



A Comparative Study of Selected Classification Algorithms of Data Mining

Ashish Kumar Dogra¹, Tanuj Wala²

¹Computer science & Career Point University, India

²Computer science & Career Point University, India

¹Ashish.dogra01@gmail.com; ²Tanu.thakur52@gmail.com

Abstract— *Data Mining is the process of extracting hidden knowledge from large volumes of raw data. Data mining has been defined as the nontrivial extraction of previously unknown, implicit and potentially useful information from data. Classification techniques are being used in different industry to easily identify the type and group to which a particular tuple belongs. This work explores the five classification algorithms (Zero, PART, OneR, Prism, J48) are compared on the bases of two techniques (cross validation and percentage split). Experimental comparison was performed by considering the car evaluation dataset and analyzing them using data mining open source WEKA tool.*

Keywords— *“Data Mining, Classification Algorithms, Car evaluation data set”*

INTRODUCTION

Data mining is the motion of all methods and techniques, which allow analyzing very large data sets to extract and discover previously unknown structures and relations out of such huge heaps of details. This information is filtered, prepared and classified so that it will be a valuable aid for decisions and strategies [2]. With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, if not necessary, to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting that could help in decision-making [3].

The title Data mining is the process of extracting patterns from data. As more data are gathered, with the amount of data doubling every three years [1]. Data mining is the process of discovering knowledge from large amounts of data stored either in databases or warehouses [2]. Data mining is becoming an increasingly important tool to transform these data into information. Data mining can also be referred as knowledge mining or knowledge discovery from data. Many techniques are used in data mining to extract patterns from large amount of database [1]. Classification is a data mining (machine learning) technique used to predict group membership for data instances.

Classification is a supervised procedure that learns to classify new instances based on the knowledge learnt from a previously classified training set of instances [5]. It takes a set of data already divided into predefined groups and searches for patterns in the data that differentiate those groups supervised learning, pattern recognition and prediction. Classification consists of predicting a certain outcome based on a given input. It is one of the Data

Mining techniques that is mainly used to analyze a given dataset and takes each instance of it and assigns this instance to a particular class with the aim of achieving least classification error. It is used to extract models that correctly define important data classes within the given dataset. It is a two-step process. In first step the model is created by applying classification algorithm on training data set. Then in second step, the extracted model is tested against a predefined test dataset to measure the model trained performance and accuracy. So, classification is the process to assign class label for this dataset whose class label is unknown [7]. Find a model that describes the data classes or concepts. This derived model is based on analysis of set of training data. The derived model can be presented in the following forms Classification (IF-THEN) Rules, Decision Trees, Mathematical Formulae and Neural Networks.

Classification Algorithm

Classification is a classic data mining technique based on machine learning. Basically classification is used to classify each item in a set of data into one of predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network and statistics [9]. Classification algorithm finds relationships between the values of the predictors and the values of the target. Components of Data Mining Algorithms i) Model Representation (Knowledge Representation) is the language for describing discoverable patterns / knowledge ii) Model Evaluation is estimating the predictive accuracy of the derived patterns iii) Search methods we have two type of broadly classified search method one is parameter search when the structure of a model is fixed, search for the parameters which optimise the model evaluation criteria and another is model search when the structure of the model is unknown iv) Learning Bias consist of feature selection and Pruning algorithm [11].

Data classification is a two steps process in which first step is the training phase where the classifier algorithm builds classifier with the training set of tuples and the second phase is classification phase where the model is used for classification and its performance is analyzed with the testing set of tuples.

The algorithm can differ with respect to accuracy, time to completion, and transparency. In practice, it sometimes makes sense to develop several models for each algorithm, select the best model for each algorithm, and then choose the best of those for deployment.

Algorithms Evaluated

Five Classification algorithms considered for our study are

- ZeroR
- Part
- Prism
- OneR
- J48

ZeroR:- ZeroR is the simplest classification method which relies on the target and ignores all predictors. ZeroR classifier simply predicts the majority category (class). Although there is no predictability power in ZeroR, it is useful for determining a baseline performance as a benchmark for other classification methods [10]. It is the simplest method which relies on the frequency of target. ZeroR is only useful for determining a baseline performance for other classification methods.

