Diabetes prediction based on ensemble learning

Zhenzhen Wang

School of Management, Shanghai University, Shanghai 201800, China

Abstract

Diabetes not only poses a threat to the life and health of patients, but also brings a heavy economic burden to patients. Various machine learning algorithms are widely used in the construction of medical models, so that patients can be classified and tested for diabetes. This article uses the Pima Indians diabetes data set, using an integrated learning algorithm, including RF (Random Forest), Adaptive Boosting (AdaBoost), Gradient Boosting Decision Tree (GBDT), Extreme gradient boosting tree (Xgboost), LightGBM algorithm, the research results show that LightGBM has the best predictive effect, and also found that in the prediction of diabetes, the most important features are glucose and insulin. The research in this article can not only improve the classification and detection of diabetes, but also play a great role in the prevention and control of diabetes.

Keywords

Ensemble learning; Diabetes; prediction; machine learning.

1. Introduction

With the change of people's lifestyle and the increase of the elderly population, diabetes has become an important global public health problem. According to the statistics of International Diabetes Federation (IDF), the number of diabetes patients in the world was 38.2 billion in 2013, and it is expected that the number of diabetes patients will reach 592 million in 2035 [1]. At the same time, the premature death and disability caused by diabetes also brings a heavy economic burden. The IDF estimates that 10 per cent of global healthcare spending goes on diabetes management. And 50.1% of adults worldwide with diabetes don't know they have it. Due to lack of access to health services, the proportion of undiagnosed patients is as high as 66.8 per cent in low-income countries, but 38.3 per cent in high-income countries.

The development of artificial intelligence has brought about machine learning techniques that do not require strict assumptions. Classical single machine learning methods (single classifiers) such as artificial neural networks, support vector machines and decision trees have emerged, which have been widely used in the construction of medical models. Ensemble learning methods have become a hot research topic because they can integrate the results of multiple single classifiers to achieve "secondary learning", and the common ensemble methods include Bagging, Boosting, etc. The predictive effect of ensemble learning model is often better than that of single classification model, which can better classify and detect diabetes patients [2].

Qian et al. [3] used LVQ network to study the classification of diabetes/abnormal glucose tolerance and compared it with traditional discriminant analysis. It is found that LVQ network not only can obtain better prediction effect, but also can be used directly and conveniently without considering the data characteristics or data conversion. Nie et al. [4] combined the rough set theory and random forest algorithm to realize the classification. At the same time, they realized the classification with the help of R language and tested the subset data of diabetes complications. Luo et al. [5] applied the data mining C4.5 algorithm to a large number of measured data of type 2 diabetes, obtained several effective rules after processing, and demonstrated the effectiveness of C4.5 algorithm in data processing of type 2 diabetes through tests. Chen et al. [6] add hemoglobin A1C as one of the feature inputs to improve the accuracy of the model. Meanwhile, K-Nearest Neighbor (KNN) and neural network are used to classify diabetes, so that the accuracy of neural network is slightly higher than KNN.

Heydari et al. [7] compared support vector machines, artificial neural networks, decision trees, nearest neighbor and Bayesian networks on the data set of type 2 diabetes to find the best algorithm for diagnosis of this disease. The results show that the performance of artificial neural network is the best on the selected data set. Pekel et al. [8] used a classification regression tree optimized by artificial neural network and genetic algorithm to diagnose diabetes. The results on the Pima Indian diabetes dataset showed that The proposed classification regression tree method of genetic algorithm is superior to artificial neural network and artificial neural network method based on classification and regression tree and genetic algorithm in classification accuracy. Alharbi et al. [9] used extreme learning machine neural network for classification and evolutionary genetic algorithm for feature extraction in the real data set of diabetes patients in Saudi Arabia. Through genetic algorithm, the dimension of feature space was reduced and only effective features were selected. Then the data is input to the extreme Learning machine neural network for classification, and the final research results show the effectiveness of this method. Chen et al. [10] chose two relatively common algorithms, Adaboost.M1 and LogitBoost, according to the diabetes clinical trial data, to establish the diabetes diagnosis machine model based on these two algorithms, and the results show that LogitBoost has a slightly better classification effect than Adaboost. Choubeyd et al. [11] conducted a comparative analysis and performance evaluation on the results obtained with and without the use of genetic algorithm in the same group of classification technology. The research results showed that genetic algorithm can help to remove irrelevant attributes, reduce costs and computing time, and improve accuracy.

