Performance Evaluation of Rule Based Classification Algorithms

Aditi Mahajan[1], Anita Ganpati[2]
[1]Research Scholar, Department of Computer Science, Himachal Pradesh University Shimla, India
[2]Associate Professor, Department of Computer Science, Himachal Pradesh University Shimla, India

Abstract-**The growth of the internet has created a vast new arena for information generation. There is huge amount of data available in Information Industry. Databases today can range in size of the terabytes or more bytes of data. To address these issues, researchers turned to a new research area called Data Mining. Data mining is a technology that blends traditional data analysis methods with sophisticated algorithms for processing large volume of data. Experimental evaluation of rule based classification algorithm is performed using WEKA open source tool. Five rule based classification algorithm considered are OneR, PART, Decision Table, DTNB and Ridor algorithms. Chess End Game data set is used in experimental evaluation of algorithms .Cross validation testing technique is considered for experiment. For comparing the five algorithm three performance parameters number of classified instances, accuracy and error rate are considered. The results of experiment are presented in tabular and graphical form. From this study it is found that PART is best algorithm for classification.**

Keywords: Classification, Cross Validation Data Mining, Performance, Rule based

I. INTRODUCTION

We are in an age often referred to as the information age. Lots of data is being collected and warehoused. Computers have become cheaper and more powerful. Data is being collected and stored at enormous speed. Today, we have far more information than we can handle. The term "Data Mining" appeared around 1990 in the database community which later on became more widespread in the business and press communities. Currently, Data Mining and Knowledge Discovery are used interchangeably [6].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" [13].Data mining is cyclical process, creating a data mining model is a dynamic and iterative process. There are many types of data mining, typically divided by the kind of information known and the type of knowledge sought from the data-mining model. Data mining is generally divided into two main categories predictive and descriptive. Predictive Modelling is used when the goal is to estimate the value of a particular target attribute and there exist sample training data for which values of that attribute are known. Classification, regression, time series analysis and prediction use predictive modeling. Descriptive Modeling divides the data into groups, it do not predict a target value, but focus more on the intrinsic structure, relations, interconnectedness, etc. of the data. Clustering, summarization, association rules and sequential patterns analysis use descriptive modeling.

II. RULE BASED CLASSIFICATION ALGORITHMS

A large number of algorithms and data mining tools have been developed and implemented to extract information and discover knowledge patterns that prove to be advantageous for decision making. Classification is a supervised procedure that learns to classify new instances based on the knowledge learnt from a previously classified training set of instances. 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 . Typical Classification Algorithms are Decision trees, rule-based induction, neural networks, genetic algorithms and bayesian networks.Rule based classification algorithm also known as separate-and-conquer method is an iterative process consisting in first generating a rule that covers a subset of the training examples and then removing all examples covered by the rule from the training set. This process is repeated iteratively until there are no examples left to cover [8]. Following are the rule based algorithms considered for study:

I) 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 [9].To create a rule for an attribute, the most frequent class for each attribute value must be determined [14]. 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 [11]. The OneR algorithm creates a single rule for each attribute of training data and then picks up the rule with the least error rate [2].

II) PART is a separate-and-conquer rule learner proposed by Eibe and Witten [14]. 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 [1].

III) Decision Table algorithm builds and using a simple decision table majority classifier as proposed by Kohavi [14]. 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. Decision Table employs the wrapper method to find a good subset of attributes for inclusion in the table [1]. By eliminating attributes that contribute little or nothing to a model of the dataset, the algorithm reduces the likelihood of over-fitting and creates a smaller and condensed decision table.

IV) DTNB this is for building and using a decision table/naive bayes hybrid classifier. At each point in the search, the algorithm evaluates the merit of dividing the attributes into two disjoint subsets: one for the decision table, the other for naive Bayes. A forward selection search is used, where at each step, selected attributes are modelled by naive Bayes and the remainder by the decision table and all attributes are modelled by the decision table initially [3]. At each step, the algorithm also considers dropping an attribute entirely from the model.

