EVALUATING THE EFFICIENCY OF RULE TECHNIQUES FOR FILE CLASSIFICATION

S. Vijayarani[1], M. Muthulakshmi[2]

[1] Assistant Professor, [2] M. Phil Research Scholar, Department of Computer Science, School of Computer Science and Engineering, Bharathiar University, Coimbatore, Tamilnadu, India.
vijimohan_2000@yahoo.com*. *abarajitha.uma@gmail.com

Abstract

Text mining refers to the process of deriving high quality information from text. It is also known as knowledge discovery from text (KDT), deals with the machine supported analysis of text. It is used in various areas such as information retrieval, marketing, information extraction, natural language processing, document similarity, and so on. Document Similarity is one of the important techniques in text mining. In document similarity, the first and foremost step is to classify the files based on their category. In this research work, various classification rule techniques are used to classify the computer files based on their extensions. For example, the extension of computer files may be pdf, doc, ppt, xls, and so on. There are several algorithms for rule classifier such as decision table, JRip, Ridor, DTNB, NNge, PART, OneR and ZeroR. In this research work, three classification algorithms namely decision table, DTNB and OneR classifiers are used for performing classification of computer files based on their extension. The results produced by these algorithms are analyzed by using the performance factors classification accuracy and error rate. From the experimental results, DTNB proves to be more efficient than other two techniques.

Index Terms: *Data mining, Text mining, Classification, Decision table, DTNB, OneR*

----------***----------

1. INTRODUCTION

Data mining is the practice of searching through large amounts of computerized data to find useful patterns or trends. Data mining is the process of knowledge discovery where knowledge is gained by analyzing the data store in very large repositories which are analyzed from various perspectives and the result is summarized it into useful information. Data mining is also known as Knowledge Discovery in Data (KDD). There are various research areas in data mining such as web mining, text mining, image mining, statistics, machine learning, data organization and databases, pattern detection, artificial intelligence and other areas.

Text mining or knowledge discovery from text (KDT) deals with the machine supported analysis of text. It uses methods from information retrieval, natural language processing (NLP) information extraction and also connects them with the algorithms and methods of Knowledge discovery of data, data mining, machine learning and statistics. Current research in the area of text mining tackles problems of text representation, classification, clustering, or the search and modeling of hidden patterns. [5]

Text mining usually involves the process of structuring the input text (usually parsing, along with the accumulation of some derived linguistic features and the removal of others, and consequent insertion into a database), deriving models within the structured data, and to finish evaluation and interpretation of the output. High quality in text mining typically refers to some combination of relevance of relevance, innovation, and interestingness. Various stages of a text-mining process can be combined together into a single workgroup. [10]

Some of the important applications of text-mining include Data Mining Competitive Intelligence, Records Management, Enterprise Business Intelligence, National Security, E-Discovery, Intelligence Scientific discovery especially Life Sciences, Search or Information Access and Social media monitoring. Some of the technologies that have been developed and can be used in the text mining process are information extraction, topic tracking, summarization, classification, clustering, concept linkage, information conception, and question answering [4]. In this new and current era of technology, developments and techniques, efficient and effective, document classification is becoming a challenging and highly required area to capably categorize text documents into mutually exclusive categories. Text categorization is an upcoming and vital field in today's world which is most importantly required and demanded to efficiently categorize various text documents into different categories.

The rest of this paper is organized as follows. Section 2 illustrates the review of literature. Section 3 discusses the classification rule classifier and the various algorithms used

for classification. Experimental results are analyzed in Section 4 and Conclusions are given in Section 5.

2. LITERATURE REVIEW

C. Lakshmi Devasena et al [9] discussed the effectiveness of Rule-Based classifiers for classification by taking a sample data set from UCI machine learning repository using the open source machine learning tool. An evaluation of different rule based classifiers used in data mining and a practical guideline for selecting the most suited algorithm for a classification is presented and some empirical criteria for describing and evaluating the classifiers are given. The performances of the classifiers were measured and results are compared using the Iris Data set. Among nine classifiers (Conjunctive Rule Classifier, Decision Table Classifier, DTNB Classifier, OneR Classifier, JRIP Classifier, NNGE Classifier, PART Classifier, RIDOR Classifier and ZeroR Classifier) NNGE Classifier performs well in the classification problem. OneR classifier, RIDOR Classifier and JRIP classifier are coming in the next category to classify the data.

