Journal of Computer Science 9 (7): 883-888, 2013
ISSN: 1549-3636
© 2013 Science Publications
doi:10.3844/jcssp.2013.883.888 Published Online 9 (7) 2013 (http://www.thescipub.com/jcs.toc)

Using Feature Selection as Accuracy Benchmarking in Clinical Data Mining

Jafreen Hossain, Nor FazlidaMohdSani, Aida Mustapha and Lilly SurianiAffendey

Department of Computer Science, Faculty of Computer Science and Information Technology,
Universiti Putra Malaysia, 43400 UPM Serdang, Selangor DarulEhsan, Malaysia

Received 2013-02-27, Revised 2013-04-18; Accepted 2013-06-20

ABSTRACT

Automated prediction of new patients' disease diagnosis based on data mining analysis on historical data is proven to be an extremely useful tool in the medical innovation. There are several studies focusing on this particular aspect. The objective of this study is two-fold. First, we look into three different classifiers, which are the Naïve Bayes, Multilayer Perceptron (MLP) and Decision Tree J48 to predict the diagnosis results. Next, we investigate the effects of feature selection in such experiments. We also compare the experimental results with the study of Comparative Disease Profile (CDP) using the same dataset. Results have shown that the Naive Bayes provides the best result in terms of accuracy in our experiments and in comparison with CDP. However, we suggest using Multilayer Perceptron since the variables used in our experiments are inter-dependent among each other. In addition, MLP has shown better accuracy than CDP.

Keywords: Data Mining, Healthcare, Heart Disease, Multilayer Perceptron, Naive Bayes, J48

1. INTRODUCTION

Data mining is the process of finding previously unknown patterns and trends in databases and using that information to build predictive models. Data mining in medical science is critical and is more sensitive than other domains because of its complexity of nature. On the other hand the significance of data mining in medical science can play a vital role if it is utilized for prediction and decision making. Healthcare industry today generates large amount of complex data about patients, hospitals resources, disease diagnosis, electronic patient records or medical devices. The large amount of data is a key resource to be processed and analyzed for knowledge extraction that enables support for cost-savings and decision making (Bhatla and Jyoti, 2012). Data mining provides a set of tools and techniques that can be applied to this processed data to discover hidden patterns and also provides healthcare professionals an additional source of knowledge for making decisions.

As has been highlighted in Wei and Altman (2004), historical clinical data is the critical source to support information to help diagnosis of patient's disease. They propose a Comparative Disease Profile (CDP), which is a set of distinguished features derived from historical medical dataset. Once established, CDP is claimed to have helped the process of manual decision making by providing useful diagnosis guidelines. Motivated by their work, this study focuses on classification approach for diagnosis of cardiac patients. The datasets were sourced from the Cleveland Heart Disease Datasets of UCI Repository of Machine Learning databases and domain theory which is available for download at: http://archive.ics.uci.edu/ml/datasets/Heart+Disease.

The remaining of this study proceeds as follows. The second part of this studydescribes the methods and techniques used, whereas the following sections discussed the experiments and results respectively.

2. MATERIALS AND METHODS

In this study, three classification algorithms are chosen for the purpose of accuracy benchmarking in clinical data, which are the Naïve Bayes, Multilayer Perceptron (MLP) and Decision Tree J48.

Corresponding Authors: Jafreen Hossain, Department of Computer Science, Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor DarulEhsan, Malaysia

A naive Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model". In simple terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature, given the class variable (Han *et al*., 2011).

A Multilayer Perceptron (MLP) is a feed-forward artificial neural network model that maps sets of input data onto a set of appropriate output. An MLP model consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one. Except for the input nodes, each node is a neuron (or processing element) with a nonlinear activation function. MLP utilizes a supervised learning technique called back-propagation for training the network. MLP is a modification of the standard linear perceptron and can distinguish data that is not linearly separable.

J48 is an open source Java implementation of the C4.5 algorithm in the Weka data mining tool. C4.5 is an algorithm used to generate a decision tree developed by Quinlan (1993). C4.5 is an extension of Quinlan's earlier ID3 algorithm. The decision trees generated by C4.5 can be used for classification and for this reason; C4.5 is often referred to as a statistical classifier.

To investigate further the classifier performance in accuracy benchmarking, this study also looks into feature selection algorithm via Weka filtering method called the attribute select classifier to reduce the dimensionality of the data. This limits the number of attributes by choosing the ones that are more likely to impact the target class label. However, in principle there is no guarantee that feature selection will yield a result better than that with the full attribute range.

