Rule Based Classifiers for Vote Pattern Analysis

AUNG NWAY OO[1], THIN NAING[2]
[1, 2] *University of Information Technology, Myanmar*

Abstract- ***Classification is the operation of determining class of the data by forming a model that makes use of data whose categories are previously determined. Data mining techniques are frequently used to form a classifier that determines belonging class of a new data among the predetermined classes. This paper intends to provide comparative analysis of the rule based classifiers for vote pattern analysis. Analyzing the performance of rule based classifiers algorithms namely Decision Table, JRip, OneR, PART and Ridor. The goal of this paper is to specify the best technique from the rules classification technique under the vote dataset and also provide a comparison result each classifier.***

Indexed Terms- Rule based classifier, Decision Table, JRip, OneR, PART, Ridor

I. INTRODUCTION

Data mining is an essential step of knowledge discovery process by analyzing the massive volumes of data from various perspectives and summarizing it into useful information. Data mining is widely used in various application domains such as market analysis, credit assessment, stock market, fraud detection, fault diagnosis in production system, hazard forecasting, medical discovery, buying trends analysis, knowledge acquisition and science exploration.

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 Modeling 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.

Classification is one of the most common tasks in machine learning where given two or more different sets of example data, the learner needs to construct a classifier to distinguish between the different classes. Classification enables us to categorize records in a large dataset into predefined set of classes. The classes are defined before studying or examining records in the dataset. It also enables us to predict the future behavior of that sample data. 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 [1].

The rest of the paper is organized as follows. Section 2 provides the related work and section 3 presents the overview of rule based classifiers. Experimental results of each classifier are discussed in section 4. Finally, conclusion of this study was provided in section 5.

II. LITERATURE REVIEW

Rule based classification algorithms are popular in classification task of data mining. Vivek kshirsagar, et al. proposed the intrusion detection system using the rule based classifiers. They used the data mining algorithms to compute activity patterns from system audit data and extract predictive features from the patterns. Machine learning algorithms are then applied to the audit records that are processed according to the feature definitions to generate intrusion detection rules [2]. In paper [3], discussed a new predictive modeling approach known as rule based classification as an alternative technique to Neural Networks, Bayesian Networks, Decision Trees and Association Rules along with its advantages and working principle. Comparative analysis of hypothyroid dataset was performed by using some benchmarking classification algorithms like Naïve Bayes, Bayesian Network, JRip, OneR and PART. These classification algorithms are applied on hypothyroid health database for the purpose of finding better techniques for classification [4].

The study of Namrata Singh and Pradeep Singh [5] performed a comparative analysis of rule based classifiers in order to generate human interpretable rules for diagnosing CKD. Various rule-based approaches for comparison that have been used in the paper are JRip, CART, Conjunctive Rule, C4.5, NNge, OneR, Ridor, PART, and Decision Table-Naive Bayes (DTNB) hybrid classifier. The study concludes that among all the conventional classifiers cited, DTNB is best rule-based classifier. V. Veeralakshmi [6] analyzed the performance of 3 Rules classifiers algorithms namely JRIP, RIDOR, Decision Table. The Iris datasets are used for calculating the performance by using the cross validation parameter. The goal of this paper is to specify the best technique from the rules classification technique under the Iris datasets and also provide a comparison result which can be used for further analysis.

In [7] experimental evaluation of rule based classification algorithm is performed using WEKA open source tool and 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 and cross validation testing technique is considered for experiment. Dr. Neeraj Bhargava, et al. [8] presented the rule based classification algorithms for malicious detection. The work deals with classifications of malicious code per their impact on user's system & distinguishes threats on the muse in their connected severity.

In this paper rule based classifiers (Decision Table, JRip, OneR, PART and Ridor) are used for vote pattern analysis.

III. RULE BASED CLASSIFIER

Rule-based machine learning (RBML) is a term in computer science intended to encompass any machine learning method that identifies, learns, or evolves 'rules' to store, manipulate or apply. The defining characteristic of a rule-based machine learner is the identification and utilization of a set of relational rules that collectively represent the knowledge captured by the system. This is in contrast to other machine learners that commonly identify a singular model that can be universally applied to any instance in order to make a prediction [9]. Following rule based classifier are used in this paper.

Decision Table: Decision Table is an accurate method for numeric prediction from decision trees and it is an ordered set of If-Then rules that have the potential to be more compact and therefore more understandable than the decision trees [10].

Jrip: JRip (RIPPER) is one of the basic and most popular algorithms. Classes are examined in growing 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 decision 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 [11].

OneR: OneR, short for "One Rule", is a simple classification algorithm that generates a one-level decision tree. OneR is able to deduce typically simple, yet precise, classification rules from a set of instances. OneR is also able to handle missing values and numeric attributes showing flexibility in spite of simplicity. The OneR algorithm creates one rule for each attribute in the training data, then selects the rule with the minimum error rate as its „one rule [12].

PART: PART is a separate-and-conquer rule learner. The algorithm producing sets of rules called decision lists, which are planned set of rules. A new data is compared to each rule in the list in turn, and the item is assigned the class of the first matching rule. PART builds a partial C4.5 decision tree, in each iteration and makes the "best" leaf into a rule [13].

Ridor: Ripple Down Rule learner (Ridor) is also a direct classification method. It constructs the default rule. An incremental reduced error pruning is used to find exceptions with the smallest error rate, finding the best exceptions for each exception, and iterating. The most excellent exceptions are created by each exceptions produces the tree-like expansion of exceptions. The exceptions are a set of rules that predict classes other than the default. IREP is used to create exceptions [14].

IV. EXPERIMENTAL RESULTS

The data set includes votes for each of the U.S. This data set contains the 435 instances and 2 classes: democrats and republicans. There are 267 vote patterns for democrats and 168 for republicans. The attributes lists are described in the following table.

