Advanced Materials Research Vols. 931-932 (2014) pp 1427-1431
© (2014) Trans Tech Publications, Switzerland
doi:10.4028/www.scientific.net/AMR.931-932.1427

Ensemble Learning Model for Diabetes Classification

Nongyao Nai-arun[a] and Punnee Sittidech[b*]

Department of Computer Science and Information Technology

Faculty of Science, Naresuan University, Phitsanulok, Thailand

[a]nongyao25@hotmail.com, [b*]punnee@nu.ac.th

Keywords: Diabetes, Classification, Ensemble Learning, Bagging, Boosting

Abstract. This paper proposed data mining techniques to improve efficiency and reliability in diabetes classification. The real data set collected from Sawanpracharak Regional Hospital, Thailand, was fist analyzed by using gain-ratio feature selection techniques. Three well known algorithms; naïve bayes, k-nearest neighbors and decision tree, were used to construct classification models on the selected features. Then, the popular ensemble learning; bagging and boosting were applied using the three base classifiers. The results revealed that the best model with the highest accuracy was bagging with base classifier decision tree algorithm (95.312%). The experiments also showed that ensemble classifier models performed better than the base classifiers alone.

Introduction

Diabetes is a chronic disease in which a person has high fasting blood sugar. The disease occurs either when the pancreas does not produce enough insulin or when the body cannot effectively respond to the insulin that is produced. Diabetes increases the risk of stroke and heart disease. In addition, it can damage heart, blood vessels, kidneys, eyes and nerves. The World Health Organization (WHO) reported that 347 million people worldwide had diabetes. They also pointed that more than 80% of diabetes deaths occurred in low and middle income countries. Moreover, WHO forecasts that diabetes will be the 7th leading causes of death in 2030 [1]. Nowadays, the number of diabetes patients in Thailand is increased. In 2011, 3.5 million Thais had diabetes and 1 of 13 adults Thais had diabetes [2].

Data mining is the analysis methodologies for knowledge discovery from database. It is a computational process of extracting interesting patterns in large data sets [3]. Data mining tasks include data clustering, data association and data classification. Classification is a process of finding a model to predict appropriate classes of the unknown data. The basic concept of classification consists of 2 steps; model construction and prediction. There are several classification techniques, such as naïve bayes, k-nearest neighbors, decision tree, regression, and artificial neural networks. These classification algorithms have been employed on medical, business and industry applications. For example, Meng et al. [4] compared logistic regression, artificial neural networks and decision tree models for predicting diabetes or prediabetes using risk factors. The results found that decision tree model had the best classificaion accuracy. Yeh et al. [5] used decision tree, bayesian classifier and back propagation neural network to construct classification models for cerebrovascular disease. After analyzed and compared, the decision tree model was chosen as the most appropriate prediction model.

When construct the classification model, the data used to construct model may have noise or imbalanced information. To improve classification accuracy, the ensemble methods were introduced. We can combine multiple models that lead to bias and vaiance reductions. The ensemble methods such as bagging and boosting have been presented [6]. They can be complied with individual base classifier. Researches have accomplished the use of ensemble methods. For instance, Liang and Zhang [7] proposed the effectiveness of bagging predictor by comparing statistical tests of 12 bagging classifiers for each medical dataset. The results revealed that bagging with decision tree performs well on the extremely imbalance and high dimensional large datasets. Abellán [8] examined ensemble mothods with decision tree classifier based on imprecision probabilities and uncertainty measures. The results show that boosting is an excellent method to combine with decision tree.

In this research, we investigated on real patients' data collected from Sawanpracharak Regional Hospital. The objective is to classify whether the patient has diabetes. Three popular base classifiers combined with feature selection technique were applied. Then, two ensemble methods; bagging and boosting with those base classifiers, were applied to observe the performance of models. Ten-fold cross validation was used to avoid model over-fitting and accuracy measurement was used for model evaluation. The rest of this paper is organized as follows: base classifiers and ensemble methods will be introduced. Experiments and their results will be presented. Finally the results will be discussed and concluded.

