RANDOM FOREST ANALYSIS ON DIABETES COMPLICATION DATA

Punnee Sittidech, Nongyao Nai-arun
Department of Computer Science and Information Technology
Faculty of Science, Naresuan University, Phitsanulok 65000, Thailand
punnees@nu.ac.th, nongyao25@hotmail.com

ABSTRACT
This paper discusses how Random Forests, ensembles of weak decision trees, can be improved by excluding less important features from the model. Gain Ratio Feature Selection was used as the basis for tuning the algorithm parameters. Backwards elimination of the features to obtain the minimum subset with the highest accuracy was the key methodology of this experiment. The results of the proposed model were better in terms of accuracy and number of features used. The objective of this paper was to create a base-line, which will be useful for the classification on diabetes complications data. We recommend using the Random Forest with Feature Selection technique for other type of classification problems. Future work also includes an extension study of the different types of learning settings to improve the feature construction process.

KEY WORDS
Diabetes complications, classification, decision tree, bagging, random forest, feature selection

1. Introduction

Diabetes is caused by disorders of the body's insulin production, resulting in levels of blood sugar that are too high. Diabetic symptoms occur because the body cannot use glucose properly. For people who have diabetes, the body cannot use glucose efficiently; as a result, their blood sugar levels rise. In the long term, if not treated properly, this will result in the destruction of blood vessels and may lead to serious complications. The International Diabetes Federation (IDF) reported that over 371 million people have diabetes. However, 50% of people with diabetes are undiagnosed. In 2012, an estimated 4.8 million people died due to diabetes and over 471 billion USD were spent on healthcare for diabetics [1]. The World Health Organization (WHO) also reported that diabetes is a leading cause of serious health problems. Patients who lack knowledge about diabetes, combined with insufficient access to health services and essential medicines, can end up with complications such as blindness, amputation and kidney failure [2].

Several researchers have applied classification models to analyze medical data and this has led to a substantial amount of useful information. Classification is a technique used for discovering classes of unknown data. There are various methods for classification, such as Decision Trees, Rule Based, Neural Networks, etc. Decision tree supervised learning is one of the most popular methods because it is easy to understand and interpret by the end user [3, 4]. To improve the accuracy, ensemble methods such as Bagging can be applied by combining the results of induced classifiers with different training subsets. This methodology can be easily parallelized. These independent methods aim at either improving the predictive power of classifiers or decreasing the total execution time [5]. Each simple base classifier is trained on a sample set taken with a replacement from the training set. Then some form of voting is used to combine all base classified outputs [6]. Bagging ensemble classification improves predictive performance by using a randomized training subset, with replacement in all attribute predictors. Another ensemble classification, Random Forest, improves predictive performance by randomly selecting features in each decision split when building several decision trees and then determining the output from the out of bag result [6, 7]. In high dimensional and large quantities of raw data, Bagging and Random Forest usually give better improvement. Moreover, feature selection can help improve classification performance with minimal effort [8]. The basic idea of the algorithms is to search through all possible combinations of features in the data to find the subset of features that works best for prediction. The selection is done by reducing the number of features of the feature vectors, keeping the most meaningful and discriminating ones, while removing the irrelevant or redundant ones [9, 10, 11].

In this article, we concentrate on the classification performance of Decision Tree, Bagging with Decision Tree based classification and Random Forest with feature selection. The objective of this comparison is to create a base-line, which will be useful for the classification on diabetes complications data. The diabetes data set used in this experiment was collected from Sawanpracharak Regional Hospital, Thailand. The remainder of this paper is organized as follows. Overviews of classification techniques are introduced. Then, the experimental studies are presented and discussed. Finally, conclusions are drawn and further work is indicated.