PART:- PART is a partial decision tree algorithm, which is the developed version of C4.5 and RIPPER algorithms .PART is a separate-and-conquer rule learner proposed by Eibe and Witten [54]. 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 (a default is applied if no rule successfully matches). PART builds a partial C4.5 decision tree in its each iteration and makes the best leaf into a rule. The algorithm is a combination of C4.5 and RIPPER rule learning [2].

Prism:- This simple and straightforward covering algorithm work by first picking a class from the dataset for which to create a new rule having the class as its conclusion, and selectively adding tests to the condition of the rule, striving for maximum number of instances covered and 100% accuracy. The accuracy of a test is measured by the ratio of the number of positive instances p to the total number of instances covered by the rule. The positive instances covered by the new rule then are removed from the dataset for further rule generation. Then, negative instances should remain in the data set to await a later iteration of the process. This process continues until no more instances remain to covered.

OneR:- OneR or “One Rule” is a simple algorithm proposed by Holt. The OneR builds one rule for each attribute in the training data and then selects the rule with the smallest error rate as its one rule. The algorithm 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 [54]. The most frequent class is simply the class that appears most often for that attribute value. A rule is simply a set of attribute values bound to their majority class. OneR selects the rule with the lowest error rate. In the event that two or more rules have the same error rate, the rule is chosen at random [8]. The OneR algorithm creates a single rule for each attribute of training data and then picks up the rule with the least error rate [4].

J48:- J48 classifier is a simple C4.5 decision tree for classification. It creates a binary tree. The decision tree approach is most useful in classification problem. With this technique, a tree is constructed to model the classification process. Once the tree is built, it is applied to each tuple in the database and results in classification for that tuple[1][3].

Tool Used for Performance Evaluation Platform

SOFTWARE USED

- Operating system Window 8
- JAVA version 7
- Tool-WEKA 3.7

Introduction to WEKA

Waikato Environment for Knowledge Analysis (WEKA) tool is used for the performance evaluation. WEKA is a collection of machine learning algorithms for data mining tasks. WEKA is created by researchers at University of Waikato in New Zealand. It is written in Java and runs on almost any platform . The algorithms in WEKA can either be applied directly to a dataset or called from own Java code. It is also well-suited for developing new machine learning schemes. WEKA is open source software issued under the GNU General public License. It is a collection of open source ML algorithms pre-processing, classifiers, clustering, association rule and visualization [6].



Data Set

The dataset is car evaluation data set has been used for the experiment. In this data set there are 1728 number of instances and 7 number of attributes.

Result and Discussion

Cross Validation Testing Method

Cross-Validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one used to learn or train a model and the other used to validate the model. In cross validation, the training and validation sets must cross-over in successive rounds such that each data point has a chance of being validated against. The basic form of cross-validation is k-fold cross-validation. Other forms of cross-

validation are special cases of k-fold cross-validation or involve repeated rounds of k-fold cross-validation. Ten Fold cross validation method is used in our experiment. In ten folds cross-validation; classifier constructed 10 identical instances of the data set and then split the data in each of these instances in 10 % for training and 90% for testing.

Number of Instances Classified

Three data are used for evaluation of five rule based algorithms. Firstly all these five algorithms are compared on the basis of number of classified instances. Number of instances consists of number of correctly classified instances and number of incorrectly classified instances by algorithms using ten fold cross validation technique in WEKA.

Car Evaluation

Total numbers of instances in car evaluation data set are 1728 instances. Ten fold cross validation technique is used to obtain the number of classified and in classified number of instances in car evaluation data sets.

Table 1: Number of Classified Instances for Car Evaluation

Algorithm	Correctly Classified Instances	Incorrectly Classified Instances
ZeroR	1210	518
PART	1655	73
OneR	1210	518
Prism	1550	54
J48	1596	132

Table 1 shows the number of correctly classified and incorrectly classified instances for five classification algorithms. PART algorithm with the highest number of 1655 classified instances and OneR algorithm with greatest number of incorrectly classified instances i.e 518. OneR 518 incorrect instances are very high as compare to number of incorrectly classified instances of other four studied algorithms.