Base Classifiers

Three popular classification techniques, namely naïve bayes, k-nearest neighbors and decision tree were used in this paper. The first technique**,** naïve bayes, is based on Bayes' theorem which is the statistical classifier. The prediction relies on class membership probabilities. This algorithm has also exhibited high accuracy and speed when applied to large database. The classifier assumes that the effect of each attribute value on a given class is independent of the value of the other attributes [3]. The second technique is k-nearest neighbors which is widely used for large training sets. This method searches the pattern space for the k training points that are closest to the unknown object. These k points are so called the k nearest neighbors. To find the nearest neighbors, Euclidean distance, as the most popular distance metric, is calculated [9]. Then, the majority vote is used to make the decision to the class of unknown object. The last technique is decision tree. It is the most popular algorithm since the algorithm is easy and understandable. The method aims to find the best split attribute and assigned it as the root node. Then each sub-tree is recursively computed to find the lower level best split and so on until all sub-trees reach acceptable purity. The final decision tree will be used as the classification model. Mostly, we found that this technique quality is highly associated with the classification accuracy [10, 11].

Ensemble Methods

The two ensemble methods; bagging and boosting were applied in this paper. The idea of ensemble learning model is to generate multiple versions of a predictor (base classifier) and uses these results to get an aggregated predictor. Bagging (Bootstrap aggregating) was first developed by Leo Breiman in 1994 to improve classification accuracy and unstable classification problems [12]. This method used multiple versions of a training set which is generated by random. Each of these data sets is used to construct a classifier. The outputs of the models are combined by voting to make decision for the last output. It also reduces variance and helps to avoid overfitting. It is normally applied to decision tree algorithm, however it can be used with other algorithms as well [13, 14]. Boosting was first introduced by Schapire et al. [15] for improving the performance of classification algorithm. The basic idea of boosting is to repeatedly construct weak classifiers which is slightly correlated with the true classification. Boosting technique can be used to significantly reduce the error of weak classifier. Despite the potential benefits of boosting promised by the theoretical results, the true practical value of boosting can only be assessed by testing the method on classification problems. In 1996, Freund and Schapire developed boosting method called AdaBoost which is designed for binary classification [16, 17].

Models Evaluation

Each classification model was fitted using 10-fold cross validation to avoid model over-fitting. The evaluation metric used in this work is accuracy to evaluate the effectiveness of the analysis models. This metric computed from number of elements in confusion matrix which are commonly evaluated [18]. The confusion matrix, in case of two classes prediction contains true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Accuracy is defined by the following formula:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Experiments and Results

The classification algorithms were applied to the data set to find the best diabetes classifier model. The conceptual framework of 5 steps is shown in Figure 1.

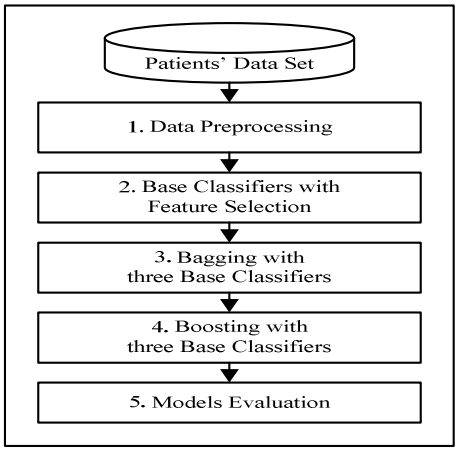

Fig. 1 Conceptual Framework

Step 1 Data preprocessing: The patients' data were collected from 27 Primary Care Units (PCU) in Sawanpracharak Regional Hospital during 2011-2013. There were 48,763 records including 20,743 diabetes and 28,020 non-diabetes. The data set composed of 15 input attributes and 1 output attribute. Descriptions of each attribute are presented in Table 1.

Table 1. Patients' data set.

No.	Attributes	Description	Values
1	AGE	Age (year)	Mean(49.44), S.D.(18.43), Min/Max (16/109)
2	BMI	Body Mass Index (kg/m^2)	Mean(26.51), S.D.(6.05), Min/Max (15/59)
3	WEIGHT	Weight (kg)	Mean(68.00), S.D.(16.10), Min/Max (36/160)
4	HEIGHT	Height (cm)	Mean(160.12), S.D.(7.79), Min/Max (130/190)
5	WAIST_CM	Waist Circumference (cm)	Mean(90.83), S.D.(14.72), Min/Max (50/183)
6	BPH	BPH/Systolic Blood Pressure (mmHg)	Mean(119.86), S.D.(15.28), Min/Max (70/198)
7	BPL	BPL/Diastolic Blood Pressure (mmHg)	Mean(76.03), S.D.(10.35), Min/Max (41/149)
8	SEX	Sex	1:Male, 2:Female
9	STATUS	Status	1:Single, 2:Marry, 3:Widow, 4:Divorce, 5:Separate, 6:Ordain
10	EDUCATION	Education	1:Before Primary, 2:Primary, 3:High School, 4:Diploma, 5:Bachelor, 6:Higher Bachelor
11	BGROUP	Blood Group	1:A, 2:B, 3:AB, 4:O, 5:A-Negative, 6:B-Negative, 7:AB-Negative, 8:O-Negative, 9:A-Posiitve, 10-B:Positive, 11-AB-Positive, 12:O-Positive
12	SMOKE	Smoke	1:No Smoke, 2:Rarely, 3:Occasionally, 4:Often
13	ALCOHOL	Drink Alcohol	1:No Drink, 2:Rarely, 3:Occasionally, 4:Often
14	DMFAMILY	Diabetes Family	1:Have History of Diabetes, 2:No have History of Diabetes
15	HTFAMILY	Hypertension Family	1:Have History of Hypertension, 2:No have History of Hypertension
16	***CLASS***	***1 : Diabetes 2 : Non-Diabetes***	