DOI: 10.2316/J.2014.216.818-0047

2. Methodology

2.1 Decision Tree Model

Decision tree learning is one of the most popular methods in data mining classification because it is easy to understand. The model based tree was proposed by Quinlan [12]. Decision Tree is a supervised learning algorithm by using the data which the answers are already known and used for building the tree. Its quality is highly associated with the classification accuracy reached on the training data set, as well as the size of the tree [13]. Classification is an important task of assigning objects to one of several predefined categories. It is the process of modeling different data classes in training a data set to predict the class of objects or the expected value of unknown attribute [14, 15]. Training data sets used in classification includes attributes that can have discrete and/or continuous properties. The class label, on the other hand, must be a discrete attribute.

Decision tree classifier is a systematic approch to building classification models from a training data set. Decision tree structures are built or constructed in a top-down recursive divide-and-conquer strategy manner [12]. Its structure includes nodes and branches modeling from the training data. The algorithm will find the most powerful features that will be used to separate training data into two or more subsets based on the values of that feature. The first node is called the root node. Each data subset is then separated until a termination criterion is satisfied. The resulting decision tree consist of four primary features, which are (1) Root node: an attribute selected as the base to build the tree upon, (2) Internal node: attributes that resides on inner part of the tree, (3) Branches descending of a node: possible values for the attribute the branch initiates, and (4) Leaf nodes: the predefined classes. There are many exiting decision tree algorithms, including ID3, C4.5, and CART. Each technique employs different measures for selecting the best split in order to identify the most appropriate fit to build the tree [16].

2.2 Bagging

Bagging (bootstrap aggregating) is a well-known ensemble method introduced by Leo Breiman to reduce the variance of a predictor [5, 6]. It aims to increase accuracy by generating multiple versions of a predictor and using these to get an aggregated prediction. A training set is then generated by a random draw with the replacement of examples. Each of these data sets is used to train with base classifiers. The outputs of the models are combined to create a single output. Usually the aggregated prediction come from the predicted results that is the chosen most often (voting method) in case of categorical data. The aggregation averages over the versions in case of numerical data. The prediction can be obtained by changing the way that combines several classification results. Bagging usually produces a combined model that often performs better than the single model built from the original single data. It is easy to implement and has not too many parameter to tune. More classifiers trend to get more accuracy. The model is good even though the data is noisy [5, 17].

Bagging has been applied to a lot of research. Machová, et.al. [18] explored a bagging method on binary decision trees which enable an improvement of the classification performance. Ling and Sheng [19] investigated the performance of bagging in terms of learning from imbalanced medical data. Their experiment indicated that bagging performs better when using the base classifier decision tree.

2.3 Random Forest Model

Random Forest is a method of classification which is part of the ensemble learning model which combines predictions of weak classifiers. It was introduced by Leo Breiman [6, 7, 16, 20] and was widely believed to be the best classifiers for high-dimensional data. It builds a predictor ensemble with a set of decision trees that grow in randomly selected subspaces of data, where each tree in the ensemble is grown in accordance with a random parameter. It is fast and easy to implement, produces highly accurate predictions and can handle a very large number of input variables without over-fitting. Each tree in the collection is formed by selecting at random, at each node, a small group of input coordinates to split on and by calculating the best split based on these features in the training set. The tree is grown without pruning. This subspace randomization scheme is blended with bagging to resample with replacement the training data set each time a new individual tree is grown. These random trees are combined to form the aggregated regression estimated. Finally, predicted class label for unseen data by aggregating the predictions of the ensemble [6, 7, 20, 21].

2.4 Feature Selection Algorithms

Feature selection is one of the most important preprocessing steps in pattern classification. Its objective is to find a minimum set of attributes such that the resulting probability distribution of the data classes is as close as possible to the original distribution obtained using all attributes. Data mining on a reduced set of attributes has an additional benefit. It reduces the number of attributes appearing in the discovered patterns, helping to build the patterns easier [22, 23, 24]. It is also an effective dimensionality reduction technique and an essential preprocessing method to remove noise features. Therefore, it can reduce the cost of the classifier [25]. This is normally approached by searching the space of attribute subsets and evaluating each one. This is achieved by combining attribute subset evaluators with a search method [9]. Feature selection, when used along with any learning model, can help improve model performance with minimal effort. Hence, by selecting useful features from the data set, we essentially reduce the number of features or attributes needed for the classification problem of interest. There are many feature selection algorithms and also several approaches to evaluate the goodness of a feature subset. Huang et al. [26] used feature selection and

classification model construction on type 2 diabetic patients' data. Their results showed that feature selection via supervised model construction was used to rank the attributes affecting diabetes. In their experiment, Naïve Bayes processes the data fastest and C4.5 is the most stable classifier.

