A Method for Selecting Diabetes Features based on Random Forest

Shaofu Lin1*, Wei Ji1 and Jiangtao Pei1

1Faculty of Information Technology, Beijing University of Technology, Beijing, 100124, China

*Corresponding author’s e-mail: linshaofu@bjut.edu.cn

Abstract. This paper studies the optimal feature subset screening for diabetes according to the health check data based on the random forest algorithm. The paper takes the real physical examination records of the same batch of people in a local health check-up center from 2010 to 2015 as the data source, and evaluates the importance of the features. The preliminary fitting finds that 28 features have an impact on the response results. The AUC performance of the classifier finally selects the optimal feature subset containing 9 characteristic variables in multiple feature subsets, which provides scientific evidence and decision support for medical expert's prediction intervention, clinical diagnosis, treatment plan determination and medical research on diabetes.

1. Introduction

Diabetes is a group of lifelong metabolic diseases characterized by multiple causes of chronic hyperglycemia. Long-term blood sugar increase causeA damage to large blood vessels and microvessels and endangers the heart, brain, kidney, peripheral nerves, eyes, feet, etc. The complications are as high as 100, which is the most common complication[1]. According to the 2017 Global Diabetes Map (8th Edition) report published by the International Diabetes Federation (IDF), approximately 425 million adults worldwide have diabetes in 2017, and it is estimated that by 2045, diabetes patients may reach 629 million[2]. At present, there is no cure for diabetes, so the study of risk factors and early intervention of diabetes is of great significance.

There are usually many factors that can influence the occurrence and development of diseases, collectively referred to as disease risk factors. Disease risk factors analysis and disease type prediction are important research directions in the field of medical data mining. These risk factors are divided into controllable risk factors and uncontrollable risk factors. If the risk factors can be accurately identified and targeted prevention and treatment can be achieved, the morbidity and mortality of the disease can be greatly reduced[3]. In the mining and analysis of medical datasets, the feature selection technique can eliminate redundant features, noise features and irrelevant features, identify the feature variables or the most relevant feature subsets that are most relevant to the target variable. It has a great value for the study of disease risk factors[4].

The paper based on the continuous batch of the same batch of health check data sets, uses the random forest algorithm and the Filter idea. The random forest variable importance score and the classification AUC are used as evaluation criteria to screen out the subset of features most relevant to diabetes, and provide scientific evidence for the research, prediction and diagnosis of diabetes.
2. Overview of feature selection and random forest

2.1 Overview of feature selection
Feature selection is also called feature subset selection or attribute selection. In the practical application of data mining, data sets usually have a large number of features, among which there may be correlations between features or unrelated features with mining targets. Feature selection refers to selecting several features to optimize the specific indicators of the system, that is, selecting the most effective features from the original features to reduce the dimension of the data set, the feature selection not only obtains better precision, but also explores diseases. An important means of risk factors[5]. The main process of feature selection is shown in Figure 1.

2.2 Random forest introduction
The random forest is an integrated classifier \( R = \{ h(x, \theta_k), k=1,2...K \} \) composed of a set of decision tree classifiers, where \( \{ \theta_k \} \) is a random vector obeying independent and identical distribution, \( K \) is the number of decision trees in a random forest. Given the independent variable \( X \), each decision tree classifier determines the optimal classification result by voting[6]. The basic steps for training a random forest model are as follows:

- Suppose that given the original training data set \( (X, Y) \), the number of samples is \( n \), marked as \( (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) \), using the bootstrap method through \( n \) times random sampling samples a new sample set \( (X^*, Y^*) \), the samples in the set are represented as \( (x_1^*, y_1^*), (x_2^*, y_2^*), ..., (x_n^*, y_n^*) \), not extracted samples consists of a collection of data outside the bag.
- Train a classification regression tree on the data set \( (X^*, Y^*) \).
- Repeat steps 1-2 to generate multiple trees to form a forest.
- Each tree in the forest is classified as a sample \( x_i \) in OOB.
- The number of votes obtained for each category is counted, and the category with the highest number of votes is considered to be the category of the sample \( x_i \).

