Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/6/20230749

Research on using Adaboost with K-Means and SMOTE to predict the incidence of diabetes

Hongquan Gao[1, 3, †], Dan Zuo[2, †]

[1]Faculty of Engineering, the University of Hong Kong, Guangzhou, 511300, China
[2]College of Business, City University of Hong Kong, Guangzhou, 511300, China

[3]hongquan@connect.hku.hk
[†]These authors contributed equally

Abstract. Diabetes is one of the most diseases in the world. In the last 40 years, the number of persons worldwide with diabetes has tripled. There were 108 million patients over the age of 18 in 1980 and 422 million in 2014, accounting for 8.5% of the entire population at that time. Diabetes directly caused 1.5 million fatalities worldwide in 2012, with hyperglycemia-related illnesses accounting for 2.2 million deaths. Diabetes is expected to be the 7th greatest cause of death by 2030 according to the World Health Organization. As the risk of diabetes increases, machine learning algorithms are used to improve early diagnosis of diabetes, and various researchers have also done some corresponding algorithms for predicting diabetes machine learning. As a commonly used machine learning algorithm, AdaBoost integrated learning algorithm is superior in the diagnosis and prediction of diabetes mellitus. In this paper, it is proposed that a hybrid model to detect the risk of diabetes. This hybrid model is detected and eliminated by K-means-based outliers, synthesizing the distribution of minority data oversampling techniques (SMOTE), and Adaboost to classify diabetes. According to the final experimental result, the model prediction accuracy is 0.950 after using the hybrid model in the PIMA dataset. In the future, if a larger number of sample training data are utilized for training, the model's accuracy will improve.

Keywords: classification, K-Means, SMOTE, Adaboost.

1. Background introduction

Diabetes is a metabolic condition characterized by excessive blood sugar and is one of the most dangerous chronic diseases. Diabetes can cause persistent damage to a wide variety of tissues, particularly the eyes, kidneys, heart, blood vessels, neurons, and nerve function. If a method that can effectively improve the accuracy of diabetes prediction and diagnosis is discovered, it will be able to detect and treat diabetes in its early stages using a variety of methods [1].

Because of the large number of indicators for diagnosis of diabetes, if it is to be analyzed from the group, the amount of data would be so large and the data may be missing. It is difficult to achieve satisfactory results by single machine learning model (LR, SVM). In recent years, the accuracy of most diabetes classification predictions has improved greatly. Karol Grudzinski used the KNN algorithm to make the accuracy of diabetes predictions reach 75.5% [2]. The accuracy rate obtained by the neural network is 75.4%, and the final rate of the classification prediction using the Bayesian method is 79.5%.

© 2023 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/6/20230749

The mixed neural network (Artificial Neural Network (ANN) and Fuzzy Neural Network (FNN) model) proposed by Allahverdi can reach to a much higher number, 84.24% [3]. Although the prediction accuracy obtained by using these models has improved step by step, in most experimental scenarios, the integrated learning method is better than the single machine learning model. In addition, although the Pima Diabetes dataset was used as a sample in most diabetes experiments, most of them did not detect and process data outliers during the data preprocessing phase, which greatly affected the final analysis results. Therefore, this paper proposes a hybrid prediction model based on K-Means outlier detection, synthetic minority oversampling technique (SMOTE) and Adaboost to analyze the diabetes data, so that the accuracy and AUC of the prediction model enhanced.

This paper will focus on the classification performance of various tree models, such as decision trees, random forests, random forests based on automatic parameter adjustment, and Adaboost based on hybrid models. Various types of tree model comparison goal are accuracy and AUC (Area Under Curve) baseline, these two indicators able to characterize the effect of the final classification of diabetes data. The structure of the full text is as follows. Next section will introduce the related work, the third chapter will carry out data preprocessing, modeling, test process description and test results analysis, the final conclusion will be stated in the fourth section.