Gain Ratio Feature Selection is one of the feature selection techniques that was able to solve the drawback of information gain applied to attributes that can take on a large number of distinct values. The information gain measure prefers to select attributes having a large number of values. The information gain ratio is a modification of the information gain that reduces its bias by taking the number and size of branches into account when choosing the significant attributes [27]. Therefore, it is the ratio between the information gain and the intrinsic value. The attribute with the highest gain ratio is selected as the splitting attribute [3].

2.5 Model Evaluation

Evaluation is the processes to calculate the effectiveness of the results for data analysis models. Accuracy of the performance of a classification model is based on the count of the test records correctly and incorrectly predicted by the model. These counts are tabulated in a table known as a confusion matrix. In the case of two classes, the accuracy of a classifier on a given test set is the percentage of test set tuples that are correctly classified by the classifier. Accuracy is a popular evaluation performance of a classifier. Most classification algorithms seek models that attain the highest accuracy when applied to the test set.

3. Experimental Analysis

In this section, we demonstrate the use of a combination of data mining techniques to predict diabetes complications.

3.1 Data Set

In this study, all diabetes data were collected from Sawanpracharak Regional Hospital, which consisted of 27 Primary Care Units (PCU) during 2009-2013. The data consists of 7,498 instances, divided into 4 classes as follows: eye disease (1,918 instances), kidney disease (2,807 instances), heart disease (1,225 instances) and stoke diabetes (1,548 instances), and 18 input attributes as shown in Table 1. In this experiment, the original data set was divided into two subsets. One subset was used for training, while the other was used for testing to avoid over-fitting. Typically, the usual ratio for dividing the training set is 2:1. Therefore, the training data set and the test set consisted of 4,948 records and 2,550 records, respectively. After applying each technique, the accuracy was computed from the same test set and can also be used to compare the performance of different classifiers.

Table 1
Attributes Description

No	Attributes	Description	Values
1.	Sex	Sex	0: Male 1: Female
2.	Status	Status	0: Single, 1: Married
3.	Age	Age	Mean (66.29), S.D.(11.85), Min/Max (24/106)
4.	Creatinine	Creatinine	Mean (1.284), S.D.(1.24), Min/Max (0.1/3.4)
5.	Cholesterol	Cholesterol	Mean (191.86), S.D.(48.20), Min/Max (107/394)
6.	Triglycerides	Triglycerides	Mean (132.49), S.D.(10.08), Min/Max (27/262)
7.	HDL_C	High Density Lipoprotein Cholesterol	Mean (47.32), S.D.(15.65), Min/Max (35/84)
8.	LDL_C	Low Density Lipoprotein Cholesterol	Mean (94.07), S.D.(14.15), Min/Max (62/193)
9.	HbA1C	Hemoglobin A1c	Mean (9.64), S.D.(6.28), Min/Max (5.8/11..6)
10.	RDW	Red Cell Distribution width	Mean (13.46), S.D.(6.53), Min/Max (7/32)
11.	Sodium	Sodium	Mean (138.97), S.D.(4.59), Min/Max (108/179)
12.	Potassium	Potassium	Mean (4.04), S.D.(0.72), Min/Max (1.78/6.4)
13.	HGB	Hemoglobin	Mean (11.72), S.D.(2.19), Min/Max (10/46)
14.	HCT	Hematocrit	Mean (35.29), S.D.(6.42), Min/Max (7.2/59.7)
15.	MCV	Mean Cell Volume	Mean (54.97), S.D.(8.36), Min/Max (34.1/85.7)
16.	PLT	Platelet Count	Mean (165.21), S.D.(10.36), Min/Max (125/540)
17.	Chloride	Chloride	Mean (102.90), S.D.(5.81), Min/Max (67/143)
18.	CO2	Carbon Dioxide	Mean (24.157), S.D.(4.26), Min/Max (20/43)
19.	***CLASS***	***1 : Eye disease*** ***2 : Kidney disease*** ***3 : Stoke disease*** ***4 : Heart disease***	