3. Data collection and preprocessing

3.1 Data collection
This paper looks forward to the adoption of the data set prerequisites: the same batch of human population test data sets for two consecutive years. The time dimension is not limited. For example, the physical examination data of Zhang in 2010-2011 and the physical examination data of Li in 2012-2013 are both valid data. In order to ensure the authenticity and reliability of the research data used in this paper, the real health data of the residents after desensitization in a hospital health checkup center is used. The data form is the physical examination report. A total of 2,325 medical examiners were collected for a total of 4,650 medical examination reports for two consecutive years.
In order to facilitate the follow-up study, the “sample” is used to represent the data of the single-person dimension. Each sample contains two consecutive physical examination data of the same medical examiner. The first year data includes basic conditions, color ultrasound, blood routine, urine routine, and biochemical projects, a total of 49 indicators, the second year of data for a single indicator fasting blood glucose values.

Table 1. Physical examination data details.

| Project       | Index                  | Remarks | Project       | Index                                      | Remarks |
|---------------|------------------------|---------|---------------|---------------------------------------------|---------|
| Basic Situation | number                 | ID code | Urine Routine | ketone body                                | character |
|               | gender                 | character |              | bilirubin                                  | character |
|               | age                    | numerical value |              | glucose                                    | character |
|               | systolic blood pressure | numerical value |              | hematocrit                                 | numerical value |
|               | diastolic blood pressure | numerical value |              | platelet count                             | numerical value |
| Color Ultrasound | fatty liver           | character |              | alanine                                    | numerical value |
|               | hemoglobin concentration | numerical value |              | aspartate                                  | numerical value |
|               | red blood cell count    | numerical value |              | alkaline phosphatase                      | numerical value |
|               | hematocrit              | numerical value |              | glutamyltranspeptidase                    | numerical value |
|               | average red blood cell volume | numerical value |              | total bilirubin                           | numerical value |
|               | white blood cell count   | numerical value |              | direct bilirubin                          | numerical value |
|               | percentage of lymphocytes | numerical value |              | total protein                              | numerical value |
|               | number of neutrophils    | numerical value |              | albumin                                   | numerical value |
|               | neutrophil percentage    | numerical value |              | globulin                                   | numerical value |
| Blood Routine | platelet count          | numerical value |              | white ball ratio                          | numerical value |
|               | mean platelet volume     | numerical value |              | urea nitrogen                              | numerical value |
|               | platelet pressure        | numerical value |              | uric acid                                 | numerical value |
|               | platelet distribution width | numerical value |              | creatinine                                | numerical value |
| Urine Routine | urobinogen              | character |              | indirect bilirubin                        | numerical value |
|               | colour                  | character |              | total cholesterol                         | numerical value |
|               | leukocyte               | character |              | triglyceride                              | numerical value |
|               | sharpness               | character |              | high density cholesterol                  | numerical value |
|               | nitrite                 | character |              | low density cholesterol                   | numerical value |
According to the World Health Organization's diagnostic criteria for blood sugar in diabetes: fasting blood glucose in diabetic patients $\geq 7\text{mmol} / \text{L}$, the second year of the sample fasting blood glucose value is divided into two numerical intervals on the scale of $7\text{mmol} / \text{L}$, fasting blood glucose is less than $7\text{mmol} / \text{L}$ is recorded as 0, fasting blood glucose greater than or equal to $7\text{mmol} / \text{L}$ is recorded as 1.

3.2 Data preprocessing
Data preprocessing is an important basic work in data science research and it has great significance for feature selection. There is no standard process for data preprocessing, and different approaches are often taken for different tasks and dataset attributes[7]. For the research data set and research objectives of this paper, the common processes of data preprocessing are: data integration, data standardization, qualitative variable transformation, outlier processing, missing value processing, and removal of unique attributes.