Table 1. Attribute Information

1. Class Name: 2 (democrat, republican)
2. handicapped-infants: 2 (y,n)
3. water-project-cost-sharing: 2 (y,n)
4. adoption-of-the-budget-resolution: 2 (y,n)
5. physician-fee-freeze: 2 (y,n)
6. el-salvador-aid: 2 (y,n)
7. religious-groups-in-schools: 2 (y,n)
8. anti-satellite-test-ban: 2 (y,n)
9. aid-to-nicaraguan-contras: 2 (y,n)
10. mx-missile: 2 (y,n)
11. immigration: 2 (y,n)
12. synfuels-corporation-cutback: 2 (y,n)
13. education-spending: 2 (y,n)
14. superfund-right-to-sue: 2 (y,n)
15. crime: 2 (y,n)
16. duty-free-exports: 2 (y,n)
17. export-administration-act-south-africa: 2 (y,n)

10 fold cross validation was used to test the dataset. The testing results of each classifier are described in the following tables.

Table 2. Statistic of rule based classifiers

Classifiers	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
Decision Table	0.966	0.077	0.952	0.966	0.959	0.981	democrat
	0.923	0.034	0.945	0.923	0.934	0.981	republican
JRip	0.951	0.042	0.973	0.951	0.962	0.942	democrat
	0.958	0.049	0.925	0.958	0.942	0.942	republican
OneR	0.948	0.030	0.981	0.948	0.964	0.959	democrat
	0.970	0.052	0.921	0.970	0.945	0.959	republican
PART	0.966	0.083	0.949	0.966	0.957	0.949	democrat
	0.917	0.034	0.945	0.917	0.931	0.949	republican

Ridor	0.936	0.048	0.969	0.936	0.952	0.944	democrat
	0.952	0.064	0.904	0.952	0.928	0.944	republican

Table 3. Average statistic of rule based classifiers

Classifiers	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
Decision Table	0.949	0.061	0.949	0.949	0.949	0.981
JRip	0.954	0.044	0.955	0.954	0.954	0.942
OneR	0.956	0.039	0.958	0.956	0.957	0.959
PART	0.947	0.064	0.947	0.947	0.947	0.949
Ridor	0.943	0.054	0.944	0.943	0.943	0.944

Table 4. Accuracy results of rule based classifiers

	Decision Table	JRip	OneR	PART	Ridor
Accuracy	94.9425 %	95.4023 %	95.6322 %	94.7126 %	94.2529 %

V. CONCLUSION

This paper analyses the rule based classifiers for vote pattern analysis. The performance of the vote dataset is calculated using the cross validation technique. The results indicate that accuracy is achieved above 94 %. From the experimental results, it is observed that the OneR algorithm performs better than other algorithms.

REFERENCES

[1] Aditi Mahajan, Anita Ganpati, "Performance Evaluation of Rule Based Classification Algorithms", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 10, October 2014.

[2] Vivek kshirsagar, Madhuri S.Joshi, "Rule Based Classifier Models For Intrusion Detection System", International Journal of Computer Science and Information Technologies, Vol. 7 (1) , 2016, 367-370

[3] Srinivas Murti, Mahantappa, "Using Rule Based Classifiers for the Predictive Analysis of Breast Cancer Recurrence", Journal of Information Engineering and Applications www.iiste.org ISSN 2224-5782 (print) ISSN 2225-0506 (online) Vol 2, No.2, 2012.

[4] Dr. Vaishali S. Parsania, Dr. N. N. Jani, Navneet H Bhalodiya, "Applying Naïve bayes, BayesNet, PART, JRip and OneR Algorithms on Hypothyroid Database for Comparative Analysis", INTERNATIONAL JOURNAL OF DARSHAN INSTITUTE ON ENGINEERING RESEARCH & EMERGING TECHNOLOGIES Vol. 3, No. 1, 2014.

[5] Namrata Singh and Pradeep Singh, "Rule Based Approach for prediction of Chronic Kidney Disease: A Comparative Study", iomedical & Pharmacology Journal Vol. 10(2), 867-874 (2017).

[6] V. Veeralakshmi, "Ripple down Rule learner (RIDOR) Classifier for IRIS Dataset", International Journal of Computer Science Engineering (IJCSE), May 2015.

[7] Aditi Mahajan, Anita Ganpati," Performance Evaluation of Rule Based Classification Algorithms", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 10, October 2014.

[8] Dr. Neeraj Bhargava, Aakanksha Jain, Abhishek Kumar, Dr. Dac-Nhuong Le, " Detection of Malicious Executables Using Rule Based Classification Algorithms",

Proceedings of the First International Conference on Information Technology and Knowledge Management, 2018.

[9] https://en.wikipedia.org/wiki/Rule-based_machine_learning

[10] Murat Koklu, et al., "APPLICATIONS OF RULE BASED CLASSIFICATION TECHNIQUES FOR THORACIC SURGERY", Joint International Conference (Technology Innovation and Industrial Management), 2015 May.

[11] Anil RAJPUT, Ramesh Prasad Aharwal, Meghna Dubey, S.P. Saxena, (2011) "J48 and JRIP Rules for E-Governance Data." IJCSS-448.

[12] Gaya Buddhinath and Damien Derry, "A Simple Enhancement to One Rule Classification." Department of Computer Science & Software Engineering University of Melbourne, Australia, 2006.

[13] Eibe Frank, Ian H. Witten, "Generating Accurate Rule Sets Without Global Optimization". In: Fifteenth International Conference on Machine Learning, 144-151, 1998.

[14] Gaines, B.R., Paul Compton, J. 1995. Induction of Ripple-Down Rules Applied to Modeling Large Databases, Intell. Inf. Syst. 5(3):211-228.