3.2 Modeling

The training data set of 4,948 records, contains 18 input predictors, was used to model the Decision Tree, Bagging with Decision Tree based classification, and typical Random Forest. The results are shown in Table 2. It can be seen that Random Forest with 18 attributes yielded the best accuracy among the three classification models.

Table 2
Models Accuracies

Models	Accuracy
Decision Tree	91.683
Bagging with Decision Tree	93.213
Random Forest	93.840
Random Forest with Feature Selection	**94.743**

However, Random Forest utilizes the ensemble method by randomly selecting subsets of attributes to build decision trees. Thus, all 18 input predictors would have an equal chance to be in each predictor. Sometimes, input attributes may be irrelevant features, defined as those features not having any influence on the response classes. Therefore, we further analyzed the data set using feature selection algorithms to remove some irrelevant predictors from these 18 attributes.

The Gain Ratio feature selection algorithms were used in this paper. The results displayed in Figure 1 show a comparison of the ranked attributes of all the original 18 attributes with respect to the response features (diabetes complications). The ranking information was used to model our proposed Random Forest method with Feature Selection, as the following steps illustrate.

(a) Rank all variables according to a gain ratio ranking
(b) For each time (backward elimination),
Remove the last feature from the training data set
Rebuild the Random Forest model using only the remaining features.
(c) Select the features of the subset which maximizes prediction accuracy.

The classification results show that the Random Forest gave better results for the small number of attributes. From the results, the best percentage accuracy was (94.743%) by using the first 14 ranked attributes as the input attributes as shown in Figure 2.

3.3 Model Results

The proposed model, Random Forest using both Bagging and Feature Selection for tree building, contains 14 important features for four diabetes complications classification. All 14 features were transformed using min-max normalization. Then, each feature was computed for measures of central tendency, using a mode measurement for nominal scale and the median for interval scale data. These measures of central tendency values were used to form the class pattern of each Diabetes Complications into four patterns of each class as shown in Figure 3.

