International Journal of Engineering & Technology, 7 (4.36) (2018) 845-849

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A Comparative Evaluation of Meta Classification Algorithms with Smokers Lung Data

K. Kavitha[1*], K. Rohini[2], G. Suseendran[3]

[1]Research Scholar, Department of Computer Science, School of Computing Sciences, Vels Institute of Science Technology and Advanced Studies (VISTAS), Chennai, Tamil Nadu, India.
[2]Asst.Prof, Department of Information Technology, School of Computing Sciences, Vels Institute of Science Technology and Advanced Studies (VISTAS), Chennai, Tamil Nadu, India. E-mail:rrohini16@gmail.com
[3]Assistant Professor, Department of Information Technology, School of Computing Sciences, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai. Email:suseendar_1234@yahoo.co.in
**Corresponding author E-mail:kavithakannan91@gmail.com*

Abstract

Data mining is the course of process during which knowledge is extracted through interesting patterns recognized from large amount of data. It is one of the knowledge exploring areas which is widely used in the field of computer science. Data mining is an inter-disciplinary area which has great impact on various other fields such as data analytics in business organizations, medical forecasting and diagnosis, market analysis, statistical analysis and forecasting, predictive analysis in various other fields. Data mining has multiple forms such as text mining, web mining, visual mining, spatial mining, knowledge mining and distributed mining. In general the process of data mining has many tasks from pre-processing. The actual task of data mining starts after the preprocessing task. This work deals with the analysis and comparison of the various Data mining algorithms particularly Meta classifiers based upon performance and accuracy. This work is under medical domain, which is using the lung function test report data along with the smoking data. This medical data set has been created from the raw data obtained from the hospital. In this paper work, we have analyzed the performance of Meta classifiers for classifying the files. Initially the performances of Meta and Rule classifiers are analyzed observed and found that the Meta classifier is more efficient than the Rule classifiers in Weka tool. The implementation work then continued with the performance comparison between the different types of classification algorithm among which the Meta classifiers showed comparatively higher accuracy in the process of classification. The four Meta classifier algorithms which are widely explored using the Weka tool namely Bagging, Attribute Selected Classifier, Logit Boost and Classification via Regression are used to classify this medical dataset and the result so obtained has been evaluated and compared to recognize the best among the classifier.

Keywords: *Data mining, Weka, meta classifier, lung function test, bagging, attribute selected classifier, logit boost, classification via Regression.*

1. Introduction

Data mining is that the extraction of hidden forebear unknown and almost certainly helpful data from knowledge. The concept is to create programs that sift through databases, mechanically seeking regularities or patterns. Sturdy patterns if found can doubtless generalize to create correct predictions on future knowledge. The amount of data in the medical field has increased tremendously. Although, such a large volume of information is valuable and need to be analyzed for further forecast perceive and predict the complexities that may arise in future. Machine learning algorithms are widely used to analyze and process various kinds of data [1].

Now-a-days several organizations apply data mining intensively and expansively to a vast amount of data by utilizing various machine learning algorithms at its different stages. This data mining process has been done various domains [2]. In-healthcare, data mining is becoming progressively more accepted [3]. Data mining gives the methodology and technology to recognize the valuable information of data for higher cognitive process. There are various tools available to apply the machine learning algorithms to different kinds of data set. These tools allow performing various kinds of tasks from preprocessing till visualization of the obtained result. In this work the Weka tool is used for implementing and evaluating the machine learning algorithms using the medical dataset.

2. WEKA -waikato environment for knowledge analysis

"Weka is a suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. It is free software licensed under the GNU General Public License" as referred in Wikipedia. It is a collection of machine learning algorithms for performing various data mining tasks. These algorithms work when applied directly to a dataset or called from our own Java code. Weka gives user friendly environment which contains tools for processing the data and providing knowledge. It starts from data pre-processing, classification, regression, clustering, association rules, and visualization. Weka has various applications such as the Explorer, Experimenter, Knowledge Flow, Work Bench and Simple CLI. It also contains package manager which can be used to install various learning schemes and tools.[4]

3. The proposed methodology

The leading objective of this research work is to find the best classification algorithm among various Meta classifiers.

