Vector Regression Machines[J]. Advances in Neural Information Processing Systems, 1996, 28(7):391 - 394.

[16] Cortes C, Vapnik V. Support-vector networks[J]. Machine Learning, 1995, 20(3):273-297.

[17] Ajzerman M A, Braverman E M, Rozonoehr L I. Theoretical foundations of the potential function method in pattern recognition learning[J]. Automation & Remote Control, 1964, 25.

[18] Guyon I, Elisseeff A. An Introduction of Variable and Feature Selection[J]. Journal of Machine Learning Research, 2003, 3:1157-1182.

[19] Lobo J M, Real R. AUC: a misleading measure of the performance of predictive distribution models[J]. Global Ecology & Biogeography, 2008, 17(2):145-151.

[20] Willmott C J, Matsuura K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance[J]. Climate Research, 2005, 30(1):79-82.