V) Ridor algorithm was introduced by Compton and Jansen, ripple-down rule technique as a methodology for the acquisition and maintenance of large rule-based systems [15]. Ridor algorithm is the implementation of a Ripple-Down Rule learner proposed by Gaines and Compton. It generates a default rule first and then the exceptions for the default rule with the least (weighted) error rate. Then it generates the "best" exceptions for each exception and iterates until pure. Thus it performs a tree-like expansion of exceptions. The exceptions are a set of rules that predict classes other than the default. IREP is used to generate the exceptions [14].

III. LITERATURE REVIEW

Neelamadhab Padhy et al. [7] provided completely introduction to data mining. Author explained the data mining life cycle and data mining classification tasks. Paper focused on variety of techniques, approaches and crucial areas of research which have been marked as the important field of data mining technologies. This paper reviewed applications and propose feature directions for some of data mining applications.

Włodzisław Duch [4] provided the introduction to the Rule based system. Rules descried in paper are the one used in classification (Classification, Machine Learning), regression (Regression, Statistics) and association tasks. Paper explained the various forms of rules that allow expression of different types of knowledge classical prepositional logic (C-rules), association rules (Arules), fuzzy logic (F-rules), M-of-N or threshold rules (T-rules) and prototype-based rules (P-rules) .All these types of rules are explained in detail with their advantages and disadvantages.

S. Vijayarani and M. Muthulakshmi [10] provided the classification of the computer files based on their extension category using classification rule techniques. In this paper three classification algorithms namely decision table, DTNB (Decision Tree Naïve Bayes) and OneR classifiers are used for performing classification of computer files. From the experimental results, DTNB proves to be more efficient than other two techniques.

C. Giraud-Carrier and O. Povel [5] presented a general schema for the characterisation of data mining tools and defined the standard data mining software requirements .Paper focused on business requirements in data mining. Authors conducted the survey of 41 popular data mining tools and presented the results obtained.

S.Vijayarani and S.Sudha [11] focused on classification rule based algorithms Decision Table, JRip, OneR, and Part algorithm. Comparative analysis was done by using WEKA tool. The performance factors used for analysis are accuracy and error measures. By analyzing the experimental results of accuracy measure, it was concluded that the decision table classification rule technique turned out to be best classifier for heart disease prediction because it contains more accuracy. By analyzing all error rates, the Decision table and OneR classification rule algorithm contains least error rate.

Biao Qin, Yuni Xia et al. [12] proposed a new rule-based algorithm uRule for classifying and predicting both certain and uncertain data and a new measure called probabilistic information gain for generating rules. Authors extended the rule pruning measure for handling data uncertainty. Proposed algorithm uRule introduced new measures for generating, pruning and optimizing rules. These new measures were computed considering uncertain data interval and probability distribution function. Based on the new measures, the optimal splitting attribute and splitting value can be identified and used for classification and prediction. The proposed uRule algorithm can process uncertainty in both numerical and categorical data.

C. Lakshmi Devasena et al. [3] studied the effectiveness of Rule-Based classifiers for classification by taking a sample data set and comparing different rule-based classifiers Conjunctive Rule Classifier, Decision Table Classifier, DTNB Classifier, OneR Classifier, JRIP Classifier, NNGE Classifier, PART Classifier, RIDOR Classifier and ZeroR Classifier. The performances of the classifiers were measured and results are compared using the Iris Data set. The experiment was done using an open source Machine Learning Tool. The performances of the various algorithms measured in classification are based on following parameters .Accuracy, RMSE, MAE and confusion matrix. NNGE Classifier performs well in the classification problem.

S. Vijayarani et al [11] discussed the classification rule techniques in data mining and are compared for predicting heart disease. The classification rule algorithms are namely. Decision table, JRip, OneR and Part. By analyzed the experimental results of accuracy measure, it is observed that the Decision Table classification rule technique turned out to be best classifier for heart disease prediction because it contains more accuracy. By analyzed all error rates, the Decision table and OneR classification rule algorithm contains least error rate in possible two outcomes.