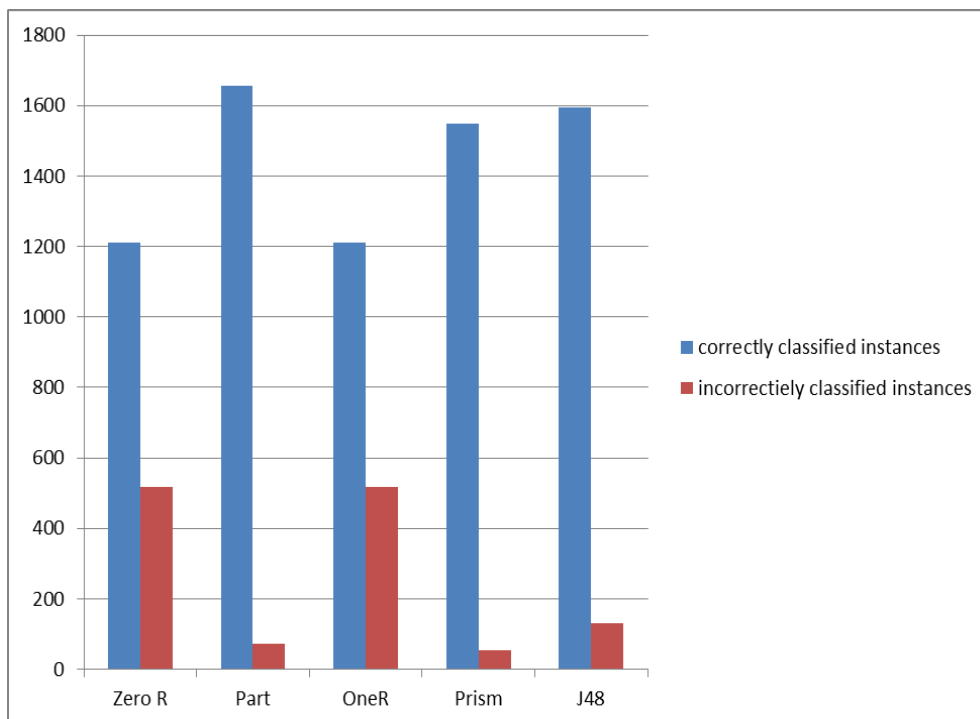


Figure 1: Number of Classified Instances for Car Evaluation Data Set

From Table 1 and Figure 1 it is evident that from car evaluation data set, PART have highest number of correctly classified instances followed by Prism and j48. ZeroR and OneR have same performance for this dataset.

Accuracy Measurement

The following table shows the accuracy measure of classification techniques. They are the True Positive rate, F Measure, Precision, Recall, Receiver Operating Characteristics (ROC) Area and Kappa Statistics.

Car Evaluation

Total numbers of instances in car evaluation data set are 1728 instances. Ten fold cross validation is used. Six parameters value is obtained when car evaluation is evaluated using a WEKA for five algorithms.

Table 2: Accuracy Parameters for Car Evaluation

Algorithm	Kappa Statistics	TP Rate	Precision	Recall	F-Measure	ROC area
ZeroR	0	.7	.49	.7	.577	.497
PART	.9091	.958	.959	.958	.958	.99
oneR	0	0.7	.49	0.7	.577	.5
Prism	0.915	.966	.964	.966	.965	.949
J48	.8343	.924	.924	.924	.924	.976

From table 2 it is evident that PART algorithm had best performance for all six studied parameters for accuracy evaluation. OneR with worst performance for all six evaluated parameters