- Data integration is the consolidation of data from multiple data sources in a single data store. The raw data collected in this paper is the resident health check report, which cannot be directly used as the input form of the data experiment. Therefore, the attributes and values in the medical report are entered into the data table for further research.
- Because some medical examination items in the medical examination report are not standard numerical representations, some symbols are added to indicate their specific conditions. For example, if “+” and “-” are used to represent positive and negative, “+-” and “+++” are used to represent the degree. For such data, the following data is processed: data containing “+” and “-” for positive and negative is subjected to dummy coding, and converted into data represented by 0 or 1, using “+-” and “+++” to represent the degree is graded and converted to numeric data.
- Qualitative variables are concepts in statistics, also known as categorical variables, which are generally expressed in non-numeric descriptions. Qualitative variables cannot be used as direct inputs, and conversion of qualitative variables to quantitative variables is required. There are five original qualitative variables in this dataset, namely gender, fatty liver, color, clarity, and nitrite, which are converted into quantitative variables by means of dumb coding.
- The outliers refer to the values that are obviously not in line with the logic and the industry consensus. The process of entering the attributes and values in the medical report into the data table finds that a few obvious abnormal values or indicators do not correspond to the attributes, which may be due to physical examination. Human error occurred during the registration of the indicator. Since the sample of this problem is small and does not affect the scale of the overall data set, this part of the data is deleted.
- The missing value refers to the situation where the value corresponding to the attribute is vacant. Because some medical examiners have not done the whole project physical examination and the reasons for human error in a small part of the entry process, some missing values appear in the original data set. For this part, three methods are adopted: if the medical examiner only participates in a small number of physical examination items, the single sample has too many missing values, and the sample is deleted; the attribute corresponding to the missing value in the sample has no obvious association with diabetes. The numerical missing values are complemented by the mean, and the descriptive missing values are complemented by the majority, the attributes corresponding to the missing values in the sample are based on common sense understanding and medical literature comparison may be related to diabetes, to ensure the rigor of the experiment, delete this sample.
- Unique identification attributes are usually ID attributes that uniquely identify a sample. This attribute does not describe the distribution of attributes in the entire sample and the weight distribution of the response features, so the ID code in the data set is deleted in the experiment.
After a series of data preprocessing processes, the number of samples of the final available medical examination data is determined to be 2064, and the pre-processed partial data is shown in Figure 2.

4. Experiments and analysis

4.1 Feature selection method

According to the characteristics of the research data set and the research objectives, this paper combines the random forest algorithm with the Filter-based feature selection idea, and uses the Filter-based feature selection method based on random forest to select features.

Filter-based feature selection technology can eliminate noise features and irrelevant features before training classification or prediction models, and identify the optimal feature subsets that are most relevant to the target variables. However, the current Filter-based feature selection algorithms are mostly based on statistical methods. For medical data sets with complex interrelationships between features, statistical tests are difficult to accurately characterize the relative importance of classification features to target variables, thus affecting the reliability of feature selection results. Random forest has the advantages of high classification accuracy, fast calculation speed and good robustness. It is good at identifying the main correlation features from data sets with small marginal effects and complex interactions. It is the preferred feature selection method in medical data mining. Helps people understand how a particular classification feature affects a target variable[8]. The use of random forest variable importance scores can be used as the evaluation criteria for feature selection. Identifying the feature subsets most relevant to the target variables through various feature search strategies is an effective way to explore the risk factors of diabetes. The Filter-based feature selection based on the importance score of random forest variables used in this paper makes use of the advantages of random forests in dealing with data with complex interactions between features, based on the importance score of random forest variables, according to the AUC performance of the classifier. Iteratively selects the optimal feature subset.

4.2 Feature selection process and results

The basic idea of feature selection in this paper is as follows: Firstly, the random forest algorithm is used to initially screen the features in the original dataset, that is, the whole dataset is fitted, all the features are sorted according to the variable importance score, and then the variable importance score is selected iteratively. The top feature, and finally the optimal feature subset is selected in multiple feature subsets according to the AUC performance of the classifier. In order to minimize the model deviation caused by parameter selection and to compare with the performance of other algorithms, a 10-fold cross-validation method is used in the process of calculating the random forest variable importance score and constructing the classification model. Specifically, the data set is first divided into 10 equal parts, and the hierarchical subset is used to ensure that the divided subset is the same as the original data set, and 9 of them are used as the training set, and the remaining 1 is used as the test
set. The iterative execution is performed 10 times, and the variable importance score generated by the highest iteration of the AUC of the random forest on the training data set is selected as the criterion for evaluating the importance of the feature, and the average value of the AUC of the test set in the 10 iterations is calculated as a feature. The evaluation criteria of the subset quality, and finally select the feature subset with the highest AUC as the optimal feature subset. The algorithm flow is shown in the following figure, where MAXAUC represents the maximum AUC on the subset, LMAXAUC represents the maximum AUC in 10-fold cross-validation, LAUC represents the AUC of each iteration in 10-fold cross-validation, Opt represents the optimal feature subset, Sort indicates the maximum feature ordering of AUC in 10-fold cross-validation. The specific process is shown in Figure 3.