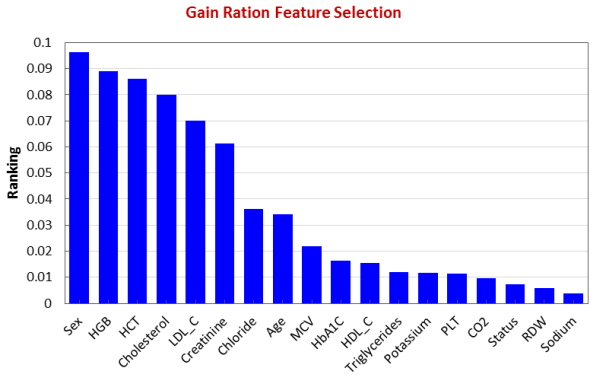

Figure 1. Gain Ration Feature Selection

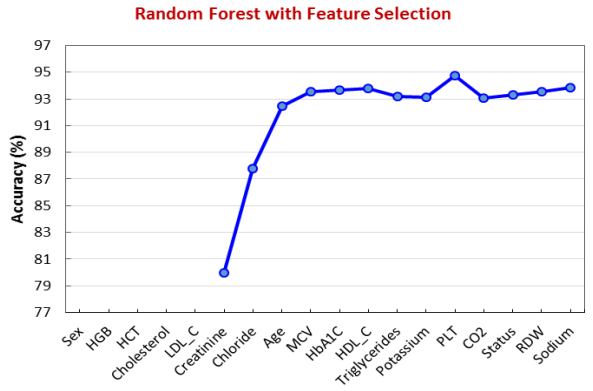

Figure 2. Random Forest with Feature Selection

4. Conclusion

In this paper, we have compared the classification results of using Decision Tree, Bagging with Decision Tree based classifier, Random Forest with all input attributes, and Random Forest with Feature Selection. The classification results showed that Random Forest with Feature Selection gave the best results. It can be concluded that the Random Forest with Feature Selection achieved increased classification performance. It also overcame the over-fitting problem generated due to missing values in the datasets. Therefore, for the classification problems, if one has to choose a classifier among the tree based classifier sets, we recommend using the Random Forest with Feature Selection. However, the data sets used in this work were small. More experiments to verify this conclusion are still needed. Possibilities for future work include an extension study of the different types of learning settings to review the feature construction process.

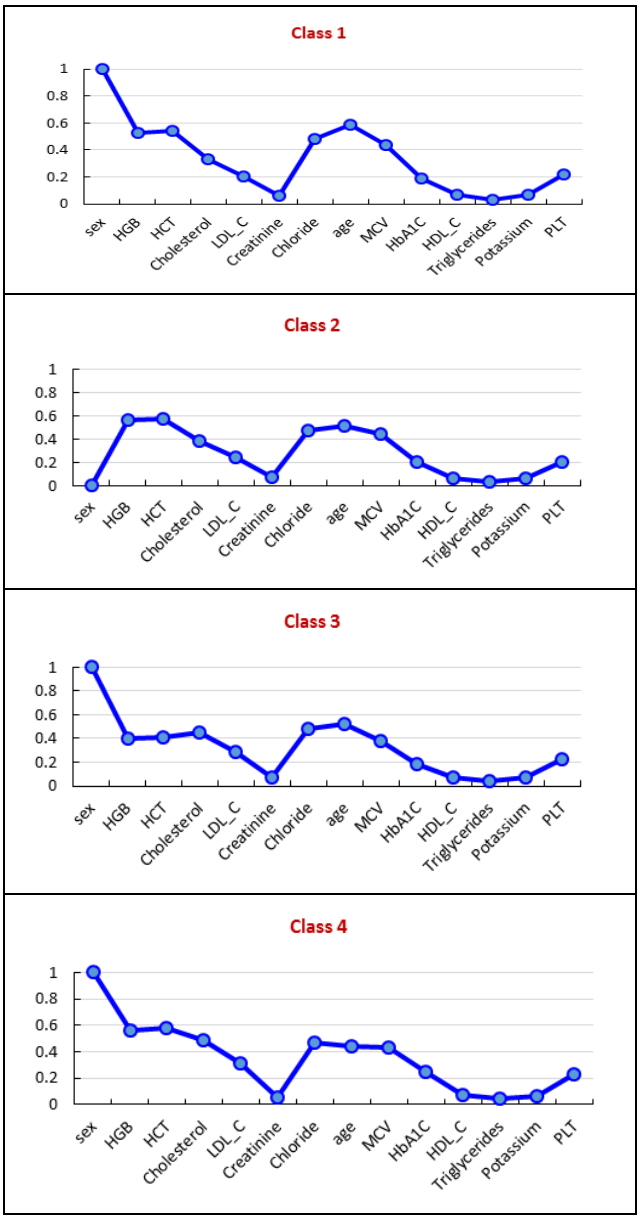

Figure 3. Class patterns of Diabetes Complications

Acknowledgments

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

References

[1] *International Diabetes Federation*. (2013). Retrieved March 12, 2013, Available: http://www.idf.org/diabetesatlas/5e/the-global burden.

[2] *World Health Organization*. (2014). Retrieved January 5, 2014, Available: http://www.who.int/diabetes/action_online/basics/en/index3.htm

[3] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques* 3rd ed (USA: Morgan Kaufman Publishers, 2012).

[4] P-N. Tan, M. Steinbach and V. Kumar, *Introduction to Data Mining* (Addison Wesley, 2006).