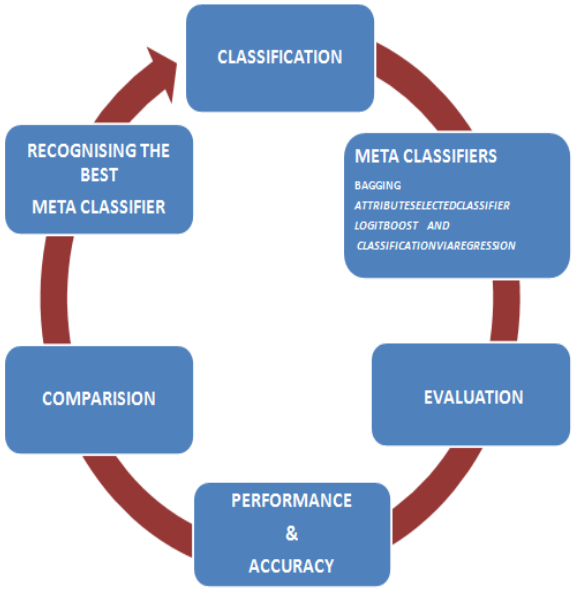


Fig. 1: Methodology of the work

The Methodology of the work is shown in the fig 1.
The steps in proposed work are as following:

- To apply different type of classification techniques such Rule and Meta algorithms using Weka tool.
- Comparing both the experimental results of all these types of classification techniques to move forward in to next level of implementation.
- The various Meta classifies shows higher accuracy than the Rule based classifiers dataset.
- Comparative analysis of results using parameters accuracy, execution time and error rate for Bagging, Attribute Selected Classifier, Log it Boost and Classification via Regression.
- Evaluation of results produced by Meta classifiers.
- Find the best classifier to build an improved classification model with maximum performance and accuracy using the medical data set.

3.1. Classification

The classification is an important data mining technique with wide spread applications [5]. It is used to categorize each item in a set of data into one of predefined set of models or classes. Classification algorithm plays an important role to identify the class name for the unknown item set. In this research, we have analyzed two classifiers namely Meta Learning and Rule based. In Meta classifier, we have analyzed and compared four classification algorithms namely BAGGING, LOGITBOOST, ATTRIBUTE SELECTED CLASSIFIER and CLASSIFICATION VIA REGRESSION.

3.2. Meta-classifier algorithms

Meta learning algorithms acquire classifiers and enhance them into additional powerful learners and create either classification model or regression model. One parameter specifies the bottom classifier others specify the quantity of iterations for schemes resembling bagging and boosting. Meta classification specifies the procedure of combination of many classifiers in which the results of base classifiers are integrated and final results are obtained.

LOGITBOOST: Log it Boost Meta-learning classification algorithm is an expansion of Adaboost algorithm. "It changes the exponential loss of Adaboost algorithm to conditional Bernoulli possibility loss". [5].