By combing relevant literature at home and abroad, it can be seen that existing studies on diabetes focus on the prediction of a single classifier model, or the comparison of several single classification models. At present, there are relatively few comparative studies on different integrated learning methods. On the other hand, most of the analysis aspects of diabetes only focus on finding models with better generalization performance, and seldom pay attention to the relationship between characteristics and whether the characteristics have diabetes, so as to provide suggestions for clinical medicine.

In this paper, the data set is the Pima Indian diabetes data set which is open to the public. Meanwhile, the selected algorithm of ensemble learning includes Random Forest, AdaBoost, GBDT, Xgboost and LightGBM algorithm. Through comparison, the most suitable prediction model is found. The most important were glucose and insulin levels. The research in this paper can not only improve the accuracy of diabetes classification detection, but also play a great role in the control and prevention of diabetes.

2. Preparatory Knowledge

2.1. Data preprocessing

Data is the basis of data mining, which determines the success or failure of data mining to a large extent. The purpose of data preprocessing is to make full use of useful data and eliminate incomplete, redundant and noisy data. The methods of data preprocessing include data cleaning, data integration, data transformation and data protocol.

The goal of data cleansing is not only to eliminate errors, redundancies, and data noise, but also to harmonize various data sets based on different, incompatible rules. Data integration is the consolidation of data from multiple data sources, which may include multiple databases, data cubes, or general files, into a consistent data store (such as a data warehouse); Data transformation is to find the feature representation of data, use dimension transformation to reduce the number of effective variables or find the invariant of data, including normalization, specification, switch and projection operations; Data code is found in the task and content of the data itself, on the basis of finding relies on found useful features of target expression data to reduce data model, thus the data same as far as possible under the premise of the smaller amount of data to the greatest extent, there are two main ways: attribute selection and data sampling, respectively for attributes and records in a database.

2.2. Evaluation of the model

The evaluation criteria used to measure the generalization ability of the model in the classification task include the value of accuracy rate, precision rate, recall rate and AUC.

2.2.1. Accuracy

Accuracy is the proportion of correctly classified samples to the total number of samples, which is applicable to both binary and multi-classification tasks. For example set D, accuracy is defined as

$$\mathrm{acc}(f;D) = \frac{1}{m}\sum_{i=1}^{m}\prod(f(x_i) = y_i) \tag{1}$$

2.2.2. Precision and recall ratio

For the dichotomy problem, the sample can be divided into true positive, false positive, true negative and false negative cases according to the combination of its real category and the learner's prediction category. Let TP, FP, TN and FN represent their corresponding sample numbers respectively, then TP+FP+TN+FN= total number of samples.

The precision ratio P and recall ratio R are respectively defined as

$$\mathrm{P}=\frac{TP}{TP+FP} \tag{2}$$

$$R = \frac{TP}{TP+FN} \tag{3}$$

2.2.3. AUC value

The full name of ROC is "Receiver Operating Characteristic" curve, and its vertical axis is "True Positive Rate (TPR)". On the horizontal axis is the "False Positive Rate," or FPR. AUC (Area Under ROC Curve) is the area under ROC curve.

Suppose the ROC curve is composed of coordinates $\{(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)\}$ points are connected in order to form $(x_1 = 0, x_m = 1)$, then the AUC can be estimated as

$$\mathrm{AUC}=\frac{1}{2}\sum_{i=1}^{m-1}(x_{i+1} - x_i)\cdot(y_i + y_{i+1}) \tag{4}$$

3. Model selection

Ensemble learning completes learning tasks by building and combining multiple learners. The individual learner is usually generated by an existing learning algorithm from the training data, such as C4.5 decision tree algorithm, BP neural network algorithm, etc. The current integrated learning methods can be roughly divided into two categories, that is, the serialization method which has strong dependence between individual learners and must be generated serially, and the parallel method which has no strong dependence between individual learners and can be generated simultaneously. The former is represented by Boosting and the latter by Bagging.