Biao Qin, Yuni Xia et al [4] proposed a new rule-based classification and prediction algorithm called uRule for classifying uncertain data. Uncertain data often occur in modern applications, including sensor databases, spatial-temporal databases, and medical or biology information systems. This algorithm introduces new measures for generating, pruning and optimizing rules. This new measures are computed considering uncertain data interval and probability distribution function. Based on the new values, the optimal splitting attribute and splitting value can be identified and used for classification and prediction. The uRule algorithm can process uncertainty in both numerical and categorical data. The experimental results show that uRule has excellent performance even when data is highly uncertain.

Mohd Fauzi bin Othman et al [13] investigated the performance of different classification or clustering methods for a set of large data. The algorithm tested are Bayes Network, Radial Basis Function, Rule based classifier, Pruned Tree, Single Conjunctive Rule Learner and Nearest Neighbors Algorithm. The dataset breast cancer with a total data of 6291 and a dimension of 699 rows and 9 columns will be used to test and justify the differences between the classification methods or algorithms. Consequently, the classification technique that has the potential to significantly improve the common or conventional methods will be suggested for use in large scale data, bioinformatics or other common applications. Among the machine learning algorithm tested, Bayes network classifier has the potential to significantly improve the conventional classification methods for use in medical or bioinformatics field.

Dr. S. Vijayarani et al [19] discussed the classification rule techniques in data mining 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.

3. METHODOLOGY

Document similarity is one of the main tasks in the area of text mining. The essential step of document similarity is to classify the documents based on some criteria. [7] In this section, various classification algorithms are used to classify the files based on their extension which are stored in the computer hard disk. (For example: pdf, doc, txt and so on). The methodology of the research work is as follows.

1. Dataset – Computer Files can be collected from the system hard disk.
2. Classification Rule Algorithms
   - Decision Table
   - DTNB
   - OneR
3. Performance factors
   - Classification accuracy
   - Error rate
4. Best Technique among classification rule algorithms
   - DTNB

3.1 Dataset

To compare these data mining classification techniques, computer files can be collected from the system hard disk and a synthetic data set is created. This dataset has 31994 instances and four attributes namely file name, file size, file extension and file path.

3.2 Classification Rule Techniques

Classification of documents involves assigning class labels to documents indicating their category. Classification algorithms are widely used in various applications. 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. [12] There are various classification rule algorithms such as Decision table, JRip, Ridor, DTNB, PART, OneR, and so on. In this research, we have analyzed classification rule algorithms namely Decision table, DTNB and OneR.

3.2.1 Decision Table

The algorithm, decision table, is found in the Weka classifiers under Rules. The simplest way of representing the output from machine learning is to put it in the same form as the input. The use of the classifier rules decision table is described as building and using a simple decision table majority classifier. The output will show a decision on a number of attributes for each instance. The number and specific types of attributes can vary to suit the needs of the task.

Two variants of decision table classifiers are available. The first classifier, called DTMaj (Decision Table Majority) returns the majority of the training set if the decision table cell matching the new instance is empty, that is, it does not contain any training instances. The second classifier, called DTLoc (Decision Table Local), is a new variant that searches for a decision table entry with fewer matching attributes (larger cells) if the matching cell is empty. This variant therefore returns an answer from the native region. [14]

3.2.2 DTNB

This is for building and using a decision table/naive bayes hybrid classifier. Every point in the search, the algorithm estimates the value of dividing the attributes into two disjoint subsets: one for the decision table, and the other for naive Bayes. A forward selection search is used at each step, then the selected attributes are exhibited by naive Bayes and the remainder by the decision table and all attributes are modeled by the decision table initially. At each step, the algorithm dropping an attribute entirely from the model. [16]

The algorithm for learning the combined model (DTNB) proceeds in much the same way as the one for stand-alone DTs. At each point in the exploration it estimates the merit associated with splitting the attributes into two disjoint subsets: one for the DT, the other for NB. The class probability estimates of the DT and NB must be combined to generate overall class probability estimates. Assuming X> is the set of attributes in the DT and X⊥ the one in NB, the overall class probability is computed as

$$Q(y \mid X) = \alpha \times QDT(y \mid X>) \times QNB(y \mid X\perp)/Q(y),$$

Where QDT(y | X>) and QNB(y | X ⊥) are the class probability estimates obtained from the DT and NB respectively, α is a normalization constant, and Q(y) is the prior probability of the class. All probabilities are predictable using Laplace corrected observed counts.