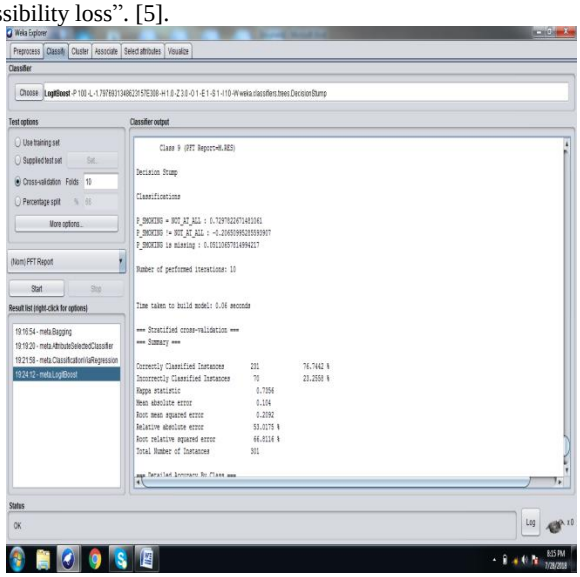


Fig. 2: experimental screen

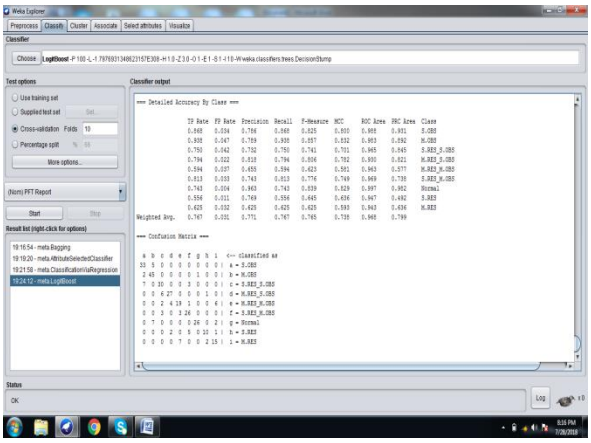


Fig. 3: experimental screen

The Figure2 shows the experimental screen shots which depicts the time taken to build the model and the Correctly classified instances which is76.74%, which is the lowest among the all the four Meta algorithms implemented. The confusion matrix is in figure 3 also shows accuracy and performance of the classifier.

3.3. Bagging

Bagging uses bootstrap aggregation and generates bootstrap samples of the training data. It builds the distinctive training set consists of frequent and plentiful data sets. Classification results are based upon the highest number of votes [6].

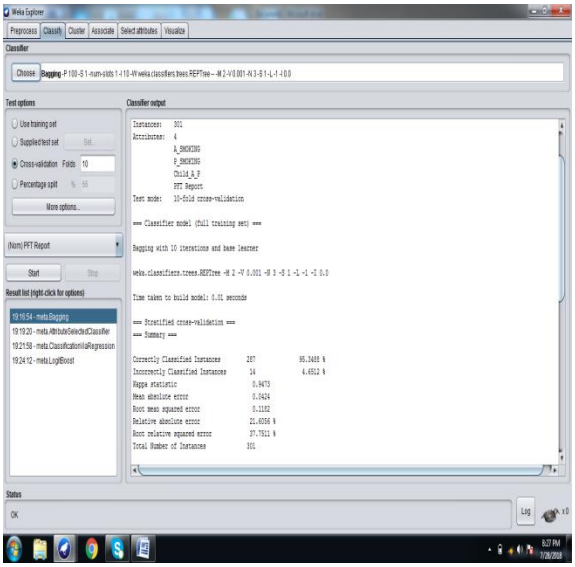


Fig. 4: experimental screen

3.4. Bagging vs boosting

Variation is reduced and performance is enhanced for unsteady classifiers which diverge significantly with small changes in the dataset [7].

Meta-Bagging algorithm is very much effective for this medical dataset than boosting algorithm such as AdaBoost and Logitboost. In this work Logitboost shows considerably higher performance than Adaboost. The Figure 4 shows the experimental screen shots which depicts the time taken to build the model and the correctly classified instances which is 95.34%. The confusion matrix is in figure 5 also shows accuracy and performance of the classifier.

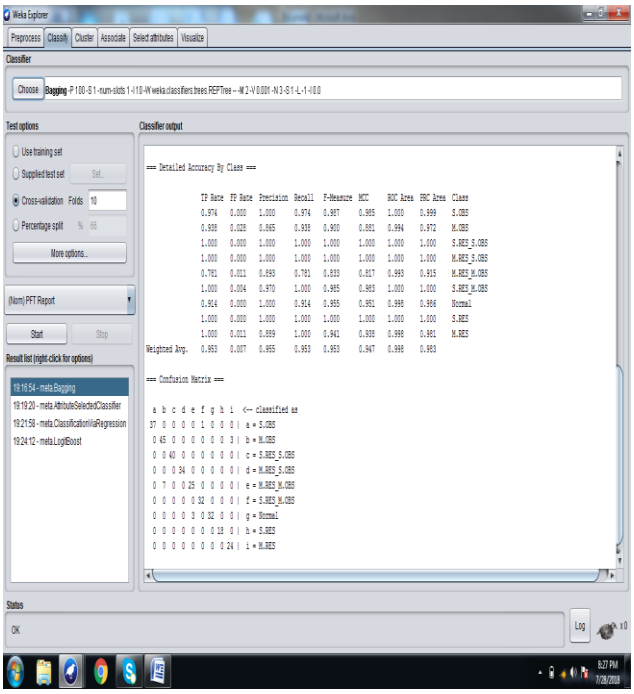


Fig. 5: experimental screen

3.5. Attribute selected classifier

In Attribute Selected Classifier the dimensionality of training and test data is reduced by attribute selection before being passed on to a classifier. It is options specific to classifier in tree classifier J48 in weka. The J48 is an un pruned tree [8].