3.1. Random Forest algorithm

Random Forest algorithm is based on Bagging integration built by decision tree based learner and further introduces random attribute selection in the training process of decision tree. To be specific, the traditional decision tree selects an optimal attribute in the attribute set of the current node (assuming there are d attributes) when choosing partition attributes. In the random forest, for each node of the base decision tree, a subset containing K attributes is randomly selected from the attribute set of the node, and then an optimal attribute is selected from this subset for division. Here, the parameter K controls the introduction degree of randomness: if K=d, the construction of the base decision tree is the same as that of the traditional decision tree; If K=1, an attribute is randomly selected for division [12].

3.2. AdaBoost algorithm

AdaBoost algorithm can obtain different test sample sets by changing the distribution weight of samples. After each training, the weight of misclassified samples is increased, and the weight of correct classified samples is reduced to obtain a new training sample set. Therefore, the next weak classifier training should focus on the misdivided samples. After each cycle, a weak classifier corresponding to a certain feature is obtained, and its weight is calculated according to its classification error. After T cycles, T weak classifiers are obtained, and these T weak classifiers are connected according to their respective weights to form the final strong classifier [13].

3.3. GBDT algorithm

GBDT is an integrated machine learning algorithm, whose core is to take the negative gradient of the loss function as the approximate value of the current model data residuals, fit a CART regression tree according to the approximate residuals, and constantly repeat this process to reduce errors. Since the output value of the classification problem is discrete value, the output error of the category cannot be directly fitted from the output category, so the square loss function commonly used for continuous values in the regression tree is abandoned, and the exponential loss function or logarithmic likelihood loss function is adopted, which is equivalent to the difference between the predicted probability of the category and the true probability value to fit the residual [14].

3.4. Xgboost algorithm

The full name of Xgboost is eXtreme Gradient Boosting, namely extreme Gradient Boosting tree, which is an extension of Gradient Boosting Machine algorithm. Boosting classifier belongs to ensemble learning model, whose basic idea is to combine hundreds of tree models with lower classification accuracy into a model with higher accuracy. The model iterates constantly and generates a new tree each iteration. Boosting classifier has the core of how to generate a reasonable tree at each step. Gradient Boosting Machine algorithm adopts the idea of gradient descent when generating each tree and moves forward in the direction of minimizing the given objective function based on all the trees generated in the above step. Under reasonable parameter Settings, a certain number of trees need to be generated to achieve the expected accuracy. When the data set is large and complex, the Gradient Boosting Machine algorithm requires a huge amount of computation. Xgboost is an implementation of the GradientBoosting Machine, which can automatically parallel using multiple threads of the CPU and improve the algorithm to improve accuracy. Xgboost base learner has both gb-tree and linear classifier (gblinear), so as to obtain linear regression or logistic regression with L1+L2 penalty. Its loss function adopts second-order Taylor expansion, which has the characteristics of high accuracy, not easy to overfit, scalability, and can be distributed to process high-dimensional sparse features. Therefore, under the same circumstances, Xgboost algorithm is more than 10 times faster than similar algorithms [15].

3.5. LightGBM algorithm

LightGBM is a new boosting framework model proposed by Microsoft in 2015, which introduces two new technologies on the basis of the traditional GBDT: gradient unilateral sampling technology and independent feature combination technology. The gradient unilateral sampling technique can eliminate a large part of the data with low gradient, and only use the remaining data to estimate the information gain, so as to avoid the influence of the long tail of low gradient. Since data with large gradient is more important to information gain, this technique can still obtain fairly accurate estimates under the premise of much less data than traditional GBDT. Independent feature merging technology implements the binding of mutually exclusive features to reduce the number of features. In addition, in the traditional GBDT algorithm, the most time-consuming step is to use the Pre-Sorted method to enumerate all possible feature points on the sorted feature value, and then find the optimal partition point. However, LightGBM uses histogram algorithm to replace the traditional Pre-Sorted to reduce memory consumption [16].

4. Experimental results and analysis

4.1. Data preprocessing

The data set is an open Pima Indian dataset from the National Institute of Diabetes and Digestive and Kidney Diseases. This dataset can be used to predict whether a patient has diabetes or not based on existing diagnostic information. Pregnanci (number of pregnancies), Glucose (glucose content), BloodPressure (diastolic blood pressure), SkinThickness (skin thickness index), Insulin (insulin content), BMI(body mass index), DiabetesPedigreeFunction, Age. The label is Outcome (whether the patient has diabetes), with a value of 0 meaning the patient does not have diabetes and a value of 1 meaning the patient does.