3.2.3 OneR

The OneR algorithm creates a single rule for each attribute of training data and then picks up the rule with the least error rate. To generate a rule for an attribute, the most recurrent class for each attribute value must be established. The most recurrent class is the class that appears most frequently for that attribute value. A rule is a set of attribute values destined to their most recurrent class with which the attribute based on. Pseudo-code for OneR algorithm is

```
For each attribute A,
For each value VA of the attribute, make a rule as follows:
Add up how often each class appears
Locate the most frequent class Cf
Generate a rule when A=VA; class attribute value = Cf
End For-Each
Compute the error rate of all rules
End For-Each
Select the rule with the smallest
```

The number of training data instances which does not agree with the binding of attribute value in the rule produces the error rate. OneR selects the rule with the least error rate. If two or more rules have same error rate, then the rule is selected at random. [2]

4. EXPERIMENTAL RESULTS

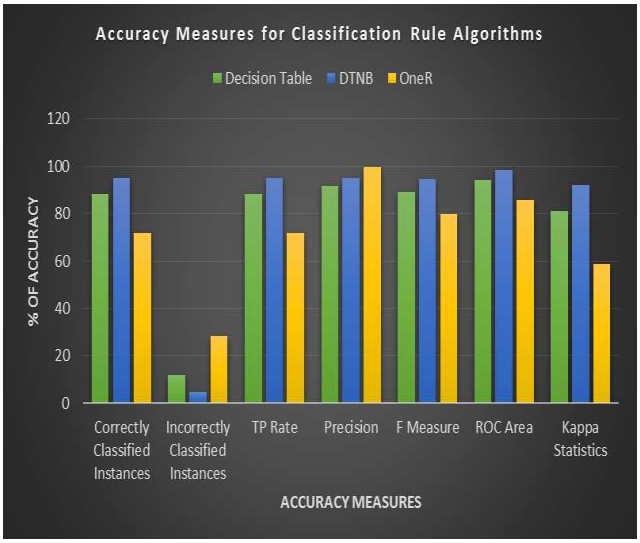
4.1 Accuracy Measure

The following table shows the accuracy measure of classification techniques. They are the correctly classified instances, incorrectly classified instances, True Positive rate, F Measure, Precision, Receiver Operating Characteristics (ROC) Area and Kappa Statistics. TP Rate is the ratio of play cases predicted correctly cases to the total of positive cases. F Measure is a way of combining recall and precision scores into a single measure of performance. [18] Precision is the proportion of relevant documents in the results returned. ROC Area is a traditional to plot the same information in a normalized form with 1-false negative rate plotted against the false positive rate.

Table-1: Accuracy Measure for Classification Rule Algorithms

Algorithm	Classifier		
	Decision Table	DTNB	OneR
Correctly Classified Instances	88.21	95.09	71.72
Incorrectly Classified Instances	11.79	4.91	28.28
TP Rate	88.20	95.10	71.70
Precision	91.70	95.20	99.60
F Measure	89.10	94.80	79.70
ROC Area	94.30	98.60	85.80
Kappa Statistics	81.09	92.08	58.95

Chart-1: Accuracy Measure for Classification Rule Algorithms



From the graph, it is observed that DTNB algorithm performs better than other algorithms. Therefore the DTNB classification algorithm performs well because it contains highest accuracy when compared to Decision table and OneR.

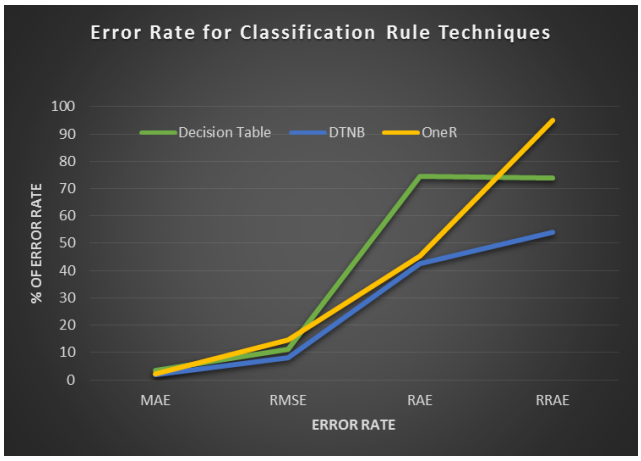
4.2 Error Rate

They are the 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). They are the 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). The mean absolute error (MAE) is defined as the quantity used to measure how close predictions or forecasts are to the eventual outcomes.[19] The root mean square error (RMSE) is defined as frequently used measure of the differences between values predicted by a model or an estimator and the values actually observed. Relative Absolute Error is a measure of the uncertainty of measurement compared to the size of the measurement. The root relative squared error defined as a relative to what it would have been if a simple predictor had been used. More specifically, this predictor is just the average of the actual values.