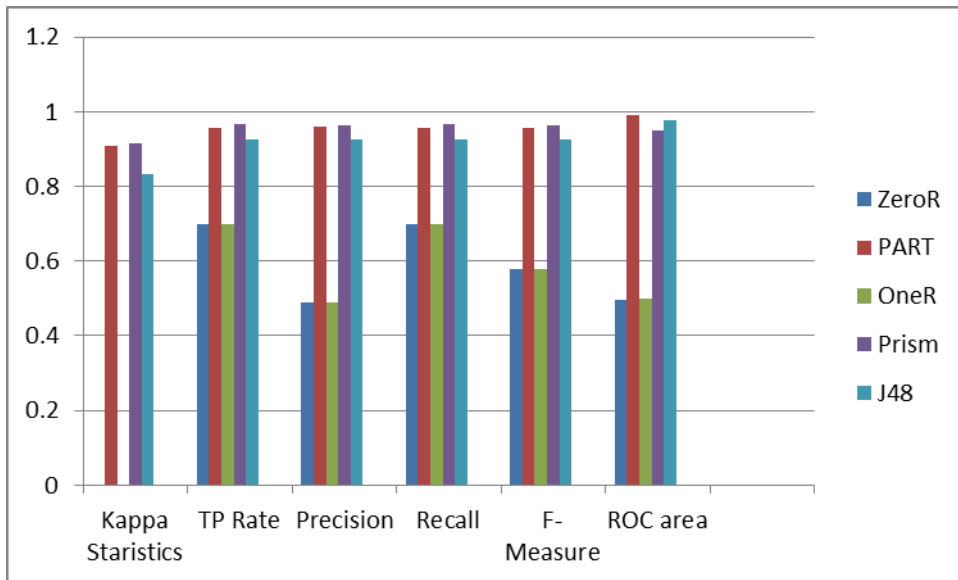


Figure 2: Various Accuracy Parameters for Car Evaluation Data Set

From table 2 and figure 2 it is evident that for car evaluation data set, from over all evaluation Prism and PART algorithm performance good in terms of accuracy. ZeroR shows average performance and OneR and J48 shows poor accuracy performance for all parameters.

Error Rate

Mean Absolute Error (M.A.E), Root Mean Square Error (R.M.S.E), Relative Absolute Error (R.A.E) and Root Relative Squared Error (R.R.S.R).

Car Evaluation

Four basic error rate parameters for the evaluation of five classification algorithms for car evaluation data set.

Table 3: Error Evaluation Parameters for Car Evaluation

Algorithm	MAE	RMSE	RAE	RRSE
ZeroR	0.229	0.3381	100	100
PART	0.0241	0.1276	10.5343	37.7421
OneR	0.1499	0.3871	65.4464	114.5017
Prism	0.0168	0.1297	8.4112	42.394
J48	0.0421	0.1718	18.3833	50.8176

Table 3 shows the Prism algorithm had the least value for two parameters i.e MAE and RAE. PART algorithm having least value for other two parameters i.e RMSE and RRSE