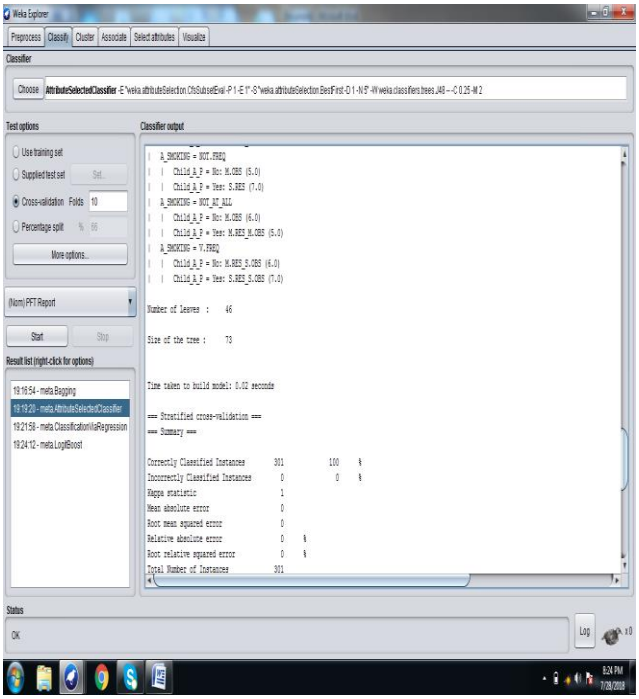


Fig. 6: experimental screen

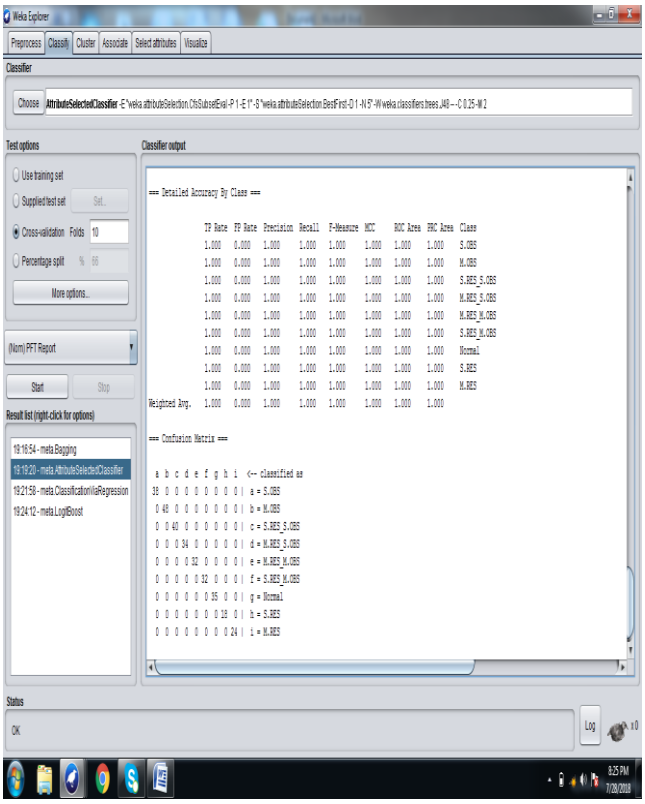


Fig. 7: experimental screen

Attribute Selected Classifier is a public class which extends Single Classifier Enhancer and implements Option Handler, Additional Measure Producer, Weighted Instances Handle etc.

The range of the training data and testing data is minimized before the actual classification process. The classifier lifts various exploring approaches during the attribute selection process [9].

