



(Print)

JCIT Vol. 9(4), 46-49 (2018). Periodicity-2-Monthly

(Online)



JOURNAL OF COMPUTER & INFORMATION TECHNOLOGY
An International Open Free Access Peer Reviewed Research Journal of Computer
Science Engineering & Information Technology
website:- www.compitjournal.org

Comparative Study of Rule Mining Classifiers with Hapatitus Datasets

RAMESH PRASAD AHARWAL¹ and MANMOHAN SINGH²

¹Assistant. Professor, Department of Mathematics Govt. P. G. College, Damoh (M.P.) (India)

²Associate Professor, Department of Computer Science & Engineering,
Dr. A.P.J. Abdul Kalam University, Indore (India)

Corresponding Author Email: ramesh_ahirwal_neetu@yahoo.com
<http://dx.doi.org/10.22147/jucit/090402>

Acceptance Date 20th July 2018,

Online Publication Date 2nd August, 2018

Abstract

This paper deals with comparison of the rule classifiers based on the evaluation parameters. The classification is a step by step procedure for designating a given piece of input data into any one of the given categories. We have used five Rule classifiers algorithms, namely Decision Table, JRIP, PART, OneR and ZeroR for experimental setup. The hepatitis datasets are used for calculating the performance by using the split percentage parameter. Finally find out the comparative analysis based on the performance factors such as the classification accuracy, Kappa statistic and execution time is performed on all the algorithms. The goal of this paper is to find out which classifier is best to other rule classifier in WEKA under hepatitis data.

Key words : Classification, WEKA, Data Mining.

Introduction

The aim of our work is to investigate the performance of different classification classifier using tool WEKA for hepatitis dataset. Is data mining software which is developed by the University of Waikato in New Zealand that implements data mining algorithms? In our work we will try to find, which classifier is best for hepatitis dataset. The classification is done based on rule classifier. It is used to automatically classify the variable. The rule classifier is generated for predict the best value. The features of rule based classifier are commonly exclusive rules and exhaustive rules. In this paper comparison is made to find out the

optimal result for Rules classifier algorithm such as Decision Table, JRIP, PART, OneR and ZeroR. In WEKA software there are four kinds of parameter like training set, supplied test set, cross validation and percentage split are available. In this paper we have used the cross validation parameter to calculate the data set values. The hepatitis dataset is using for comparison of those algorithms.

Data Mining :

Gartner Group, the information technology research firm defines "Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories,

using pattern recognition technologies as well as statistical and mathematical techniques". It is popular due to the successful applications in telecommunication, marketing, healthcare, medical and many other fields. In now a day, the usefulness of the methods has been proven also in medical field. Data mining is also known as knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging.

Classification :

Classification is a supervised learning method. Data classification is two-step process. In the first step, a model is built by analyzing the data tuples from training data having a set of attributes.

Decision Table: This classifier builds and using a simple decision table majority classifier as proposed. It summarizes the dataset with a decision table which contains the same number of attributes as the original dataset. Then, a new data item is assigned a category by finding the line in the decision table that matches the non-class values of the data item.

JRip Rule Classifier :

JRip is one of the basic and most popular algorithms. Classes are examined in increasing size and an initial set of rules for the class is generated using incremental reduced error JRip (RIPPER) proceeds by treating all the examples of a particular judgment in the training data as a class, and finding a set of rules that cover all the members of that class. Thereafter it proceeds to the next class and does the same, repeating this until all classes have been covered.

OneR or "One Rule" is a simple algorithm proposed by Holt. This classifier builds one rule for each attribute in the training data and then selects the rule with the smallest error rate as its one rule. This classifier is based on ranking all the attributes based on the error rate. To create a rule for an attribute, the most frequent class for each attribute value must be determined. The most frequent class is simply the class that appears most often for that attribute value.

PART : is a separate-and-conquer rule learner proposed by Eibe and Witten. The algorithm producing sets of rules called decision lists which are ordered set of rules. A new data is compared to each rule in the list in turn, and the item is assigned the category of the first matching rule. The PART algorithm is used to represent the knowledge/pattern identified to identify significant rules.

ZeroR - ZeroR is the simplest classification method which relies on the objective and ignores all predictors. ZeroR classifier simply predicts the majority class. Although there is no predictability power in ZeroR, it is useful for determining a baseline performance as a target for other classification methods. This Algorithm Construct a frequency table for the target and select its most frequent value.

About WEKA Tool :

WEKA is open source data mining software written in java, is a collection of machine learning algorithms for data mining tasks. WEKA is stand for The Waikato Environment for Knowledge Analysis is a unified workbench that allows researchers easy access to state-of-the-art techniques in machine learning. It was developed by the University of Waikato in New Zealand and implements data mining algorithms using the JAVA language. The WEKA project provides a comprehensive collection of machine learning algorithms and data preprocessing tools. It includes algorithms for regression, classification, clustering, association rule mining and attribute selection. Moreover, it also includes tools for data visualization. The data is usually imported from the ARFF file format, which consists of special tags to indicate different attribute names, attribute types, attributes values and the data itself.