IV. NEED OF STUDY

Data is being collected and stored at enormous speed. Today we have far more information than we can handle Manual data analysis has been around for some time, but it creates a bottleneck for large data analysis. Data mining is a technology that blends traditional data analysis methods with sophisticated algorithms for processing large volume of data. Classification algorithms are important and widely studied area under research from many years. Many new and improved classification algorithms have been proposed over past years. Many number of research paper and articles of experimental evaluation of classification algorithms like decision tress, clustering and neural network are present. As compared to other categories of classification algorithms rule based classification algorithms are least studied category of classification algorithm. Our study focus of experimental evaluation of rule based classification algorithm. Very few research paper of classification focus on rule based algorithms. Experiment evaluation of present rule based algorithms and introduction of new approaches in this category of classification algorithm in an important aspect.

V. OBJECTIVE OF STUDY

- To have a general knowledge of data mining classification technique and algorithm.
- To experimentally analyze and compare five Classification Rule based algorithm OneR, PART, Decision table, DTNB and Ridor.

VI. RESEARCH METHODOLOGY

Firstly a theoretical study of classification algorithms was done by retrieving the information from studying and reviewing secondary data acquired from research journals, thesis, books and internet. Then an empirical study was performed to evaluate the classification Rule based algorithms. Experimental comparison by considering the Chess End-Game King+Rook Versus King+Pawn dataset taken from the UCI repository[15]. For analyzing algorithms open source WEKA tool is used.

VII. ANALYSIS

A) Approach followed

I) Five rule based algorithms are considered for study i.e OneR, PART, Decision Table, DTNB and Ridor rule based classification algorithms

II) WEKA tool is used for an experimental study

III) Data set Chess End-Game King+Rook Versus King+Pawn having 3196 instances and 36 attributes is considered for experiment.

IV) Tenfold cross-validation is used in this experimental study. In this testing technique, data set is equally divided into 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. In each test 9 folds of data are used for training and one fold is for testing. The test procedure is repeated 10 times. The final accuracy of an algorithm will be the average of the 10 trials.

V) Three performance parameters have been considered for experimental evaluation i) Number of classification instances ii) Accuracy Parameters iii) Error rate .

B) Results and Discussion

I) Number of Classified Instances consisting of number of correctly classified and incorrectly classified instances by five Rule based algorithms using WEKA tool.

Table1: Number of Classified Instances for Chess End Game

Algo.	Correctly Classified Instances	Incorrectly Classified Instances
OneR	2124	1072
PART	3166	30
Decision Table	3107	89
DTNB	3088	108
Ridor	3152	44

From Table 1 it is evident that PART algorithm has highest number of correctly classified instances and OneR algorithm has highest number of incorrectly classified instances.

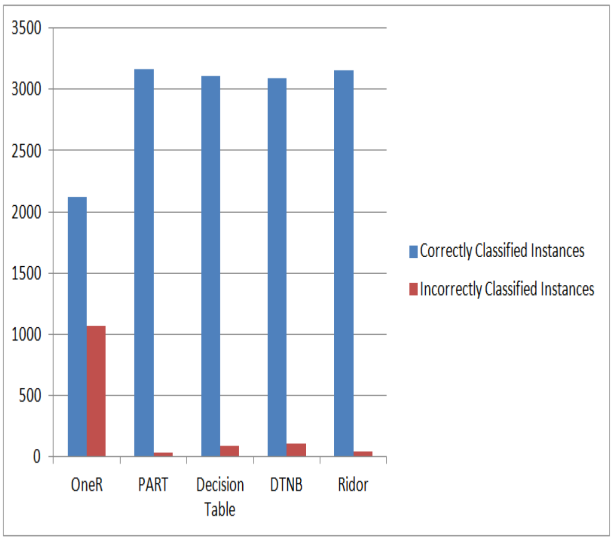

Figure1: Number of Classified Instances

From Figure 1 it is evident that PART shows the best performance as compare to other studied algorithms. PART has highest number of correctly classified instances followed by Ridor and Decision Table. Ridor algorithm shows good performance .DTNB algorithm has an average performance in terms of correctly classification of instances and OneR show poor classification performance.