2. Related work

2.1. K-Means outlier detection

Clustering is a popular technique and is generally used to group data points in groups or clusters [4]. The K-Means algorithm based on the partitioning method has become the most widely used clustering algorithm because of its simple, fast and efficient processing of large-scale data. In this paper, the K-Means algorithm is used to detect outliers. K-Means clustering is an excellent outlier detection method. The core idea of the K-Means algorithm based on outlier detection is, firstly detect the outliers of the dataset by using the distance-based outlier detection method, and then randomly select K data points as clusters in the non-outliers. The initial seed uses the traditional K-Means algorithm to cluster non-outliers, and finally divides the outliers into corresponding clusters. The idea of the algorithm is shown in Figure 1.

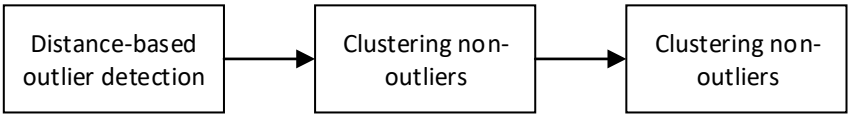

Figure 1. K-Means outlier detection step.

2.2. SMOTE data balancing algorithm

SMOTE (Synthetic Minority Oversampling Technique), a new approach based on the random oversampling algorithm. The general classification data set has a large distribution difference between the number of most classes and a few classes. This phenomenon is called data imbalance [5]. Learning through unbalanced data sets is a problem that must be faced in supervised learning because the standard classification algorithm is designed to explain the balance class distribution. One such method is called oversampling, which creates a balanced class distribution by creating artificial data.

2.3. Decision Tree Model

The decision tree model is widely used in data classification for its ease of understanding, which was proposed by Quinlan [6]. The decision tree algorithm is a supervised learning algorithm that uses known answers and is used to build tree data. The most essential aspect influencing the quality of the results is the classification accuracy attained on the training dataset, as well as the size of the tree. Classification is the process of modeling different data categories while acquiring expected values for object categories or unknown properties during training on the dataset, and it is a vital duty of allocating objects to one of several predetermined categories. Decision tree algorithms that are currently accessible include ID3,

DOI: 10.54254/2755-2721/6/20230749

C4.5, and CART. Each algorithm employs a distinct set of rules to select the optimum split for the goal of selecting the best constructing tree.

2.4. Boosting

Boosting is a landmark algorithm in the field of machine learning, which can improve the performance of any given learning algorithm. The "Probably Similar Correct" learning model proposed by Valiant in 1984 gave birth to the idea of Boosting. The concepts of strong learning algorithm and weak learning algorithm are defined in the PAC model [7]. If a learning algorithm learns a set of samples and the recognition rate is high, it's known as a strong learning algorithm. If the accuracy performance is only marginally higher than the random guess, which is 50%, it is a weak learning algorithm.

Boosting is a powerful technique for enhancing classification performance. The weak classifiers are recombined in a certain way to create a strong classifier with much enhanced classification performance. This approach successfully translates rough rules of thumb into highly accurate prediction rules. The strong classifier enhances the result of classifying data by voting and then pick the best number of votes on the weak classifier. The algorithm is a simple weak classification algorithm lifting process, which is continuously trained to improve the ability to classify data [8].

2.5. Adaboost

Freund and Schapire updated the Boosting algorithm in 1999, naming it the Adaboost algorithm, which does not demand prior knowledge about weak learning algorithms and has almost the same computing efficiency as Freund's Boosting algorithm proposed in 1991 [9, 10]. Adaboost is an acronym for Adaptive Boosting, which implements:

(1) The error rate of weak learning algorithms can be lowered using adaptive methods and mechanisms. The error rate can achieve the desired impact and goal after several iterations.

(2) The exact spatial distribution of the samples is not required. Adjusting the sample space distribution after each weak learning, updating the weights of all training samples, and reducing the weights of successfully categorized samples in the sample space can meet the objective. Misclassified sample weights are improved so that the next time when learned weakly, we will be more concerned with these misclassified samples. The algorithm can be easily applied to practical problems, so it has become the most popular Boosting algorithm.

The main principle of AdaBoost is to use the same training set to train multiple weak classifiers, and then combine these weak classifiers in a certain way to obtain a strong classifier.

Given : $(x_1, y_1),...,(x_m, y_m)$ *where* $x_i \in X, y_i \in \{-1,+1\}$.