Dataset and Source of Dataset :

In this experiment we have used hepatitis dataset. Dataset have extract from UCI machine learning data repository. The hepatitis dataset comprises 155 patients' records with 20 attributes.

Experimental Setup

To evaluate algorithms, performance several measures have been employed in this paper. Such as correctly classification, mean absolute error, Root Mean Squared Error and Kappa statistics. Using Performance parameters are calculated from following formula.

Correctly Classification = $(TP+TN) / (TP+TN+FP+FN)$

Mean absolute error = $\frac{|p_1 - a_1| + \dots + |p_n - a_n|}{n}$, where

p and a are predicted value and actual values

Root Mean Squared Error = $\sqrt{\frac{(p_1 - a_1)^2 + \dots + (p_n - a_n)^2}{n}}$

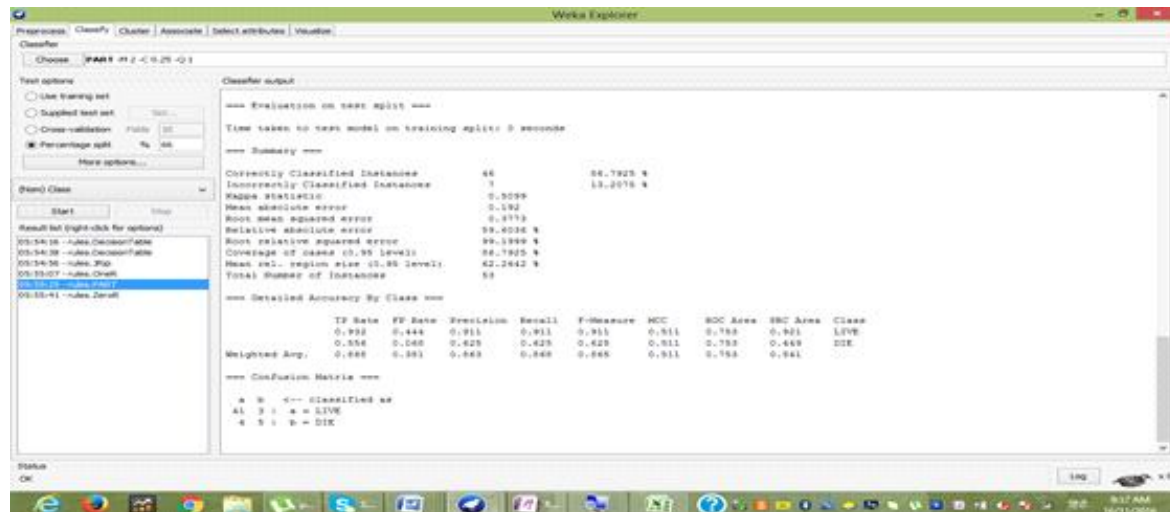


Fig.1 Screen shot of weka explorer which generate during the experiment

Experimental Result

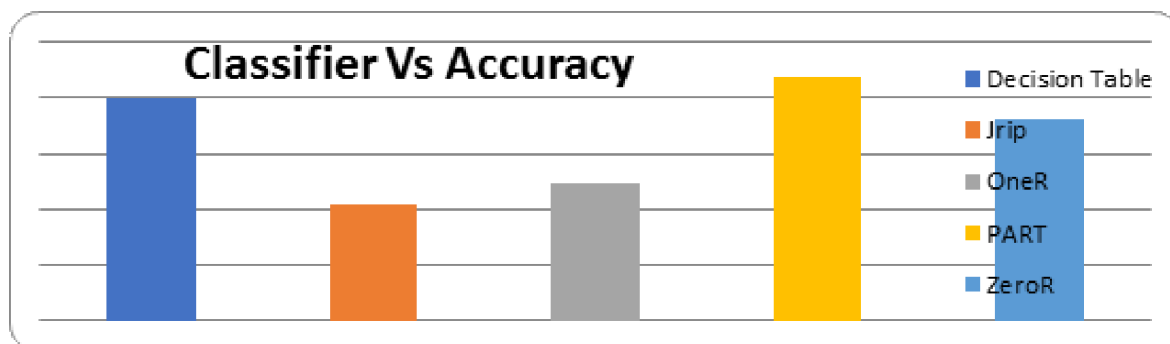


Fig. 2 Simulation result of correctly classification with classifier

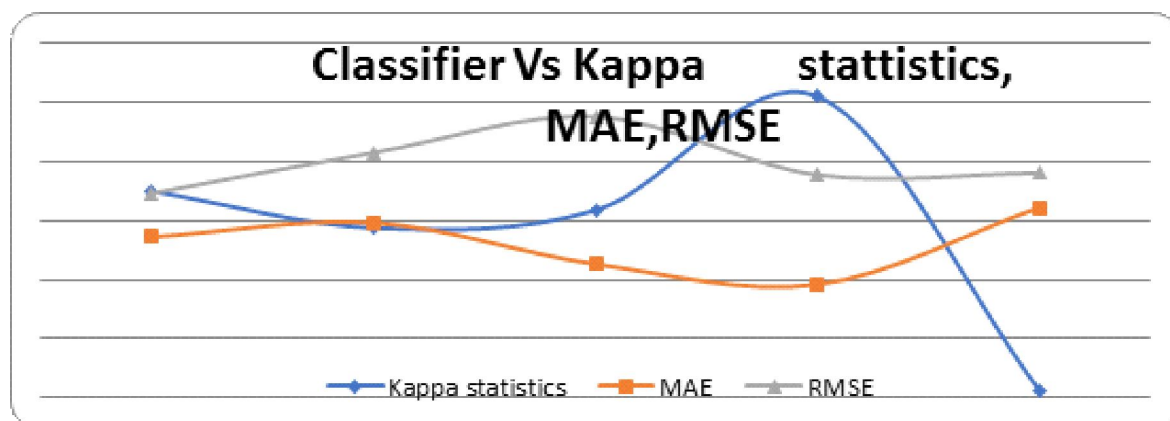


Fig. 3 Various Error Rate Evaluation Parameters with classifier

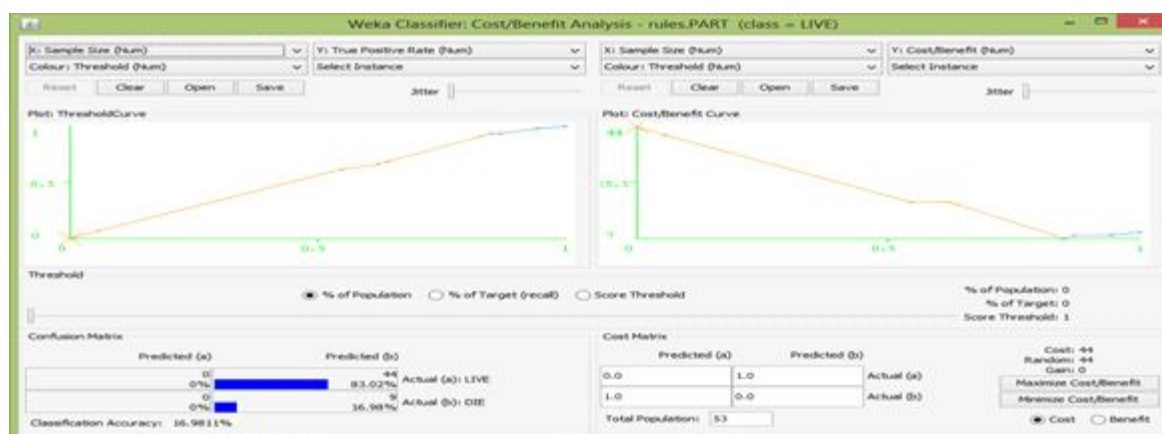


Fig. 4 Simulation result of Cost/Benefit Analysis of PART Classifier for Live Class

Conclusion

Five Classification rule based classifiers are considered in this study and experimentally evaluated using hepatitis data. The rule based classification algorithms are experimentally compared based on number of classified