To measure the performance, this study focuses on Receiver Operating Characteristics (ROC) area to compare the accuracy of the different classifiers. ROC graph organizes classifiers and helps visualize their performance. ROC graphs are commonly used in medical decision making and in recent years have been used increasingly in machine learning and data mining research (Robin *et al*., 2011). Basically, ROC is a two-dimensional graph in which true positive is plotted on the Y-axis and false positive is plotted on the X-axis. The classifier that is nearest to the perfect point (0, 1) or the top left corner in the graph shows the best accuracy.

2.1. Experiments

In this study, we set up a series of classification experiments focusing three algorithms in Weka 3.7.4 data mining tool (Hall *et al*., 2009), which are Naïve Bayes, Multilayer Perceptron (MLP) and Decision Tree J48. The task is to predict and diagnose cardiac patients based on the given symptoms and information from the Cleveland Heart Disease Dataset.

This dataset contains 13 attributes and one class variable "Label" that is used to categorize between 'sick' and 'not-sick'. The 13 attributes are all numeric and they are: age, sex, Chest pain type (Cp), resting blood pressure (Trestbps), serum Cholesterol (Chol), Fasting blood sugar (Fbs), resting electrocardiographic results (Restecg), maximum heart rate achieved (Thalach), the occurrence of Exercise induced angina (Exang), ST depression induced by exercise relative to rest (Oldpeak), slope of peak exercise ST segment (slope), number of major vessels colored by fluoroscopy (Ca) and thal. **Table 1** shows the attributes and descriptions on the Cleveland data.

Table 1. Attributes and descriptions

Name of attributes	Data type and Description
Age	Age in years
Sex	Sex (1 = male; 0 = female)
Cp	Chest pain type -- Value 1: typical angina -- Value 2: atypical angina -- Value 3: non-anginal pain -- Value 4: asymptomatic
Trestbps	Resting blood pressure (in mm Hg on admission to the hospital)
Chol	Serum cholesterol in mg/dl
Fbs	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
Restecg	Resting electrocardiographic results -- Value 0: normal -- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) -- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
Thalach	Maximum heart rate achieved
Exang	Exercise induced angina (1 = yes; 0 = no)
Oldpeak	ST depression induced by exercise relative to rest
Slope	The slope of the peak exercise ST segment
Ca	Number of major vessels (0-3) colored by fluoroscopy
Thal	3 = normal; 6 = fixed defect; 7 = reversible defect

For the given dataset, the collection of attribute values from patients that belong to the sick group forms the distribution which presents the number of sick patients. Similarly, another distribution presents the patients who are not sick.

The experiments were carried out in two stages. The first stage is to measure the benchmark performance of Naïve Bayes, Multilayer Perceptron and J48 classifiers. The ROC areas were observed and recorded. Next in the second stage, feature selection was added to the experiments before the classification task. The ROC area were again observed and compared. Finally, the results from the second stage were then compared with findings by Comparative Disease Profile (CDP) (Wei and Altman, 2004).

3. RESULTS

The experimental results are reported in two parts, before and after feature selection is applied.

3.1. Benchmark Results

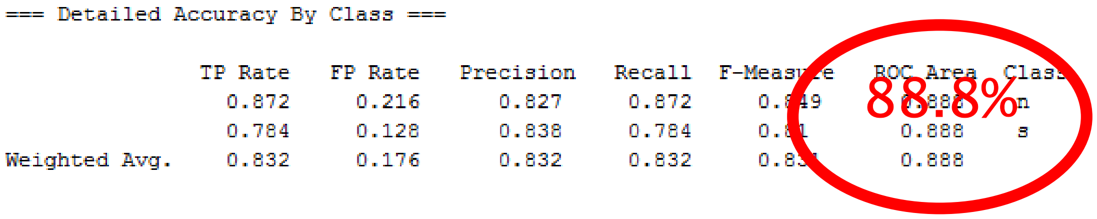
From the observation, the average of ROC area using Naïve Bayes was 88.8%, whereas 82.4% and 78.7% for Multilayer Perceptron and J48 respectively.

Figure 1-3 shows the benchmark results before feature selection.

3.2. After Feature Selection

Then the attribute select classifier was applied to find the best attributes while expecting better ROC area. Feature selection stage returned seven best attributes after using the attribute select classifier. They are Cp, Restecg, Thalach, Exang, Oldpeak, Ca and Thal.