Initialize: $D_1(i) = 1/m$ for $i = 1,2,3,...,m$.

For $t = 1,2,3,...,T$:

- Train weak learner using distribution D_t.
- Get weak hypothesis $h_t : X \to \{-1,+1\}$.
- Aim : select ht with low weighted error :

$$\varepsilon_t = \Pr_i \sim D_t[h_t(x_i) \neq y_i].$$

- Choose $\alpha_t = \frac{1}{2}\ln(\frac{1-\varepsilon_t}{\varepsilon_t})$.
- Update, for $i = 1,2,3,...,m$:

$$D_{t+1}(i) = \frac{D_t(i)\exp(-\alpha_t y_i h_t(x_i))}{Z_t}$$

where Zt is a normalization factor (chosen so that $Dt+1$ will be a distribution).

Output the final hypothesis :

$$H(x) = sign(\sum_{i=1}^{T} \alpha_t h_t(x)).$$

Figure 2. The boosting algorithm AdaBoost.

DOI: 10.54254/2755-2721/6/20230749

The process of AdaBoost is shown in Figure 2. This paper gives m as the training sample $(x_1, y_1), \cdots (x_m, y_m)$. In each round $t = 1,2,3, \cdots, T$, the distribution D_t calculates the m training samples as in Figure 2, and the given weak learning algorithm is used to find the classifier $h_t: X \rightarrow \{-1, +1\}$, where the purpose of the weak learner is to find a classifier relative to D_t the weakly weighted ε_t error. The combined hypothesis H calculates a weighted classifier which shown in formular (1):

$$F(x) = \sum_{t=1}^{T} \alpha_t h_t(x) \tag{1}$$

This is said H to be a weighted majority vote that is calculated as a classifier, where each classifier is given a weight α_t.

3. Experimental analysis

This section uses a variety of machine learning algorithms to predict diabetes classification.

3.1. Dataset

The Pima Diabetes data used in this study was derived from the UCI public data set. The data consisted of 768 cases, divided into health data (500 cases) and disease data (268 cases), including 8 attribute values as shown in Table 1. To avoid overfitting and enhance the validity of the model, the diabetes data was separated into two sub-data sets in this experiment. One sub-data set was utilized for training, while the other was used as a test. The ratio of two sub-data sets is commonly 2:1. As a result, the training and test data sets contain 514 records and 254 records respectively. Finally, through the training data set, verify the performance of the test set on different classification algorithms to compare the advantages and disadvantages of each classification algorithm model.

Table 1. Property description.

Attributes	Description
Pregnancies	Number of times pregnant
Glucose	Plasma glucose concentration at 2 hours in an oral glucose tolerance test
BloodPressure	Diastolic blood pressure
SkinThickness	Triceps skin fold thickness
Insulin	2-hour serum insulin
BMI	Body mass index
DiabetesPedigreeFunction	Diabetes pedigree function
Age	Age
Outcome	Class variable

3.2. Data pre-processing

To improve the precision of the final experimental results, data pre-processing is necessary. Since the collection process of the source data is uncontrollable, it leads to some outliers (e.g., blood pressure: 400), missing values, etc. In view of these situations, if the data is not pre-processed before data modeling, the resulting model will not perform well and affect the accuracy of the model. Therefore, data pre-processing is a very important stage in machine learning.

Data transformation and feature reconstruction: Data changes are regularization of the diabetes dataset, which can be trained and tested by normalizing the data to a uniform scale. In this dataset, this paper uses Z-score and MinMaxScaler for data regularization and convert all features into the given region. At the same time, the data features are reconstructed, and the regularized data features are taken as new features of the data set.

The equations provided in Equations (2) and (3) explain how to use regularization methods to transform data values.

DOI: 10.54254/2755-2721/6/20230749

$$MinMaxScaler\left(feature_{range} = (0,1), copy = True\right) \tag{2}$$

$$Z - Score\left(feature_{range} = (0,1), copy = True\right) \tag{3}$$

The equations provided in Equations (4), (5), and (6) explain the specific conversion steps.

$$X_{std} = \frac{X - X_{.min}}{X_{.max} - X_{.min}} \tag{4}$$

$$X_{scaled} = X_{std} * max - min + min \tag{5}$$

$$Z - Score = \frac{(x - \mu)}{\sigma} \tag{6}$$

The final converted results are shown in Table 2. (Where Feature0 represents the pre-conversion feature and Feature represents the transformed feature).