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

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     S.OBS
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     M.OBS
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     S.RES_S.OBS
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     M.RES_S.OBS
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     M.RES_M.OBS
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     S.RES_M.OBS
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     Normal
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     S.RES
               1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     M.RES
Weighted Avg.  1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000
```

Fig. 8: experimental screen

The Figure 6 shows the experimental screen shots which depicts the time taken to build the model and the Correctly classified instances which is100%, which is the highest among the all the four Meta algorithms implemented. The confusion matrix is in figure 7 also shows accuracy and performance of the classifier.
The Figure 8 shows the detailed accuracy by the class model by weighted average. This classification algorithm shows the best accuracy than other three algorithms.

3.6. Classification via regression

Regression approaches are applied for classification under this classifier. In this the public class Classification via Regression Extends the Single Classifier Enhancer and implements the Technical Information Handler. This is for doing classification using regression methods. Class is binarized and one regression model is built for each class value [10].
The Figure 9 shows the experimental screen shots which depicts the time taken to build the model and the Correctly classified instances which is 97.34%. The confusion matrix is in figure 10 also shows accuracy and performance of the classifier.

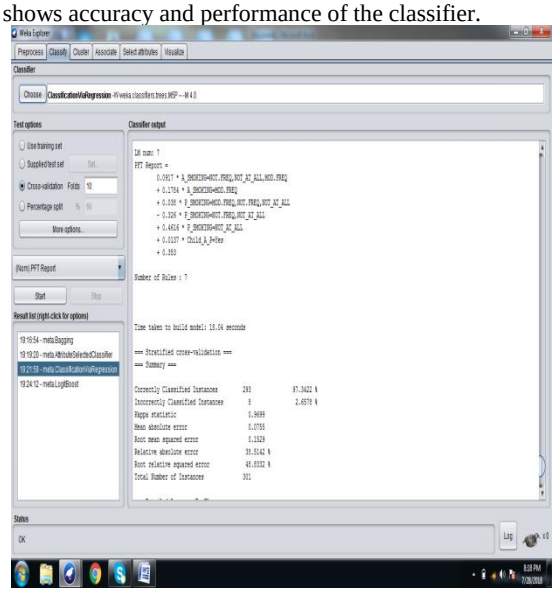


Fig. 9: experimental screen

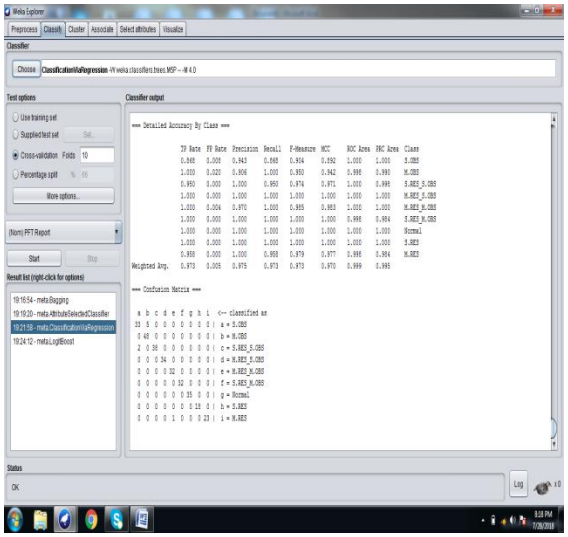


Fig. 10: experimental screen

4. Data set

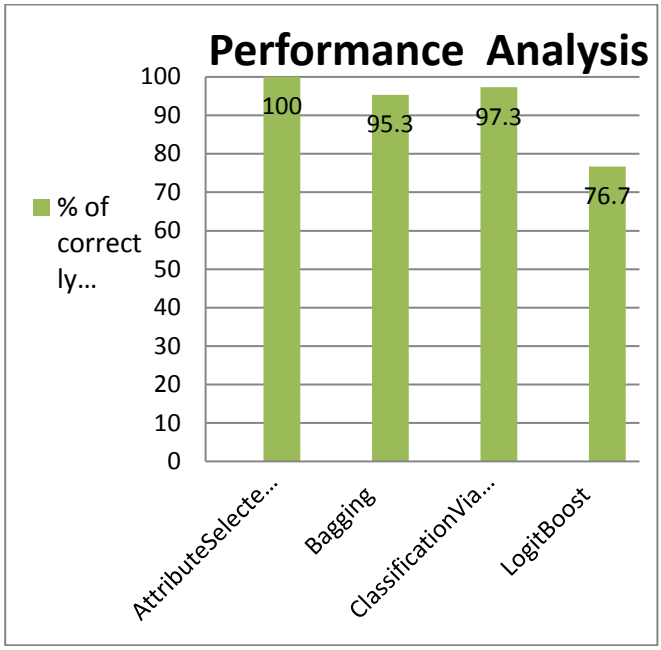
This work, have used medical dataset. This medical data set consists of two types of data. One among them is lung function test report data and another is about the smoking habits of the respective patients. The raw dataset is collected, processed and stored in .csv file format. The following table provides the details of the dataset.

Table 1: Parameters of Lung Function Data Sets

No.	Parameters	Descriptions
1	Age	Taken as numeric value (years)
2	Childhood Smoking	Taken Boolean values 1. Yes, 2. No
3	Active smoking	Taken 5 Scale categorical nominal input values.
4	Passive smoking	Taken 5 Scale categorical nominal input values.
5	Lung function Test report	Taken 9 category of nominal values

5. Experimental evaluation