II) Accuracy Parameters for Chess End Game- for accuracy measurement six parameters are considered:
i) Precision= True Negative/ True Negative + False Positive
ii) Recall = True Positive / True Positive + False Negative
iii) TP Rate is a rate of true positives (instances correctly classified as a given class)
iv) F-measurer is a combined in a single quantity derived from precision and recall, the
F1 = (2 *Precision * Recall)/ (Precision + Recall)
iv) ROC Area is defined as area under the ROC curve (*AUC*) which is the probability of randomly chosen positive instance that is ranked above randomly chosen negative one.
v) Kappa Statistic measure degree of agreement between two sets of categorized data .Kappa result
varies between 0 to 1 intervals. Higher the value of Kappa means stronger the agreement.

Table2: Accuracy Parameters for Chess End Game

Algo.	Kappa Statistics	TP Rate	Precision	Recall	F-Measure	ROC area
OneR	.319	.665	.675	.665	.655	.657
PART	.9812	.991	.991	.991	.991	.997
Decision Table	.9442	.972	.972	.972	.972	.985
DTNB	.9322	.966	.966	.966	.966	.988
Ridor	.9724	.986	.986	.986	.986	.986

Table 2 shows the six parameters for evaluating accuracy of five rule based algorithms. These parameters are Kappa Statics, TP Rate, Precision, Recall, F-measure and ROC area. PART algorithm had highest values for all accuracy parameters. OneR algorithm and having least values and performance as compare to other algorithms.

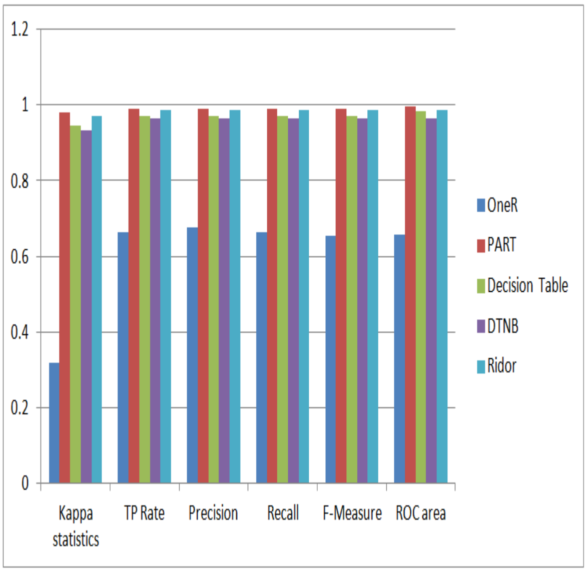

Figure 2: Various Accuracy Parameters

From Figure 2 it is evident that PART algorithm has the highest performance. PART and Ridor both algorithms perform well for all parameters and have high accuracy. Decision Table and DTNB have average performance and OneR algorithm shows poor performance for all parameters.

III) Error Rate Evaluation Parameters for Chess End Game- for error measurement four parameters are considered
i) RMSE Root mean squared error
ii) MAE Mean absolute error
iii) RRSE Root relative squared error
iv) RAE Relative absolute error

Table3: Error Rate Evaluation Parameters for Chess End Game

Algo.	MAE	RMSE	RAE	RRSE
OneR	33.54	57.92	62.2	115.9
PART	1.06	8.81	2.13	17.6
Decision Table	7.59	16.23	15.22	32.5
DTNB	1.77	24.38	35.47	48.8
Ridor	1.38	11.73	2.76	23.49

Table 3 shows the four basic error rate parameters for the evaluation of five Rule based classification algorithms. PART algorithm had the least value for all four parameters .OneR algorithm had the highest value for all four parameters.