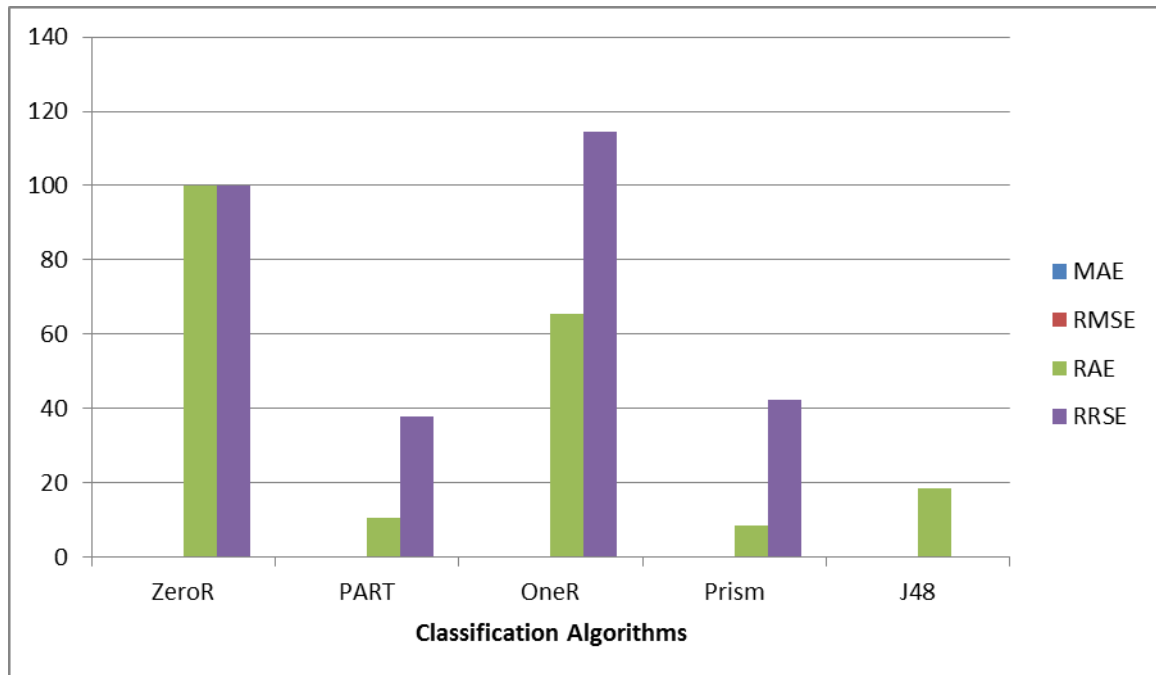


Figure 3: Various Error Evaluation Parameters for Car Evaluation Data Set

From Table 3 and Figure.3 it is evident that PART and J48 have minimum error rate and good performance. Prism algorithm have intermediate error rate thus having average performance. Both OneR and ZeroR algorithm have very high error rate and poor performance as compare to other algorithms.

Percentage Split Method

In percentage split, the database is randomly split in to two disjoint datasets. The first set, which the data mining system tries to extract knowledge from called training set. The extracted knowledge tested against the

second set which is called test set, it is common to randomly split a data set under the mining task in to 2 parts [42F]. 66% percentage split is chosen. Objects of the original database are used as a training set and the rest of objects as a test set. Once the tests is carried out using the selected datasets, then using the available classification and 66 % percentage split test mode, results are collected and an overall comparison is conducted .

Number of Instances Classified

Three data are used for evaluation of five algorithms. Firstly all these five algorithms are compared on the basis of number of classified instances. Number of instances consists of number of correctly classified instances and number of incorrectly classified instances by algorithms using 66% percentage split technique in WEKA.

Car Evaluation

Total number of classified instances for car evaluation dataset is 1728. Number of instances used for training is 1140 i.e 66% of instances used for training and left 33% of total instances i.e 588 instances are used for testing.

Table 4: Number of Classified Instances for Car Evaluation

Algorithm	Correctly Classified Instances	Incorrectly Classified Instances
ZeroR	535	53
PART	559	29
OneR	409	179
Prism	516	18
J48	409	179

Table 4 shows the number of correctly classified and incorrectly classified instances for five algorithms. PART and Ridor algorithm had the highest number of classified instances. OneR algorithm had highest number of incorrectly classified instances.