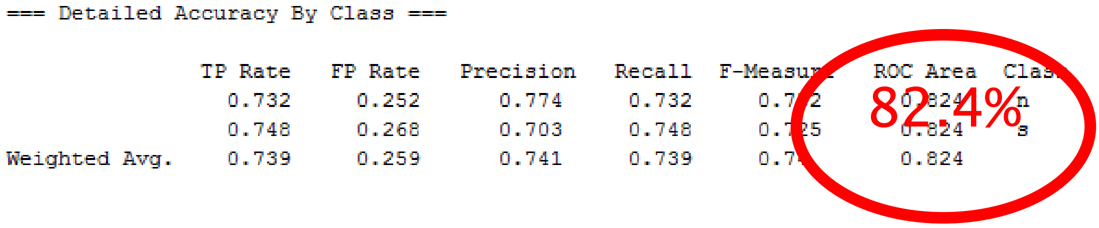
The same classifiers were used to this seven selected attributes and the ROC area were observed again. Interestingly the Naïve Bayes shows a little lower ROC area this time than using the 13 attributes. On the other hand the average of ROC area increased significantly after using Multilayer Perceptron and J48 to 86 and 79.1% respectively. **Figure 4-6** shows the results.



```
=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   0.872     0.216     0.827      0.872     0.849      0.888    n
                   0.784     0.128     0.838      0.784     0.811      0.888    s
Weighted Avg.      0.832     0.176     0.832      0.832     0.831      0.888
```

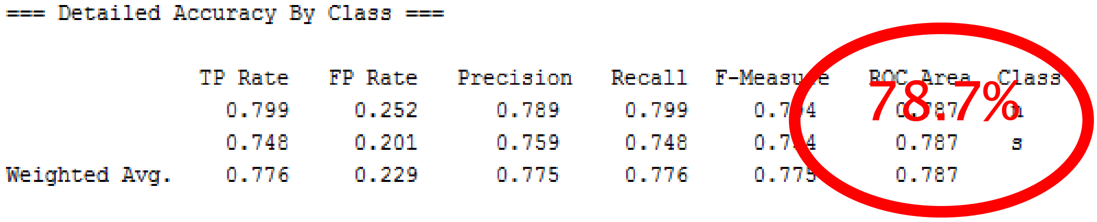
Fig. 1. Accuracy with Naïve Bayes



```
=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   0.732     0.252     0.774      0.732     0.752      0.824    n
                   0.748     0.268     0.703      0.748     0.725      0.824    s
Weighted Avg.      0.739     0.259     0.741      0.739     0.739      0.824
```

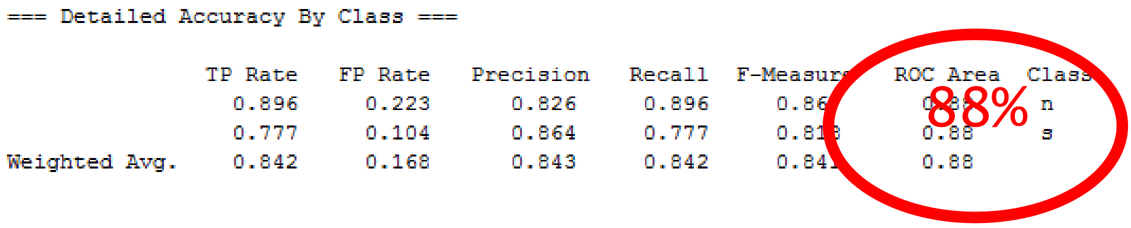
Fig. 2. Accuracy with Multilayer Perceptron



```
=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   0.799     0.252     0.789      0.799     0.794      0.787    n
                   0.748     0.201     0.759      0.748     0.754      0.787    s
Weighted Avg.      0.776     0.229     0.775      0.776     0.775      0.787
```