[5] L. Breiman, Bagging Predictors. *Machine Learning*, 24, 1996, 123-140.

[6] T. G. Dietterich, An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization, Machine Learning, 40, 2000, 139–157.

[7] L. Breiman, Random Forests. *Machine. Learning*. 45, 2001, 5–32. DOI 10.1023/A: 1010933404324

[8] R. A. Caruana and D. Freitag. How Useful is Relevance? Technical report, in *Fall'94 AAAI Symposium on Relevance*, New Orleans, 1994.

[9] K. Selvakuberanet, M. Indradevi, and R. Rajaram, Combined Feature Selection and classification: A novel approach for the categorization of web pages, *Journal of Information and Computing Science,* 3(2), 2008, 083-089.

[10] H. C. Yang and C. H. Lee, A Text Mining Approach on Automatic Generation of Web Directories and Hierarchies, *Proc. IEEE/WIC International Conference on Web Intelligence (WI'03)*, 2003.

[11] Y. Yimingand and O. P. Jan, Comparative Study of feature selection in Text Categorization, *Proc. 14th International Conference on Machine Learning (ICML'97),* 1997, 412-420.

[12] J. R. Quinlan, *Induction of Decision Tree* (Reading in Machining Learning, 1986).

[13] A. L. Symeonidis and P. A. Mitkas, *Agent Intelligence Through Data mining* (USA: Springer Science and Business Media, 2005).

[14] S. B. Kotsiantis, Supervised Machine Learning: A Review of Classification Techniques, *Informatica* , 31, 2007, 249-268.

[15] C. S. Sang, *Practical Applications of Data Mining* (USA: Jones & Bartlett Publishers, 2012).

[16] J. Ali, R. Khan, N. Ahmad and I. Maqsood, Random Forests and Decision Trees, *IJCSI International Journal of Computer Science Issues*, 9(5), No 3, 2012.

[17] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed (USA: Morgan Kaufmann Publishers, 2005).

[18] K. Machová, F. Barčák and P. Bednár, A Bagging Method using Decision Trees in the Role of Base Classifiers, in *Acta Polytechnical Hungarica*, 3(2), 2006.

[19] C. X. Ling and V. S. Sheng, Cost-Sensitive Learning and the Class Imbalance Proble, *Encyclopedia of Machine Learning, C.Sammut (Ed.)*, Springer, Canada. 2008.

[20] G. Biau, Analysis of a Random Forests Model, *Journal of Machine Learning Research,* 13, 2012, 1063-1095.
[21] P. Geurts et al., Proteomic mass spectra classification using decision tree based ensemble methods, *Bioinformatics,* 21(15), 2005, 3138–3145.
[22] S. Chakrabarti, E. Cox, E. Frank, R. H. Guting, J. Han, X. Jiang, M. Kamber, S. Lightstone, S. Nadeau, T. P. Neapolitan, R. E. Pyle, D. Refaat, M. Schneider, T. J. Teorey, and I. H. Witten. *Data Mining: Know It All* (USA: Morgan Kaufmann Publishers, 2008).
[23] A. G. K. Janecek, W. N. Gansterer, M. A. Demel, and G. F. Ecker, On the Relationship Between Feature Selection and Classification Accuracy, *JMLR Workshop and Conference*, 4, 2008, 90-105.
[24] L. Ladha and T. Deepa, Feature Selection Methods and Algorithms, 2011.
[25] B. Krishnapuram et al, A Bayesian Approach to Joint Feature Selection and Classifier Design, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9), 2004, 1105 – 1111.
[26] Y. Huang, P. McCullagh, N. Black, and R. Harper, Feature Selection and Classification Model Construction on type 2 Diabetic Patients' data, *Artificial Intelligence in Medicine*, 41, 2007, 251-262.
[27] G. K. Asha, A. S. Manjunath, and M. A. Jayaram, Comparative Study Of Attribute Selection Using Gain Ratio And Correlation Based Feature Selection, *International Journal of Information Technology and Knowledge Management*, 2(2), 2010, 271-277.