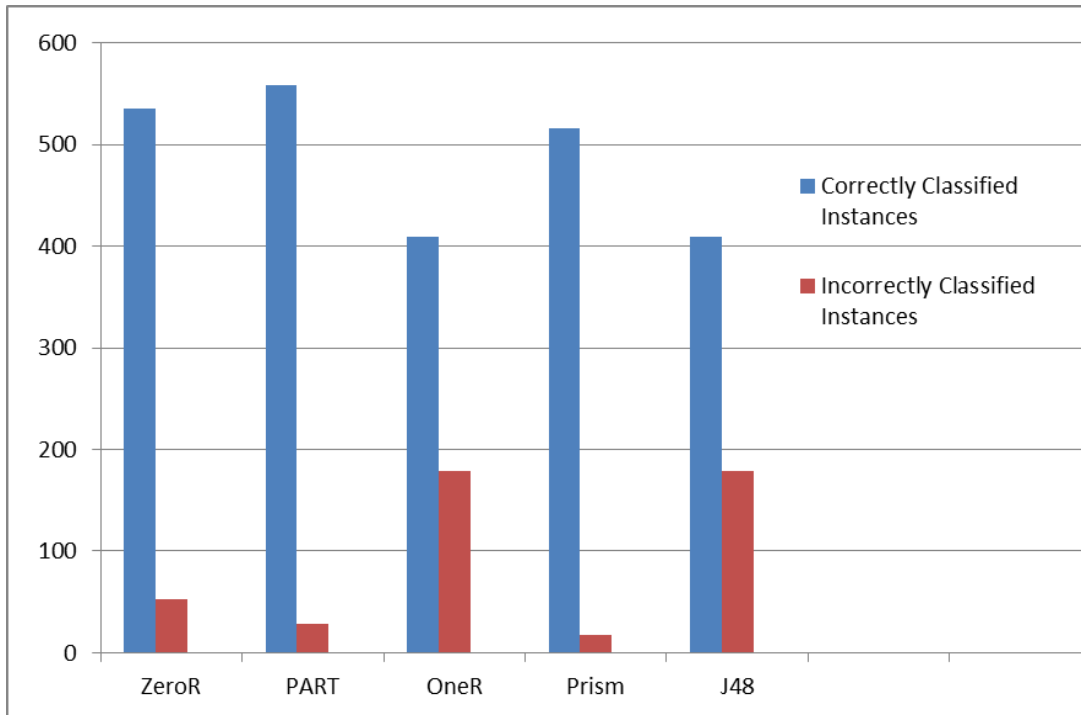


Figure 4: Number of Classified Instances for Car Evaluation Data Set

From the Table 4 and Figure 4 it is evident that from car evaluation data set PART have highest number of correctly classified instances followed by ZeroR algorithm. All the two algorithms perform well in classifying the instances. Prism algorithm show average performance. Whereas OneR and J48 algorithm have lowest performance in terms of correctly classification of instances.

Accuracy Measure

The following table shows the accuracy measure of classification techniques. True Positive rate, F Measure, Precision, Receiver Operating Characteristics (ROC) Area and Kappa Statistics.

Car Evaluation

Total numbers of instances in car evaluation data set are 1728 instances.

Table 5: Accuracy Parameters for Car Evaluation

Algorithm	Kappa Statistics	TP Rate	Precision	Recall	F-Measure	ROC area
ZeroR	0	0	0.484	0.696	0.571	0.5
PART	0.8939	0.951	0.953	0.951	0.951	0.986
OneR	0	0.696	0.484	0.696	0.571	0.5
Prism	0.9109	0.966	0.964	0.966	0.964	0.944
J48	0.8088	0.91	0.915	0.91	0.571	0.965

Table 5 shows the six parameters for evaluating accuracy of five classification algorithms. These parameters are Kappa Statics, TP Rate, Precision, Recall, F-measure and ROC area. Part and Prism algorithm had best performance and ZeroR with worst performance for all six evaluated parameters.

