International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 4 | Issue 3| ISSN: 2456-3307

Comparative Study on Various Text Mining Algorithms in Data Mining

M.Prakash[1], A.Jesudasan[2]

[1]Assistant Professor, Department of Computer Science, Shanmuga Industries Arts and Science College, Tamilnadu, India

[2]Department of Computer Science, Shanmuga Industries Arts and Science College, Tamilnadu, India

ABSTRACT

This paper describes about text mining from the source of data mining. Data mining is nothing but an extraction of hidden knowledge from the huge database. There are lot of domains in data mining as text mining, image mining, sequential pattern mining, web mining and so. Here text mining can be used for extracting the information of the text using various algorithms using data mining software called WEKA. The data sets are taken from the UCI repository for performing the text mining techniques.

Keywords: Text mining, Data mining, WEKA, UCI repository, Algorithms

I. INTRODUCTION

Data mining is a technique which can be used for extracting the hidden knowledge from the huge database. The data mining can be classified into various domains named as text mining, image mining, sequential pattern mining, and web mining and so. Now, we are going to discuss about the text mining, how the information can be extracted from the database of text mining. The text mining has various fields like information retrieval, document similarity, information extraction, clustering, classification and so. Searching the similar document has an important role in text mining and document management. Classification is one of the main tasks in document similarity. It is used to classify the documents based on their category. Text mining also referred as text data mining which is similar to data analytics. Text mining is the process of deriving the highly valuable information from the text. Text mining can involve the process of structuring the input text, deriving the patterns within the structure of the data, and finally evaluation and interpretation of the output can execute. Text mining tasks includes the following methods as text categorization, text clustering, concept extraction, sentiment analysis, document summarization, and entity relational modeling. The main goal of text mining is to turn the text into the data using the application called Natural Language Processing (NLP). The term text analytics is modified as text mining by the author named Ronen Feldman. The term text analytics is also describes the application of text mining to respond the business demerits independently with queries and analyze the fielded numerical data.

II. TEXT MINING

Text mining or knowledge discovery from text (KDT) deals with the machine supported analysis of text. The text mining is a method of extracting the text and retrieves the highly valuable information's. It uses methods from information retrieval, information extraction and natural language processing (NLP) and also connects them with the

algorithms and methods of Knowledge discovery of data, data mining, machine learning and statistics. This method of text mining process cannot do any mining process without the help of any algorithms. In this paper, there are three Meta classification algorithms have been used for text mining in a comparative manner. Finally resulted which algorithm will produce the high accuracy in execution of the information retrieved.

The three Meta classification algorithms are named below:

1. Attribute selected classifier.
2. Filtered classifier.
3. Logit Boost

The above mentioned three algorithms are used for mining process in the text. These algorithms are used for classifying the computer files based on their extension. For example, .docx, .pdf, .xls, .ppt, and so. The performance of Meta algorithms are analyzed by applying the performance factor such as classification accuracy and error rate. The current research in the area of text mining tackles the problems like text representation, classification, clustering or searching the hidden patterns. It is used to describe the application of data mining techniques to automated discovery of useful or interesting knowledge from unstructured or semi-structured text. The procedure of synthesizing the information by analysing the relations, the patterns, and the procedures among textual data semi-structured or unstructured text.

III. APPLICATIONS

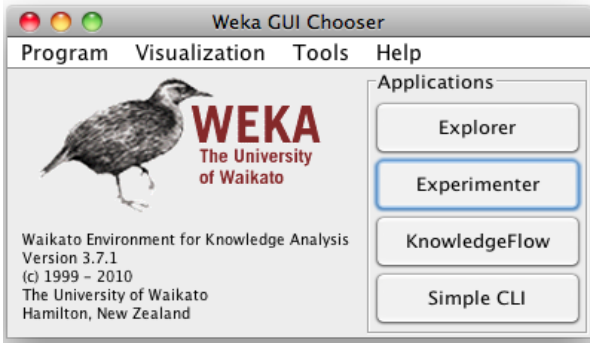
- Enterprise Business Intelligence,
- Data Mining Competitive Intelligence,
- E-Discovery,
- National Security,
- Intelligence Scientific Discovery,
- Records Management,
- Search Or Information Access And
- Social Media Monitoring.