Table-2: Error Rate of Classification Rule Algorithms

Algorithm	MAE	RMSE	RAE	RRAE
Decision Table	3.45	11.25	74.46	73.92
DTNB	1.98	8.21	42.64	53.96
OneR	2.09	14.47	45.19	95.10

Chart-2: Error Rate for Classification Rule Algorithms



From the above graph, it is observed that Decision table and OneR algorithms attains highest error rate. Therefore, the DTNB classification algorithm performs well because it contains least error rate when compared to other algorithms.

CONCLUSIONS

Data mining is the process of discovering interesting knowledge from large amounts of data stored either in databases, data warehouses or other information sources. Text Mining is the automated or partially automated processing of text. In this research work, the classification rule algorithms namely Decision table, DTNB and OneR are used for classifying computer files which are stored in the computer. By analysing the experimental results it is observed that the DTNB (Decision Tree Naïve Bayes) classification technique has yields better result than other techniques. In future we tend to improve efficiency of performance by applying other data mining techniques and algorithms.

REFERENCES

[1]. Abdullah Wahbeh, Mohammed Al-Kabi, “Comparative Assessment of the performance of three WEKA text classifiers applied Arabic text.
[2]. Anshul Goyal, Rajni Mehta, “Performance Comparison of Naïve Bayes and J48 Classification Algorithms”.
[3]. Aurangzeb Khan, Baharum Baharudin, Lam Hong Lee, Khairullah khan, “A Review of Machine Learning Algorithms for Text-Documents Classification”.
[4]. Biao Qin, Yuni Xia, Sunil Prabhakar, Yicheng Tu, “A Rule-Based Classification Algorithm for Uncertain Data”.
[5]. BS Harish, DS Guru, S Manjunath, “Representation and Classification of Text Documents: A Brief Review”.
[6]. S.Deepajothi, Dr.S.Selvarajan, “A Comparative Study of Classification Techniques on Adult Data Set”
[7]. Ian H. Witten, Eibe Frank, “Data Mining Tools and Techniques practical Machine Learning”.

[8]. Kaushik Raviya, Biren Gajjar, "Performance Evaluation of Different Data Mining Classification Algorithm Using WEKA".
[9]. C. Lakshmi Devasena, T. Sumathi, V.V. Gomathi and M. Hemalatha, "Effectiveness Evaluation of Rule Based Classifiers for the Classification of Iris Data Set.
[10]. Lan Huang, David Milne, Eibe Frank, Ian H. Witten, "Learning a Concept-based Document Similarity Measure".
[11]. Mahendra Tiwari, Manu Bhai Jha, Om Prakash Yadav, "Performance analysis of Data Mining algorithms in Weka".
[12]. Mirza Nazura Abdulkarim, "Classification and Retrieval of Research Papers: A Semantic Hierarchical Approach".
[13]. Mohd Fauzi bin Othman, Thomas Moh Shan Yau, "Comparison of Different Classification Techniques using WEKA for Breast Cancer".
[14]. Petra Kralj Novak, "Classification in WEKA".
[15]. Mrs. Sayantani Ghosh, Mr. Sudipta Roy, Prof. Samir K. Bandyopadhyay, "A tutorial review on Text Mining Algorithms".
[16]. Sunila Godara, Ritu Yadav, "Performance analysis of clustering algorithms for character recognition using Weka tool".
[17]. Trilok Chand Sharma, Manoj Jain, "WEKA Approach for Comparative Study of Classification Algorithm".
[18]. Dr. S.Vijayarani, S.Sudha, "Comparative Analysis of Classification Function Techniques for Heart Disease Prediction".
[19]. Dr. S.Vijayarani, S. Sudha, "An Effective Classification Rule Technique for Heart Disease Prediction".

BIOGRAPHIES

Dr. S. Vijayarani has completed MCA, M.Phil and PhD in Computer Science. She is working as Assistant Professor in the School of Computer Science and Engineering, Bharathiar University, Coimbatore. Her fields of research interest are data mining, privacy, security, bioinformatics and data streams. She has published papers in the international journals and presented research papers in international and national conferences.

Mrs. M Muthulakshmi has completed M.Sc in Computer Science and Information Technology. She is currently pursuing her M.Phil in Computer Science in the School of Computer Science and Engineering, Bharathiar University, Coimbatore. Her fields of interest are Data Mining, Text Mining and Semantic web mining.