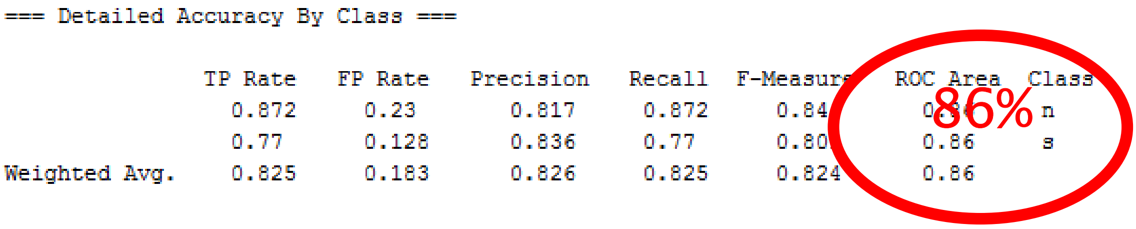
Fig. 3. Accuracy with J48



```
=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   0.896     0.223      0.826     0.896     0.86       0.88     n
                   0.777     0.104      0.864     0.777     0.818      0.88     s
Weighted Avg.      0.842     0.168      0.843     0.842     0.841      0.88
```

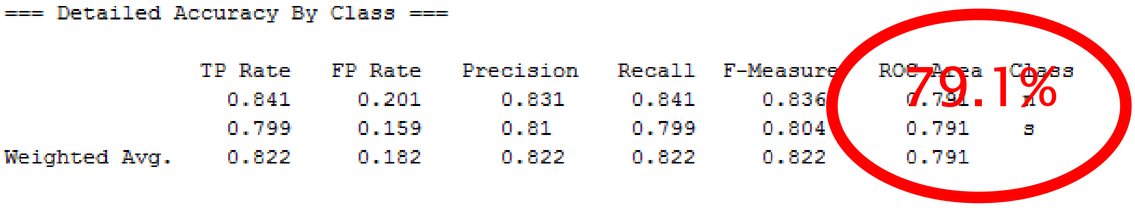
Fig. 4. Accuracy with Naïve Bayes (after feature selection)



```
=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   0.872     0.23       0.817     0.872     0.84       0.86     n
                   0.77      0.128      0.836     0.77      0.80       0.86     s
Weighted Avg.      0.825     0.183      0.826     0.825     0.824      0.86
```

Fig. 5. Accuracy with Multilayer Perceptron (after feature selection)



```
=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   0.841     0.201      0.831     0.841     0.836      0.791    n
                   0.799     0.159      0.81      0.799     0.804      0.791    s
Weighted Avg.      0.822     0.182      0.822     0.822     0.822      0.791
```

Fig. 6. Accuracy with J48 (after feature selection)