Step 2 Base classifiers with feature selection: The 15 input attributes were ranked by using gain ratio algorithm for selecting important attributes of the diabetes risk factors. The attributes were ranked as the following order; WAIST_CM, BMI, WEIGHT, DMFAMILY, AGE, HTFAMILY, EDUCATION, ALCOHOL, BPH, SMOKE, BPL, BGROUP, HEIGH, STATUS, and SEX. Then, three well known algorithms; naïve bayes, k-nearest neighbors and decision tree, were used to construct classification models using the first 13th ranked attributes. All models were tested using 10-folds cross validation to avoid model over-fitting. The result from decision tree algorithm has the highest accuracy which is 94.621% as shown in Table 2.

Step 3 Bagging with three base classifiers: The first ensemble method (bagging) was applied using three base classifiers. The method was tested with various percentage of bagging size (70-100). Then, accuracies of each base classifier were computed as shown in Figure 2. It clearly shows that the accuracy of bagging with base classifier decision tree is much higher than naïve bayes and k-nearest neighbors.

Step 4 Boosting with three base classifiers: The second ensemble method (boosting) was applied using three base classifiers. The method was tested with various values of weight of threshold (70-100). Then the accuracy of boosting with base classifier decision tree is still the highest compared to other two classifiers as presented in Figure 3.

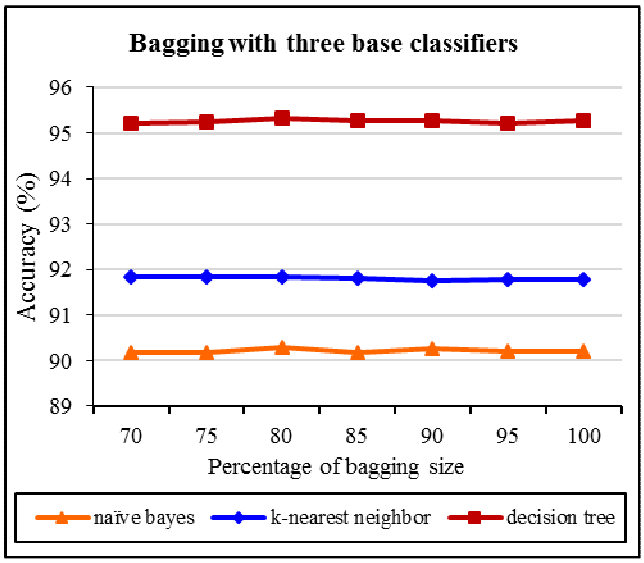

Fig. 2 Accuracies of bagging with three base classifiers

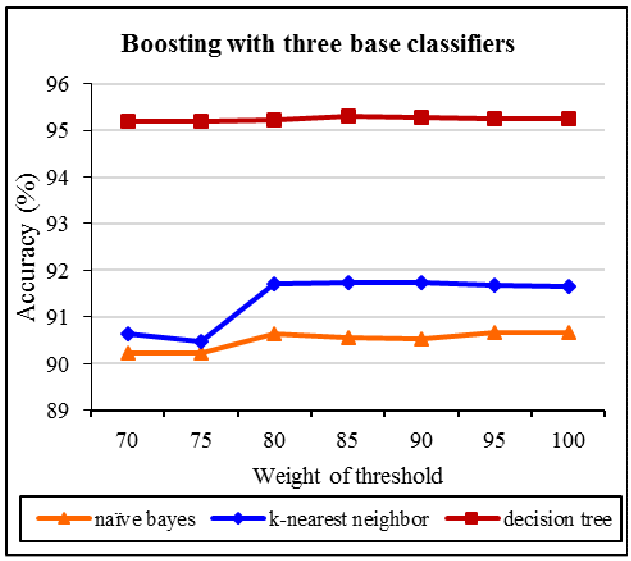

Fig. 3 Accuracies of boosting with three base classifiers

Step 5 Models Evaluation: Table 2 presents a comparison of accuracies obtained from the models of three base classifiers alone, accuracies obtained from bagging models with three base classifiers, and accuracies obtained from boosting models with three base classifiers. It can be seen that bagging method with base classifier decision tree has the highest accuracy (95.312%). As a whole, the ensemble methods give better performance than those of the base classifiers.