Table 2. Transformed features name.

Feature0	Feature1	Feature0	Feature1
Pregnancies	minMaxPreg	Insulin	zscore_insulin
Glucose	zscore_glucose	BMI	zscore_bmi
BloodPressure	zscore_pressure	DiabetesPedi-greeFunction	minMaxPedigree
SkinThickness	zscore_thick	Age	log_Age

K-Means outlier detection: This experiment chooses 2 as the K value. That is because the "Outcome" variable contains two results, and the discrete point threshold is set to threshold=2. After iteration 500 times, the outlier point is finally detected and deleted. As shown in Table 3, this detection method greatly improves the final experimental accuracy.

Table 3. K_MEANS instance number before and after.

Name	Total Instances	Attributes
K-Means_Before	768	8
K-Means_After	611	8

After 6 rounds of K-means outliers are removed, the data sets distributed in reasonable intervals are placed in the next training. Experimental results demonstrate that removing outliers can improve the accuracy of the final classification.

SMOTE data balance: The SMOTE algorithm can generate new samples by using the imputation formula (7).

$$x_{new} = x_i + (\hat{x}_i - x_i) \times \delta \tag{7}$$

When it is a random value between 0 and 1.

Finally, the proportion of positive and negative samples in the diabetes dataset (as shown in Table 4) is adjusted to 50%.

Table 4. Smote instance number before and after.

Name	Total Instances	Number of Diabetes	Number of Healthy	Attributes
Smote_Before	611	406	205	8
Smote_After	812	406	406	8

DOI: 10.54254/2755-2721/6/20230749

3.3. Modeling

In the above steps, this paper used K-Means for outlier detection, excluding the outliers from the data set, and then using the SMOTE algorithm to balance the data. After the data pre-processing step, this paper used 7 classifiers to train the data set. These classifiers include algorithms such as decision trees, SVM, Logistic Regression (LR), Random Forest (RF), Random Forest (RF) + GridSearch, Random Forest (RF) + Hyperopt, and Integrated Learning Algorithm (AdaBoost). In order to reduce the training test bias caused by data set partitioning, this paper used 50% cross-validation for training tests.

3.4. Experimental results

Tables 5 and 6 demonstrate the classification prediction results under two conditions. Table 5 represents the experimental results after data pre-processing using K-Means and SMOTE algorithms. Table 6 represents the experimental results using raw data and without any data pre-processing methods. The AUC, Accuracy, Precision, and Recall in table are used to measure the performance, and finally the A-C, Accuracy, Precision, and Recall can be applied to prove that the K-Means and SMOTE algorithms in the data pre-processing improve the final classification prediction.

The machine learning algorithm's classification rate or accuracy rate for true positives (TP-correct classification is true), false negatives (FN-error classification is false), true negative (TN-error classification is true), and false positives (FP- is correctly classified as false) are calculated as formular (8).

$$Accuracy = \frac{TP + FP}{TP + FP + TN + FN} \tag{8}$$

Table 5. Comparison of classification performance (auc, accuracy, precision, recall) of diabetes on 7 classifiers (Utilizes Data Pre-processing Methods).

	AUC	**Accuracy**	**Precision**	**Recall**
Decision Tree	0.829	0.832	0.770	0.734
SVM	0.903	0.859	0.779	0.828
LR	0.883	0.772	0.720	0.562
RF	0.911	0.875	0.847	0.781
RF+GridSearch	0.937	0.864	0.810	0.797
RF+Hyperopt	0.941	0.870	0.786	0.859
Proposed Model	**0.989**	**0.950**	**0.930**	**0.921**

Table 6. Comparison of classification performance (auc, accuracy, precision, recall) of diabetes on 7 classifiers. (Does Not Utilize Data Pre-processing Methods).