IV. WEKA TOOL

WEKA is a machine learning software written in java and developed by the University of Waikato, New Zealand. It is free open source software licensed under the GNU general public License. The term Waikato Environment for Knowledge Analysis is shortly called as WEKA. Weka contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to these functions. It can be used in many different application areas, in particular for educational purposes and research. Advantages of Weka are mentioned below:

- Free availability under the GNU General Public License.
- Portability, because it is implemented in the Java programming language and it runs on any modern computing platform.
- A comprehensive collection of data preprocessing and modeling techniques.
- Ease of use due to its graphical user interfaces.

Weka's main user interface is the Explorer, but essentially the same functionality can be accessed through the component-based Knowledge Flow interface and from the command line. There is also the Experimenter, which allows the systematic comparison of the predictive performance of Weka's machine learning algorithms on a collection of datasets. There are several features in explorer provide access to the main components such as pre-process, classify, cluster, visualize panel, select attribute.



V. METHODOLOGY

Text classification is one of the important research issues in the field of text mining where the documents are classified with supervised knowledge. The main objective of this process is to find the best classification algorithm among:

- Attribute Selected Classifier,
- Filtered Classifier and
- Logit Boost.

The methodology of this paper work is as collection of the data set from the UCI repository, and implement that data into the Meta classification algorithms, and checks about the performance factor of these algorithms and finally producing the best algorithm, which can produce the best result by having high accuracy of data and low error rate in execution.

Here, the computer files from the system hard disk can be taken as dataset for text mining process. By taken these dataset, it will be implemented into the above mentioned three Meta classification algorithms and proceed for the mining process. After the implementation of data set into the algorithms, the performance factors can be calculated by producing the classification accuracy and error rate. From this methodology, we are finalizing the best algorithm by monitoring which will produce the best result.

VI. DATASET

A dataset can be collected from the computer systems, which are stored in the hard disk. The dataset can contains minimum of 9000 instances and four attributes namely file name, file size, extension and file path. Weka data mining tool is used for analyzing the performance factor of the classification algorithms.

VII. CLASSIFICATION ALGORITHMS

Classification is an important data mining technique with broad applications. It is used to classify each item in a set of data into one of predefined set of classes or groups. Classification algorithm plays an important role in document classification. There are various Meta classification algorithms such as Attribute Selected Classifier, Bagging, Decorate, Vote, Filtered Classifier, Logit Boost, END, Rotation Forest, and so on. Now, we have analyzed three Classification Meta Algorithms. The algorithms are namely Attribute Selected Classifier, Filtered Classifier and Logit Boost.

A. Attribute Selected Classifier

Dimensionality of training and test data is reduced by attribute selection before being passed on to a classifier. Some of the important options in attribute selected classifier as Classifier, Debug, Evaluator, and Search. Here, the base classifiers are used and from debug method it set to true and the classifier may output additional information to the console. Evaluator set the attribute evaluator to use and is used during the attribute selection phase before the classifier is invoked. Set the search method for using during the attribute selection phase before the classifier is invoked.

B. Filtered Classifier

From the filtered classifier class is used for running an arbitrary classifier on data that has been passed through an arbitrary filter. Similar to classifier, the structure of the filter is based exclusively on the

training data and test instances will be processed by the filter without changing their structure. Some of the important options in Filtered classifier are Classifier, Debug, and Filter.

C. Logit Boost

Logit Boost algorithm is an extension of Ada boost algorithm. It replaces the exponential loss of Ada boost algorithm to conditional Bernoulli likelihood loss. This Class is used for performing additive logistic regression. This class performs classification using a regression scheme as the base learner, and can handle multiclass problems.