instances, accuracy and error rate using WEKA tool. We used 66% split options for our experiments. From the result we observed that PART classifier is the best rule based classification comparatively to the other studied classifiers.

References

1. Aditi Mahajan, Anita Ganpati, Performance Evaluation of Rule Based Classification Algorithms, (IJARCET) Volume 3 Issue 10, PP 3546-3550, ISSN: 2278 – 1323 (2014).
2. B.R. Gaines and P. Compton, Induction of ripple-down rules applied to modeling large databases.
3. J. Han and M. Kamber, "Data Mining: concept and techniques "first edition, Harcourt India private Limited. 2001 (2001).
4. I. H Witten., Frank E., *Data Mining, Practical Machine Learning Tools and Techniques*, 2nd Elsevier (2005).
5. P K Srimani, K Balaji, A comparative study of different classifiers on search engine based educational data, International Journal of Conceptions on Computing and Information Technology Vol. 2, Issue 1, Jan' 2014; ISSN: 2345 – 9808 (2014).
6. Zhou Z.-H., Jiang Y., Yang Y.-B., Chen S.-F., Lung cancer cell identification based on artificial neural network ensembles. *Artificial Intelligence in Medicine*, vol. 24, no. 1, 25–36 (2002).
7. U.M. Fayyad., G. Piatetsky, P. Shapiro, *et.al.*, "Advances in knowledge Discovery and Data Mining" AAAI/MIT Press., 1996 (1996).
8. ZeroR classifier <http://chem-eng.utoronto.ca/~datamining/dmc/zeror.htm>
9. Dataset <http://archive.ics.uci.edu/ml/>
10. WEKA at <http://www.cs.waikato.ac.nz/~ml/weka>.