	AUC	**Accuracy**	**Precision**	**Recall**
Decision Tree	0.807	0.844	0.788	0.703
SVM	0.903	0.866	0.779	0.811
LR	0.857	0.766	0.692	0.486
RF	0.937	0.870	0.824	0.757
RF+GridSearch	0.963	0.900	0.844	0.878
RF+Hyperopt	0.959	0.900	0.815	0.892
AdaBoost	0.963	0.903	0.849	0.838

The experimental evaluation index by combining K-Means and SMOTE technology is further improved than the evaluation index without using any data pre-processing method. As what can be seen from Table 5 and Table 6, for diabetes prediction, in addition to considering accuracy and AUC, Precision reflects the proportion of true positive samples in the positive example of classifier prediction, that is, can more accurately determine A sample of a suspected diseased group that is truly ill. Recall reflects the proportion of positive cases that the classifier correctly predicts to the total positive case. That is, it

DOI: 10.54254/2755-2721/6/20230749

can measure the predictive ability of the classifier in the positive case of diabetes prediction. The larger the value, the stronger the positive predictive ability of the classifier. In Table 3, by comparison, the performance of the AdaBoost hybrid prediction model based on K-Means and SMOTE is the best. The values of the four metrics AUC, Accuracy, Precision, and Recall are 0.989, 0.950, 0.930, and 0.921. Compared with other classifier models, the model this paper proposed can not only accurately identify the diseased samples, but also make effective judgments on the true disease samples.

4. Conclusion

In this study, the research compares decision trees, support vector machines, logistic regression, random forests, random forests with GridSearch methods, random forests with the Hyperopt method, and AdaBoost classification results with K-Means and SMOTE methods, based on the Pima Diabetes dataset. The classification result shows that AdaBoost with K-Means and SMOTE methods can achieve the best results. It is because K-means can identify and help remove misinformation, and SMOTE can amplify the features of the data to a certain extent, thereby improving the training accuracy. It may be concluded and proved here that AdaBoost combined with K-Means and SMOTE approaches can significantly increase classification performance. At the same time, it solves the problem of inaccurate fitting caused by missing values and outliers. As the result, for classification issues, if the experimenter needs to select the tree model for data classification, it is recommended to use AdaBoost with K-Means and SMOTE methods for classification. From the perspective of optimizing this research, the number of dataset samples is still relatively small, requiring more experimental data to validate our proposed hybrid model. Because the existing clustering technology and the utilized data balancing algorithm still have certain defects. For instance, when use SMOTE to cluster, it may increase the degree of overlap between classes, and some samples that cannot provide useful information will be generated. Therefore, different data balancing algorithms and clustering algorithms will be used in the future to further improve the effectiveness of data pre-processing to improve the accuracy of the model and the metrics such as AUC.

References

[1] Who.int. 2022. Diabetes. [online] Available at: <https://www.who.int/news room/factsheets/detail/diabetes> [Accessed 15 October 2022].Another reference.

[2] Grudziński, K., 2008, June. Towards Heterogeneous Similarity Function Learning for the k-Nearest Neighbors Classification. In International Conference on Artificial Intelligence and Soft Computing (pp. 578-587). Springer, Berlin, Heidelberg.

[3] Kahramanli, H. and Allahverdi, N., 2008. Design of a hybrid system for the diabetes and heart diseases. Expert systems with applications, 35(1-2), pp.82-89.

[4] Madhulatha, T.S., 2012. An overview on clustering methods. arXiv preprint arXiv:1205.1117.

[5] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, pp.321-357.

[6] Quinlan, J.R., 1996. Learning decision tree classifiers. ACM Computing Surveys (CSUR), 28(1), pp.71-72.

[7] Valiant, L.G., 1984. A theory of the learnable. Communications of the ACM, 27(11), pp.1134-1142.

[8] Freund, Y., Schapire, R. and Abe, N., 1999. A short introduction to boosting. Journal-Japanese Society For Artificial Intelligence, 14(771-780), p.1612.

[9] Freund, Y. and Haussler, D., 1991. Unsupervised learning of distributions on binary vectors using two layer networks. Advances in neural information processing systems, 4.

[10] Schapire, R.E., 2003. The boosting approach to machine learning: An overview. Nonlinear estimation and classification, pp.149-171.