This work experiments various Meta-learning classification algorithms and analyses its performance and accuracy. Main aim of this proposed work is to analyze the classification algorithms' performance and accuracy as shown in the chart and screen shots of the experimental page and result obtained. The WEKA application is used for the performance evaluation. All the four classifiers namely Bagging, Attribute Selected Classifier, Logit Boost and Classification via Regression algorithms are experimented and evaluated with 10 fold cross validation. Among these Attribute Selected Classifier, Classification via Regression and bagging algorithms shows more than 95% of performance accuracy. But the best performance accuracy has been secured by the Attribute Selected Classifier as shown in the graph. The graph describes the percentage of correctly classified instances by all the four algorithms for this medical dataset.



Graph 1: experimental screen

6. Conclusion

The classification is one of the most important and very useful tasks in the field of data mining and data analytics. Model creation by finding the interesting patterns provides useful knowledge acquired from large data repositories. This work examines and evaluates the performance accuracy of four different Meta classification algorithms such as Bagging, Attribute Selected Classifier, Log it Boost and Classification via Regression using the lung function and smoking dataset. The experimental results illustrates that the highest accuracy is established in Attribute Selected Classifier 100% and Classification via Regression and bagging algorithms shows more than 95% of accuracy while the Log it Boost algorithm shows 76% of performance accuracy. In future this work can be extended using other data mining techniques for the chosen data set.

References

[1] Rohini K & Suseendran G, "Aggregated K Means Clustering and Decision Tree Algorithm for Spirometer Data", *Indian Journal of Science and Technology*, Vol.9, No.44, (2016), pp.1-6.

[2] Srivastava S, "Weka: a tool for data preprocessing, classification, ensemble, clustering and association rule mining", *International Journal of Computer Applications*, Vol.88, No.10, (2014), pp.26-29.

[3] Khatri MD & Dhande S, "History and Current and Future trends of Data mining Techniques", *International Journal of Advance Research in Computer Science and Management Studies*, Vol.2, No.3, (2014), pp.311-315.

[4] Rohini K & Suseendran G, "Predicting lung disease severity evaluation and comparison of hybrid decision tree algorithm", *Indian Journal of Innovations and Developments* Vol.6, No.1, (2017), pp.1-15.

[5] Ian HW, Eibe F & Mark AH, *Data Mining Practical Machine Learning Tools and Techniques*. 3rd Edition, Elsevier, (2011).

[6] Breiman L, "Bagging predictors", Machine Learning, Vol.24, No.2, (1996) 123–140.

[7] Cohen WW, "Fast effective rule induction", *Twelfth International Conference on Machine Learning*, (1995), pp.115–123.

[8] Quinlan J, *C4.5: programs for machine learning*. San Mateo, CA: Morgan Kaufmann, 1993.

[9] Demiroz G & Guvenir A, "Classification by voting feature intervals", *Ninth European Conference on Machine Learning*, (1997), pp.85–92.

[10] Chye K & Gerald T, "Data Mining Applications in Healthcare", *Journal of Healthcare Information Management*, Vol.19, No.2, (2011), pp.64-72.