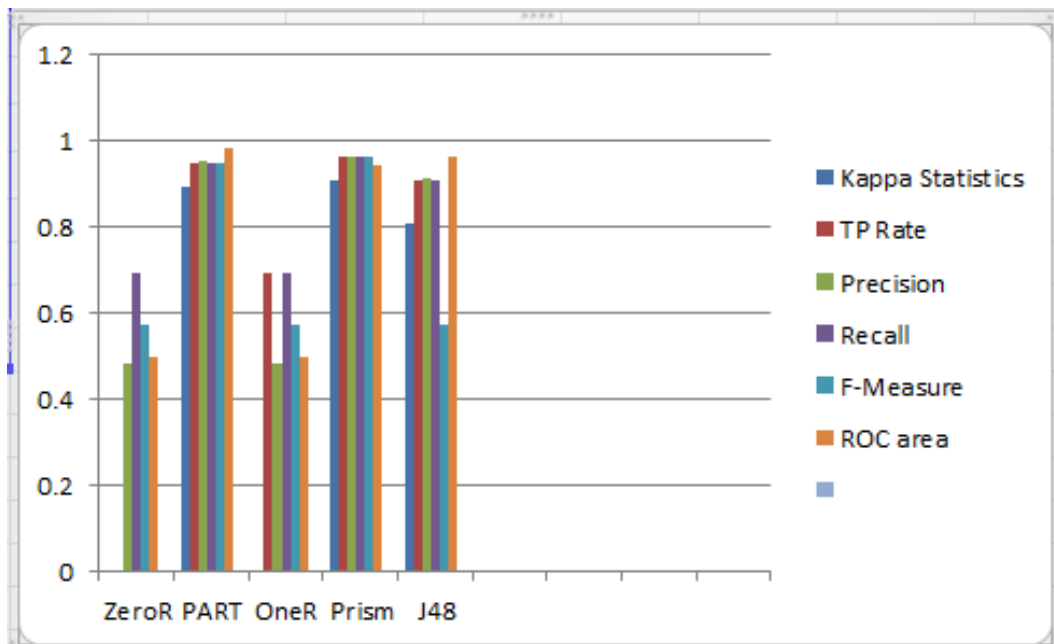


Figure 5 Various Accuracy Parameters for Car Evaluation Data Set

From the Table 5 and Figure 5 it is evident that for car evaluation data set, from over all evaluation Prism and PART algorithm performance good in terms of accuracy. J48 algorithm having performance following two algorithm OneR shows average performance and ZeroR shows poor accuracy performance for all parameters.

ERROR RATE

Mean Absolute Error (M.A.E), Root Mean Square Error (R.M.S.E), Relative Absolute Error (R.A.E) and Root Relative Squared Error (R.R.S.R).

Car Evaluation

Four basic error rate parameters for the evaluation of five classification algorithms foe car evaluation data set.

Table 6: Error Evaluation Parameters for Car Evaluation

Algorithm	MAE	RMSE	RAE	RRSE
ZeroR	0.2295	0.339	100	100
PART	0.0294	0.1391	12.8137	41.0226
OneR	0.1522	0.3901	66.3341	115.0861
Prism	0.0169	0.1298	8.778	43.7677
J48	0.0509	0.1883	22.1706	55.5396

Table 6 shows the four basic error rate parameters for the evaluation of five classification algorithms. PART algorithm had the least value for all four parameters. OneR algorithm having highest value for two parameters i.e RMSE and RRSE. DTNB algorithm having highest value for other two parameters i.e MAE and RAE.