After the data pre-processing research data set, the number of samples is 2064, the data dimension is 49, and the data set is preliminarily fitted by the random forest algorithm. The order of importance of the characteristics obtained by the AUC performance is shown in Figure 4.
It can be seen from the figure that there are 28 features in the research dataset that have influence on the response variables. In the feature selection process, each time the k*step features are selected to form a feature subset, wherein the parameter step is used to set the process of iteratively selecting feature subsets. In the case of increasing the number of features, considering the efficiency and accuracy of the selection, the default is set to 3, that is, each feature subset contains Top3, Top6, Top9... features, and the AUC performance on the test set is shown in Figure 5.

![Figure 5. Test set AUC performance.](image-url)

It can be seen that when the feature subset contains Top9 features, the AUC performance on the test set is the best, then the best feature subset after the feature selection of the data set can be determined as the data set by the random forest algorithm. The Top9 features are: fasting blood glucose, triglyceride, age, urea nitrogen, low density cholesterol, creatinine, alanine aminotransferase, high density cholesterol, mean platelet volume, in line with medicine Pathological consensus. When TOP9 features are selected, the AUC is 0.828, which is superior. The maximum difference between other data in the verification experiment is 0.093, and the minimum difference is 0.022. The difference is relatively obvious, indicating that this method is in screening the optimal feature subset has a good result and excellent fit.

5. Conclusion
The feature selection method proposed in this paper can effectively select the characteristic variables related to diabetic patients. The screening results are highly consistent with the existing medical research conclusions, and the optimal feature subset AUC performance is significantly better than other feature subsets. The clinical and experimental perspectives verify the validity and scientificity of the proposed method, and this method does not absolutely limit the specific business scenarios, which has a strong promotion. In this paper, the first 3n features with the largest variable importance score are selected to form the subset of candidate features. The iterative experiment is used to select the optimal feature subset. The step parameter in the algorithm is set to 3, which is based on the size of the data set used in this paper. Determine an empirical value. How to set the threshold more intelligently is still a subject to be studied.
References

[1] Kuller L H, Velentgas P, Barzilay J, et al. Diabetes Mellitus[J]. Arteriosclerosis Thrombosis & Vascular Biology, 2000, 20(3):823.

[2] Cho N H, Shaw J E, Karuranga S, et al. IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045[J]. Diabetes Research and Clinical Practice, 2018:S0168822718302031.

[3] Al-Lawati J A, Mohammed N B, Ibrahim A Z, et al. Control of Risk Factors for Cardiovascular Disease Among Adults with Previously Diagnosed Type 2 Diabetes Mellitus: A Descriptive Study from a Middle Eastern Arab Population[J]. Open Cardiovascular Medicine Journal, 2012, 6(1):133-140.

[4] Cheng T H, Wei C P, Tseng V S. Feature Selection for Medical Data Mining: Comparisons of Expert Judgment and Automatic Approaches[C]// IEEE International Symposium on Computer-based Medical Systems. 2006.

[5] Guyon I, Elisseeff A, Kaelbling L P. An introduction to variable and feature selection.[J]. Journal of Machine Learning Research, 2003, 3(6):1157-1182.

[6] Svetnik V, Liaw A, Tong C, et al. Random forest: a classification and regression tool for compound classification and QSAR modeling.[J]. Journal of Chemical Information & Computer Sciences, 2003, 43(6):1947.

[7] Duhamel A, Nuttens M C, Devos P, et al. A preprocessing method for improving data mining techniques. Application to a large medical diabetes database[J]. Studies in Health Technology & Informatics, 2003, 95(95):269.

[8] Ziegler A, König I R. Mining data with random forests: current options for real-world applications.[J]. Wiley Interdisciplinary Reviews Data Mining & Knowledge Discovery, 2014, 4(1):55-63.