VIII. EXPERIMENTAL RESULT ACCURACY AND ERROR RATE

There are various measures used for classification accuracy such as true positive rate, precision, F Measure, ROC Area, and kappa Statistics. The 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. 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 I
Accuracy Measures of Classification Algorithms

Parameter	Attribute selected classifier	Filtered classifier	Logit boost
Correctly classified instances	95.44	97.12	97.91
Incorrectly classified instances	4.56	2.88	2.09
TP rate	95.40	97.10	97.90
Precision	95.30	95.40	98.10
F measure	94.90	96.10	97.70
ROC area	99.00	99.80	99.90
Kappa statistics	94.25	96.37	97.37

From the above mentioned table, the accuracy measures of classification algorithms the values executed for Logit Boost algorithm produces a high performance factor in accuracy by comparing with other two algorithms.

IX. ERROR RATES

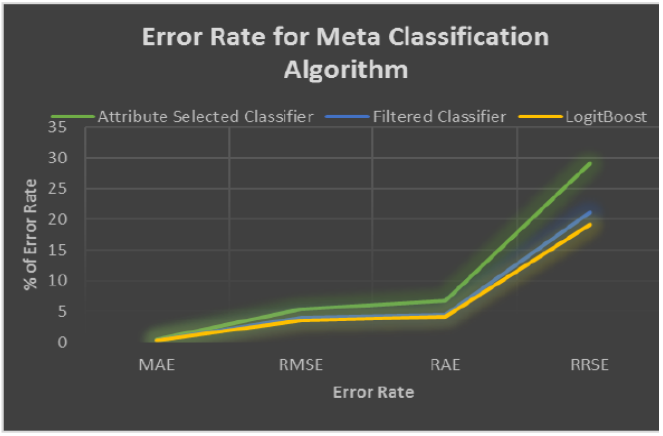
There are some types of errors as mentioned below,

- 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).

Table IIError rate

Algorithm	MAE	RMSE	RAE	RRSE
Attribute selected classifier	0.46	5.41	6.69	29.11
Filtered classifier	0.31	3.94	4.45	21.19
Logit boost	0.29	3.56	4.18	19.13

The mean absolute error (MAE) is defined as the quantity used to measure how close predictions or forecasts are to the eventual outcomes. 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 error is a measure of the uncertainty of measurement compared to the size of the measurement. The root relative squared error is defined as a relative to what it would have been if a simple predictor had been used.



From the above table and figure we conclude that the Logit Boost algorithm produces the low error rate than the attribute selected classifier and filtered classifier.

X. CONCLUSION

Data mining can be defined as the extraction of useful knowledge from large data repositories. Text mining is a technique which extracts information from both structured and unstructured data and also finding patterns which is novel and not known earlier. Here, the classification Meta algorithms are used for classifying computer files which are stored in the computer. The Classification Meta algorithms include three techniques namely Attribute Selected Classifier, Filtered Classifier and Logit Boost. By analyzing the experimental results it is observed that the Logit Boost classification technique has yields better result than other techniques.

XI. REFERENCES

[1] Abdullah Wahbeh H, Mohammed Al-Kabi., "Comparative Assessment of the Performance of Three WEKA Text Classifiers Applied to Arabic Text", Vol. 21, No. 1, pp. 15- 28, 2012.

[2] Abdullah Wahbeh H, Qasem Al-Radaideh A, Mohammed Al-Kabi N, and Emad Al-ShawakfaM., "A Comparison Study between Data Mining Tools over some Classification Methods".

[3] Artur Ferreira., "Survey on Boosting Algorithms for Supervised and Semi-supervised Learning".

[4] Christophe Giraud-Carrier., "Meta learning - A Tutorial".

[5] ChristophGoller, Joachim Löning., Thilo Will, Werner Wolff., "Automatic Document Classification: A thorough Evaluation of various Methods".