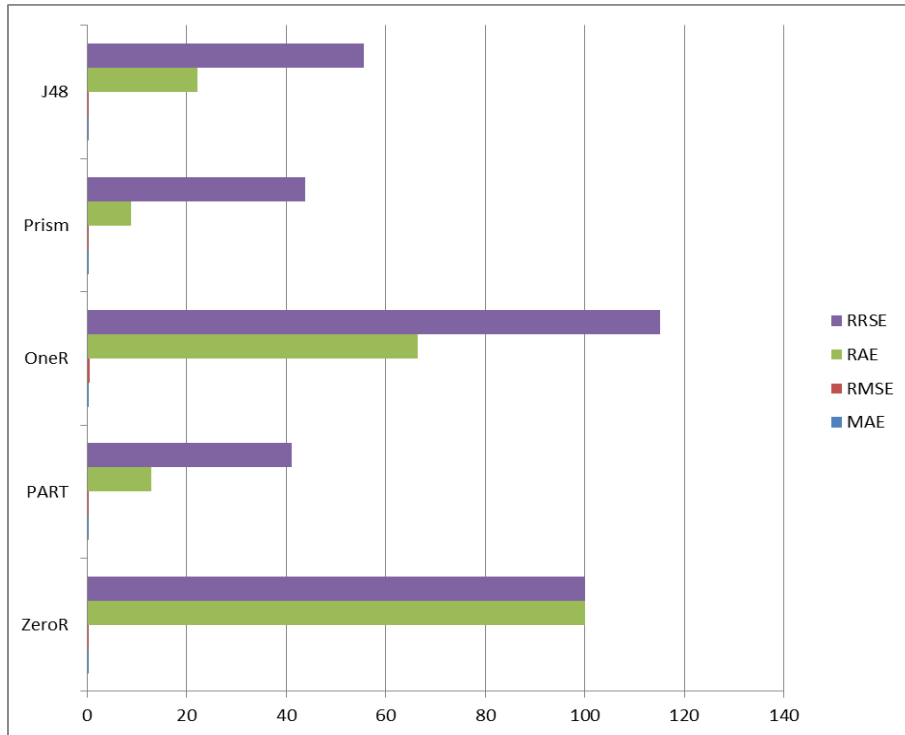


Figure6: Various Error Evaluation Parameters for Car Evaluation Data Set

From Table 6 and Figure 6 it is evident that PART algorithm has the best performance when compares to other algorithm. PART has minimum error followed by J48 algorithm. ZeroR and OneR algorithm have very high error rate and show poor performance.

Summary

From all the above figures it can be summon that PART is best algorithm for classification of car evaluation. Over all J48 table and Prism algorithm have average performance and OneR with very low performance for all three data set. Thus PART and Prism perform well for all the three performance measurement parameters i.e number of classified instances, accuracy and error rate.

Conclusion & Future Scope

This thesis work presented an introduction to data mining and data mining process. A brief introduction to classification methods and algorithms is provided. Five Classification algorithms ZeroR, OneR, PART, Prism, and J48 are introduced and experimentally evaluated with different data set using WEKA tool. The classification algorithms are experimentally compared based on number of classified instances, accuracy and error rate. We used two testing options for our experiment ten fold cross validation and 66% percentage split. From our study we found that PART is best algorithm for classification of car evaluation data set. Prism and J48 had average performance. OneR algorithm with over all low performance and thus least recommended algorithm. After considering and comparing all the tables and graphs in our study PART algorithm is recommended among all five classification algorithms. As per the future scope is considered during thesis work I analysed that a good amount of research work is devoted to various classification algorithms. And this research work can also be extended by testing the selected algorithms for other machine learning tasks such as clustering, association etc.

REFERENCES

- [1] Abdulkarim, Mirza Nazura. Classification and Retrieval of Research Papers: A Semantic Hierarchical Approach .Diss. Christ University, 2010
- [2] Ali, Shawkat, and Kate A. Smith. "On learning algorithm selection for classification." Applied Soft Computing , 2006
- [3] Apté, Chidanand, and Sholom Weiss. "Data mining with decision trees and decision rules." Future generation computer systems ,1997.
- [4] Buddhinath, Gaya, and Damien Derry."A simple enhancement to One Rule Classification." Department of Computer Science & Software Engineering. University of Melbourne, Australia , 2006.
- [5] Devasena, C. Lakshmi, et al. "Effectiveness Evaluation of Rule Based Classifiers for the Classification of Iris Data Set." Bonfring International Journal of Man Machine Interface 1.Special Issue Inaugural Special Issue , 2011.
- [6] Lilla Gula "Information Data Management for the Future of Communication" ,Luleå University of Technology Department of Computer Science, Electrical and Space Engineering.
- [7] Ragab, Abdul Hamid M., et al. "A Comparative Analysis of Classification Algorithms for Students College Enrollment Approval Using Data Mining", Proceedings of the 2014 Workshop on Interaction Design in Educational Environments. ACM, 2014.
- [8] Vijayaran S, Sudha. "An Effective Classification Rule Technique for Heart Disease Prediction",.International Journal of Engineering Associates (IJE), ISSN: 2320-0804, Vol.1, Issue 4, 2013.
- [9] <http://www.zentut.com/data-mining>, "data-mining-techniques",Retrived on:26 April 2015.
- [10]<http://www.saedsayad.com/zeror.htm>, " ZeroR",Retrived on: 14 May 2015
- [11]<http://www.doc.ic.ac.uk/~yg/course/ida2002/ida-2002-2.PPT>, "Data Pre-processing",Retrived on: 30 May 2015.