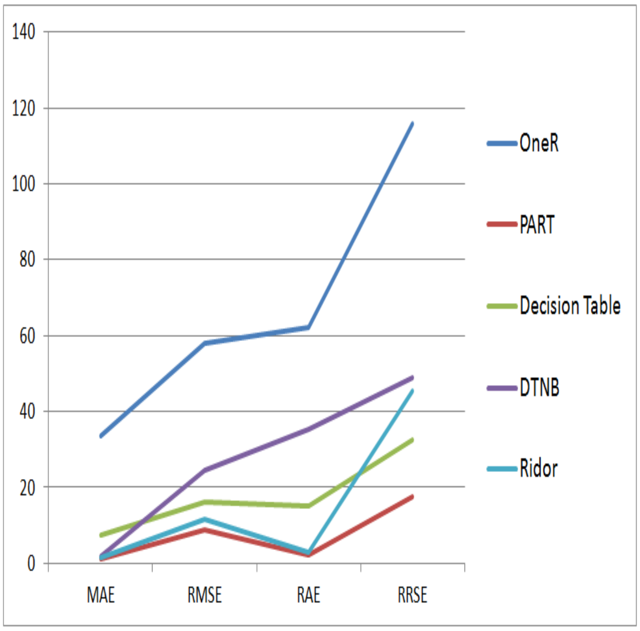

Figure3: Various Error Rate Evaluation Parameters

From Figure3 it is evident that PART algorithm have minimum error rate and have highest performance. Ridor algorithm has second minimum error rate and it also have over all good performance. Decision Table and DTNB average error rate and thus average performance. As seen in the graph OneR have high error rate and have poor performance as compare to other algorithm under study.

VIII. CONCLUSION

Five Classification rule based algorithms OneR, PART, Decision Table, DTNB and Ridor are introduced and experimentally evaluated using chess end game data sets. The rule based classification algorithms are experimentally compared based on number of classified instances, accuracy and error rate using WEKA tool. We used cross validation testing options for our experiments .From the result it is evident that PART is best rule based classification algorithm when compared to the other studied rule based algorithms. OneR algorithm had over all low performance for all the parameters.

IX. REFERENCES

[1] Ali, Shawkat, and Kate A. Smith. "On learning algorithm selection for classification." Applied Soft Computing 6.2 (2006): 119-138.

[2]Buddhinath, Gaya, and Damien Derry."A simple enhancement to One Rule Classification." Department of Computer Science & Software Engineering. University of Melbourne, Australia (2006).

[3] 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): 05-09.

[4] Duch,Włodzisław " Rule discovery" Encyclopedia of Systems Biology 2013, pp 1879-1883.

[5] Giraud-Carrier, C., and Olivier Povel. "Characterising data mining software."Intelligent Data Analysis 7.3 (2003): 181-192.

[6] Lawrence, Rick L., and Andrea Wright. "Rule-based classification systems using classification and regression tree (CART) analysis." Photogrammetric engineering and remote sensing 67.10 (2001): 1137-1142.

[7] Padhy, Neelamadhab, Dr Mishra, and Rasmita Panigrahi. "The survey of data mining applications and feature scope." arXiv preprint arXiv:1211.5723 (2012). [8] Phyu, Thair Nu. "Survey of classification techniques in data mining."Proceedings of the International MultiConference of Engineers and Computer Scientists. Vol. 1. 2009.

[9] Tayel , Salma, et al. "Rule-based Complaint Detection using RapidMiner", Conference: RCOMM 2013, At Porto, Portugal, Volume: 141 - 149

[10] Vijayarani1, S, M. Muthulakshmi ."Evaluating The Efficiency O f Rule Techniques for File Classification". International Journal of Research in Engineering and Technology eISSN: 2319-1163 | pISSN: 2321-7308.

[11] Vijayaran S, Sudha. "An Effective Classification Rule Technique for Heart Disease Prediction",.International Journal of Engineering Associates (IJEA), ISSN: 2320-0804, Vol.1, Issue 4, P.No.81-85, February 2013.

[12] Qin, Biao, et al. "A rule-based classification algorithm for uncertain data." *Data Engineering, 2009. ICDE'09. IEEE 25th International Conference on*. IEEE, 2009.

WEB REFERENCES

[12]http://www.gartnerweb.com/public/static/hotc/hc00086148.html.

[13] http://mydatamining.wordpress.com/2008/04/14/rule-learner-or-rule-induction

[14] B.R. Gaines and P. Compton. Induction of ripple-down rules applied to modeling large databases.

[15] http://archive.ics.uci.edu/ml/