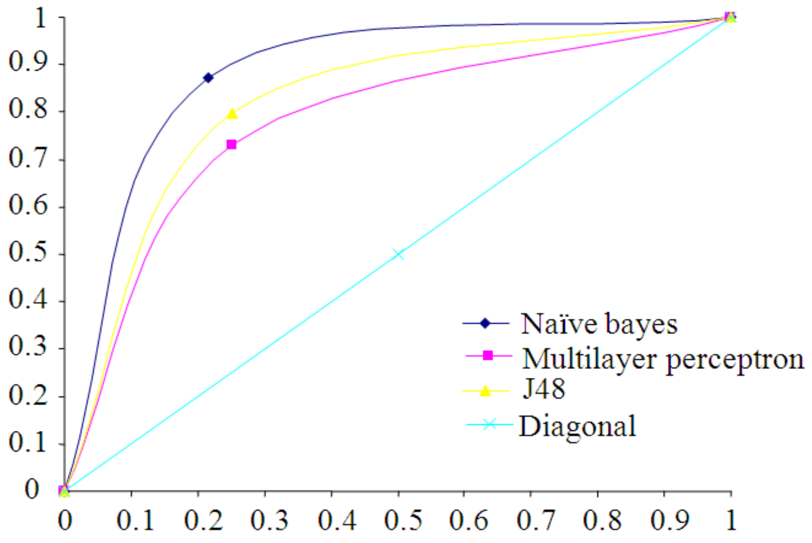


Fig. 7. ROC before attribute select classifier

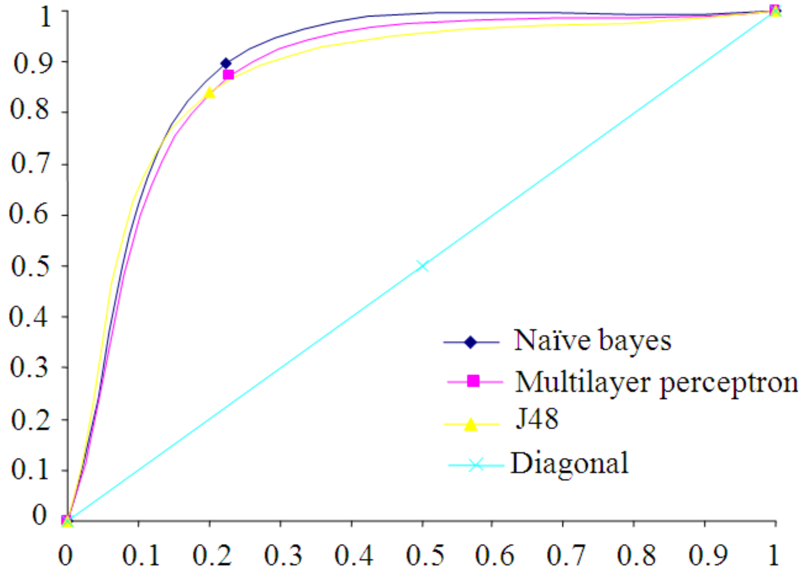


Fig. 8. ROC after attribute select classifier

3.3. ROC Area

The ROC areas below serve to present the comparative performance across three proposed classifiers. Without applying the Weka filtering feature of attribute select classifier, the results are clearly in favor of Naïve Bayes. However, using attribute select classifier, changes the comparison more in favor of the other two classifiers, as they inch forward towards Naïve Bayes. Naïve Bayes actually deteriorates a little. **Figure 7 and 8** shows the comparison of ROC areas between two experiments.

Even with improvement in accuracy of Multilayer Perceptron and J48 algorithms, as well as deterioration in Naïve Bayes accuracies, Naïve Bayes still remains the best classifier in terms of accuracy.

4. DISCUSION

Though Naïve Bayes is showing better results in our experiment, we suggest using Multilayer Perceptron since Naïve Bayes algorithm assumes independency among variables whereby in real-life situations the variables are inter-dependent among each other. We also suggest to use Multilayer Perceptron classification algorithm together with the filtering method of attribute select classifier in Weka, which resulted a significant increase in accuracy from 82.4 to 86%.

Next, this study compares the findings with the CDP accuracy (Wei and Altman, 2004). The result in the CDP study shows an accuracy of 82.2%, which is better than the performance of our J48 classifier. However, our proposed MLP shows a better result than the CDP after using the attribute select classifier.

5. CONCLUSION

In this study, we have used three different classification algorithms in a data mining tool, Weka (Hall *et al*., 2009) using the standard Cleveland heart data sets and compared the accuracy level of each method. We also compared the results of our experiments with CDP system developed by Wei and Altman (2004). It has been observed that the Naïve Bayes shows the best result in terms of accuracy in our experiment and in comparison with CDP. However, we suggest to use Multilayer Perceptron since the variable used in our experiments are inter-dependent among each other. In addition, MLP has shown better accuracy than CDP. In the future work, we hope to investigate further on attributes from other medical dataset.

6. ACKNOWLEDGMENT

Special thanks from author to financial support (Prototype Research Grant Scheme, PRGS) received from the Ministry of Higher Education (MoHE), Malaysia via Universiti Putra Malaysia. The principle investigator of the research project is Dr. NorFazlidaMohdSani.

7. REFERENCES

Bhatla, N. and K. Jyoti, 2012. An analysis of heart disease prediction using different data mining techniques. Int. J. Eng. Res. Technol., 1: 1-4.

Hall, M., E. Frank, G. Holmes, B. Pfahringer and P. Reutemann *et al*., 2009. The WEKA data mining software: An update. ACM SIGKDD Exp. Newsletter, 11: 10-18. DOI: 10.1145/1656274.1656278

Han, J., M. Kamber and J. Pei, 2011. Data Mining: Concepts and Techniques. 3rd Edn., Elsevier Inc., Burlington, ISBN-10: 0123814804, pp: 744.

Quinlan, J.R., 1993. Programs for Machine Learning. 5th Edn., Morgan Kaufmann Publishers, San Mateo, CA., ISBN-10: 1558602380, pp: 302.

Robin, X., N. Turck, A. Hainard, N. Tiberti and F. Lisacek *et al*., 2011. pROC: An open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinform., 12: 77-77. DOI: 10.1186/1471-2105-12-77

Wei, L. and R.B. Altman, 2004. An Automated System for Generating Comparative Disease Profiles and Making Diagnoses. IEEE Trans. Neural Netw., 15. 597-597.