Table 2. Accuracy results.

Models	Base Classifiers (%)	Bagging with Base Classifiers (%)	Boosting with Base Classifiers (%)
naïve bayes	90.224	90.281	90.641
k-nearest neighbor	91.721	91.850	91.742
decision tree	**94.621**	**95.312**	**95.304**

Conclusions

This research evaluates accuracies of various diabetes classification models. The experiment used the data set of 48,763 records collected from Sawanpracharak Regional Hospital, Thailand. The gain ratio feature selection technique was firstly used for attribute ranking. It could reduce the number of predictor attributes from 15 to 13. Next, three base classifiers of naïve bayes, k-nereast neighbors and decision tree algorithms were applied. Then, all three algorithms were used as base classifier in bagging and boosting ensemble methods. The experimental results revealed that the bagging technique performed better than boosting technique and the base classifiers alone. The findings of this study are helpful in guiding the choice of classification algorithm for future applications. Other classification algorithms such as stacking method might be another aspect for further work.

Acknowledgment

The authors wish to thank Sawanpracharak Region Hospital, Thailand for the data set and Naresuan University for the financial support.

References

[1] Information on http://who.int/mediacentre/factsheets/fs312/en/index.html

[2] Ministry of Public Health, Surveillance Control and Prevention System of DM and HT in Thailand: Policy to Action, Nonthaburi, Thailand, 2013, p.6.

[3] J. Han, M. Kamber., J. Pei, Data Mining Concepts and Techniques, third ed., Morgan Kaufman, USA, 2012.

[4] X-H. Meng, Y-X. Huang, D-P. Rao, Q. Zhang and Q. Liu, Comparison of three data mining models for predicting diabetes or prediabetes by risk factors, Journal of Medical Sciences, 29, 93-99, (2013)

[5] D-Y. Yeh, C-H. Cheng and Y-W. Chen, A predictive model for cerebrovascular disease using data mining, Expert System with Application, 38, 8970-8977, (2011)

[6] N. Hosseinpour, S. Setayeshi, K. Ansari-asl and M. Mosleh, Diabetes Diagnosis by Using Computational Intelligence Algorithms, Journal of Advanced Research in Computer Science and Software Engineering, 2(12), 71-77, (2012)

[7] G. Liang, and C. Zhang, Empirical Study of Bagging Predictors on Medical Data, 9th Australasian Data Mining Conference, 121, 31-40, (2011)

[8] J. Abellán, Ensemble of decision tree based on imprecise probabilities an uncertainty measures, Information Fusion, 14,423-430, (2013)

[9] P. Giudici, S. Figini, Applied Data Mining for Business and Industry, second ed., Italy, 2009.

[10] J. Quinlan, Induction of decision tree, Readings in Machine Learning, 1986.

[11] A.L. Symeonidis., P.A. Mitkas, Agent Intelligence through Data Mining, Springer, New York, 2005.

[12] L. Breiman, Bagging Predictors, Machine Learning, 24(2), 123-140, (1996)

[13] I. Syarif, E. Zaluska, A. Prugel-Bennett and G. Wills, Application of Bagging, Boosting and Stacking to Intrusion Detection, MLDM2012, LNAI7376, 513-602, (2012)

[14] T.G. Dietterich, An Experimental Comparison of Three Methods for Construction Ensembles Of Decision Trees: Bagging, Boosting, and Randomization, Machine Learning, 40, 139-157, (2000)

[15] R.T. Schapire, The Boosting Approach to Machine Learning, An Overview. Nonlinear Estimation and Classification. Springer, New York, 2003.

[16] Y. Freund, R.E. Schapire, Experiments with a New Boosting Algorithm. Machine Learning, 13th International, New Conference, 1996.

[17] J. Xu and H. Li, AdaRank: A boosting algorithm for information retrieval, SIGIR'07, July 23-27, (2007)

[18] J. Sun, B. Liao and H. Li, AdaBoost and Bagging Ensemble Approaches with Neural Network as Base Learner for Financial Distress Prediction of Chinese Construction and Real Estate Companies, Recent Patents on Computer Science, 6, 47-59, (2013)