4.1.1. Filling in missing values

In this data set, the minimum data of some features such as glucose content and body mass index should not be 0, so it is understandable that there is data missing in the data set. The idea of missing value filling is to convert these 0 values into NaN values first, and then calculate the average value of each case according to the result of "sick or not", and then use the average value to fill the missing value.

4.1.2. Feature extraction

Pearson coefficient can be used to calculate the correlation of features, which represents the degree of linear correlation between two groups of features. The value range is -1 to 1. The mutual relationship between features can be known through feature extraction.

4.1.3. Standardization of data

The purpose of data standardization is to make the data comparable, this paper adopts StandardScaler model for standardization. The StandardScaler class is a class that normalizes and standardizes data such that the data for each attribute is clustered around 0 with a standard deviation of 1.

4.2. Model Building

In this paper, the data set is divided according to the ratio of 8:2, in which 80% (including 614 pieces of data) is used as the training set for training the model, and the other 20% (including 154 pieces of data) is used as the test set to detect and evaluate the effect of performance.

In this paper, Random Forest, AdaBoost, GBDT, Xgboost and LightGBM in sklearn machine learning library are adopted. The selected performance indicators include accuracy rate, precision rate, recall rate and AUC value (area under ROC curve).

4.3. Result Analysis

4.3.1. Feature relationship

As shown in Figure 1, the relationship between each feature is positive, and the reason behind this is that these features are risk factors leading to the diagnosis of diabetes. There was a strong correlation between skin thickness index and body mass index, age and pregnancy times, glucose content and insulin content.

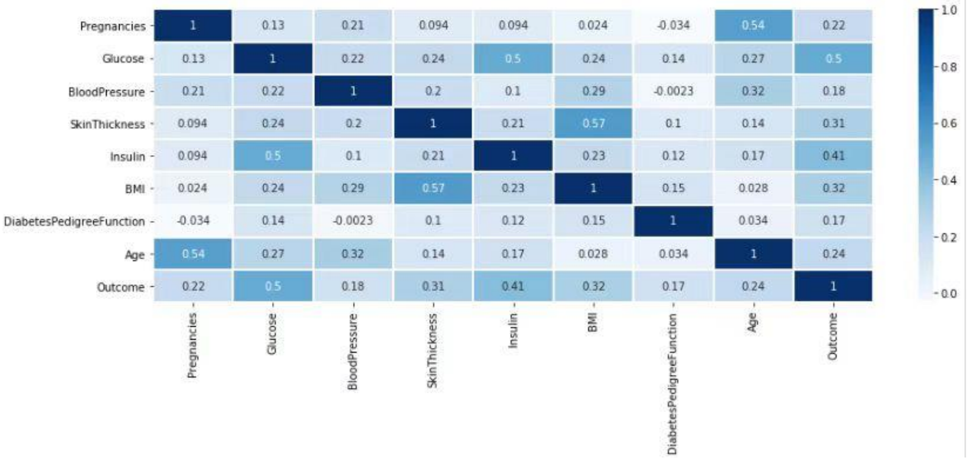

Figure 1 Relationship of diabetes characteristics

On the other hand, the correlation coefficient between each feature and "whether or not you have the disease" is also positive, and the correlation order from high to low is shown in Table 1 below. Among them, Glucose and Insulin have a greater influence on the diagnosis of diabetes.

Table 1 Feature importance ranking

feature	coefficient
Glucose	0.496
Insulin	0.411
BMI	0.315
SkinThickness	0.308
Age	0.2384
Pregnancies	0.222
BloodPressure	0.175
DiabetesPedigreeFunctiom	0.174

4.3.2. Model comparison and analysis

The following is the comparison of classification results based on diabetes data, as shown in Table 2. Through comprehensive analysis of various performance evaluation indexes, it can be seen that LightGBM has the best prediction effect among the five classification models, while Random Forest has the worst prediction effect. Take AUC value as an example: LightGBM has the best effect, followed by Xgboost, AdaBoost, GBDT, and Random Forest has a slightly worse effect.

Table 2 Comparison of classification results of ensemble learning

Classification model	Accuracy	Precision	Recall	AUC
Random Forest	0.857	0.766	0.766	0.832
AdaBoost	0.895	0.792	0.894	0.895
Xgboost	0.890	0.768	0.915	0.897
GBDT	0.896	0.804	0.872	0.889
LightGBM	0.903	0.786	0.936	0.912

5. Conclusion

In this paper, we compare and analyze the applicability of different integrated algorithms in the classification of diabetes on the publicly available diabetes data sets. The integration algorithms compared in this paper include Random Forest, AdaBoost, GBDT, Xgboost and LightGBM. The accuracy, AUC value and other evaluation indexes were used to select the model. The results show that LightGBM algorithm has better prediction effect. In addition, the results show that among the 8 features of diabetes, the most important ones are glucose content and insulin content, which can provide some help for the prevention of diabetes and make due contributions to the domestic medical industry. The limitation of this paper is that the patients in the data set are all Pima Indian females aged at least 21 years old. Subsequent studies can make the distribution of attributes in the data set more balanced, so as to improve the generalization ability of the classification model to some extent. At the same time, the ensemble learning method adopted in this paper has not been improved. Later studies can compare the improved ensemble learning algorithm with the traditional ensemble learning algorithm.

References

[1] GRP IDFDA. Update of mortality attributable to diabetes for the IDF Diabetes Atlas: Estimates for the year 2013[J]. Diabetes Res Clin Pract, 2015, 109(3): 461-5.

[2] Cao, W., C. Li, T. He, et al. Predicting Credit Risks of P2P Loans in China Based on Ensemble Learning Methods[J]. Data Analysis and Knowledge Discovery, 2018, 2(10): 65-76.

[3] Qian, L., L.-y. Shi, M.-j. Cheng. Study on the application of artificial neural network on diabetes mellitus/insulin-glucose tolerance classification[J]. Zhonghua liuxingbingxue zazhi, 2003, 24(11): 1052-6.

[4] Nie B., Z. Wang ,J.-q Du et al. The Study on Classification of Secondary Complications of Diabetes Based on Rough Set and Random Forest [J]. Journal of Jiangxi Normal University(Natural Science), 2014, 38(03): 278-81.

[5] Luo, S., H. Cheng, Y. Gu, et al. Application of C4.5 Algorithm in the Construction of the Type 2 Diabetes Classified Rules[J]. Application Research of Computers, 2004, 21(7): 174-176,179.

[6] Chen, Z., Y. Du, C. Zou, et al. Classification of diabetes based on K-Nearest Neighbor and neural network[J]. Chinese Journal of Medical Physics, 2018, 35(10): 1220-1224.

[7] HEYDARI M, TEIMOURI M, HESHMATI Z, et al. Comparison of various classification algorithms in the diagnosis of type 2 diabetes in Iran[J]. Int Diabetes Dev Ctries, 2016, 36(2): 167-73.

[8] PEKEL OZMEN E, OZCAN T. Diagnosis of diabetes mellitus using artificial neural network and classification and regression tree optimized with genetic algorithm[J]. Journal of Forecasting, 2020, 39(4): 661-70.

[9] ALHARBI A, ALGHAHTANI M. Using Genetic Algorithm and ELM Neural Networks for Feature Extraction and Classification of Type 2-Diabetes Mellitus[J]. Applied Artificial Intelligence, 2019, 33(4): 311-28.

[10] CHEN P, PAN C. Diabetes classification model based on boosting algorithms[J]. Bmc Bioinformatics, 2018, 19.

[11] CHOUBEYD K, PAUL S, SHANDILYA S, et al. Implementation and Analysis of Classification Algorithms for Diabetes[J]. Current Medical Imaging, 2020, 16(4): 340-54.

[12] Zhou Z. Machine Learning [M]. Beijing: Tsinghua University Press, 2016.01.

[13] Zhan, W., D. He, S. Shi. Recognition of kiwifruit in field based on Adaboost algorithm[J]. Transactions of the Chinese Society of Agricultural Engineering, 2013, 29(23): 140-146.

[14] Weng X., P-l LV. Subway IC Card Commuter Crowd Identification Based on GBDT Algorithm [J]. Journal of chongqing jiaotong university(natural science), 2019, 38(05): 8-12.

[15] Ye Q., H Rao., M-s J. Sales prediction of stores based on xgboost algorithm [J]. Journal of nan chang university (natural science) , 2017, 41(03): 275-81.

[16] Xie Y.,W Xiang, M.-z J et al. An application and analysis of forecast housing rental

Based on xgboost and lightgbm algorithms [J]. Computer Applications and Software, 2019, 36(09): 151-5+91.