International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 11 | Nov -2016 www.irjet.net p-ISSN: 2395-0072

A COMPARATIVE ANALYSIS OF META AND TREE CLASSIFICATION ALGORITHMS USING WEKA

T.Sathya Devi[1], Dr.K.Meenakshi Sundaram[2],
(Sathya.kgm24@gmail.com[1] , lecturekms@yahoo.com[2])

[1](M.Phil Scholar, Department of Computer Science Erode Arts and Science College, Erode, Tamilnadu, India)
[2](Associate Professor, Department of Computer Science, Erode Arts and Science College, Erode, Tamilnadu, India)

ABSTRACT - *Data mining is one of the most knowledge research areas in the field of computer science. Data mining techniques are used for separation the hidden knowledge from the large databases. There are various research domains in data mining such as text mining, web mining, visual mining, spatial mining, knowledge mining and distributed mining. The purpose of text mining is to process unstructured knowledge, extract meaningful numeric indices from the text and thus make the information contained in the text accessible to the various data mining algorithms. There are dissimilar methods in text mining such as information retrieval, information extraction, document classification, natural language processing. Searching of akin documents has an important role in text mining and document management. Classification is one of the main tasks in document similarity. Searching for similar documents is an important problem in text mining. The first and essential step of document akin is to classify the documents based on their category. In this paper work, we have analysed the performance of Meta and Tree classifiers for classifying the files. There are two algorithms in Meta classifier namely LogitBoost and AdaBoost. In Tree classifier has three algorithms namely NBTree and ADTree. The performances of Meta and Tree classifiers are analysed by applying different performance factors. From the experimental results, it is observed that the Meta classifier is more efficient than Tree classifier.*

Keywords : Text mining, Weka ,ADTree , NBTree, LogitBoost and AdaBoost, Classification accuracy .

I.INTRODUCTION

With the advent of World Wide Web, amount of data on web increased tremendously. Although, such a huge accumulation of information is valuable and most of this information is texts, it becomes a problem or a challenge for humans to identify the most relevant information or knowledge. So text classification helps to overcome this challenge. Text classification is the act of dividing a set of input documents into two or more classes where each document can be said to belong to one or multiple classes [1]. Text Classification is a text mining technique which is used to classify the text documents into predefined classes. Classification can be manual or automated. Unlike manual classification, which consumes time and requires high accuracy, Automated Text Classification makes the classification process fast and more efficient since it automatically categorizes document. Language is used as medium for written as well as spoken communication. With the use of Unicode encoding, text on web may be present in different languages. This will add complexity of natural language processing to text classification. Text Classification is combination of Text Mining as well as Natural Language Processing. It has many applications such as document indexing, document organization . This task is usually solved by combining Information Retrieval (IR) technology and Machine Learning (ML) technology which both work together to assign keywords to the documents and classify them into specific categories. ML helps to categorize the documents automatically and IR represents the text as a feature. This paper concentrates on analysis of text classifier work on different document Languages. Section II describes the classification using Weka. Section III discusses the Meta and Tree classifiers and the various algorithms used for classification.Experimental results are analysed in Section IV and Conclusions are given in Section V.

II. CLASSIFICATION USING WEKA

Weka is written in java and can run on any of the platform. We can say that Weka is a collection on of algorithms with the help of which real world problems can be solved. Algorithms can be applied either directly or to a dataset called from own java code. Data processing, classification, clustering, visualization regression and feature selection these techniques are supported by Weka. In Weka data is considered as an instances and features as attributes [6]. In this main user interface is the explorer but essential functionality can be attained by component based knowledge flow interface and command line whenever simulation is done than the result is divided into several sub items for easy analysis and evolution. One part in correctly or correctly classified instances partitioned into percentage value and

numeric value and subsequently kappa statistics mean absolute error and root mean squared error will in numeric value.

In data mining, an important problem is large data set of classification. For a database with a set of classes and number of records such that each record belongs to one of the given classes, the problem of classification is to decide the class to which a given record belongs. Here, it is concerned with a type of classification called supervised classification. In supervised classification, a training data set of records and for each of this set, the respective class to which it belongs is also known. As they represent rule, Decision tress are especially attractive in data mining environment.

Objectives of Research Work

The objectives of this research work are as following:

- ✓ To apply Meta and Tree algorithms AdaBoost , LogitBoost, ADTree and NBTree on Document dataset.
- ✓ Evaluation of results produced.
- ✓ Comparative analysis of results using parameters accuracy, execution time and error rate for AdaBoost , LogitBoost, ADTree and NBTree.
- ✓ To build a new enhanced method for classification of data.

III 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 research work is to find the best classification algorithm among Meta and Tree classifiers. The Methodology of the research work is as follows:

Classification : 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. In this research, we have analysed two classifiers namely Meta Learning and Tree . In Meta classifier, we have analysed two classification algorithms namely AdaBoost and LogitBoost , in Tree classifier we have analysed three classification algorithms such as ADTree and NBTree. The classifiers are mentioned in brief.

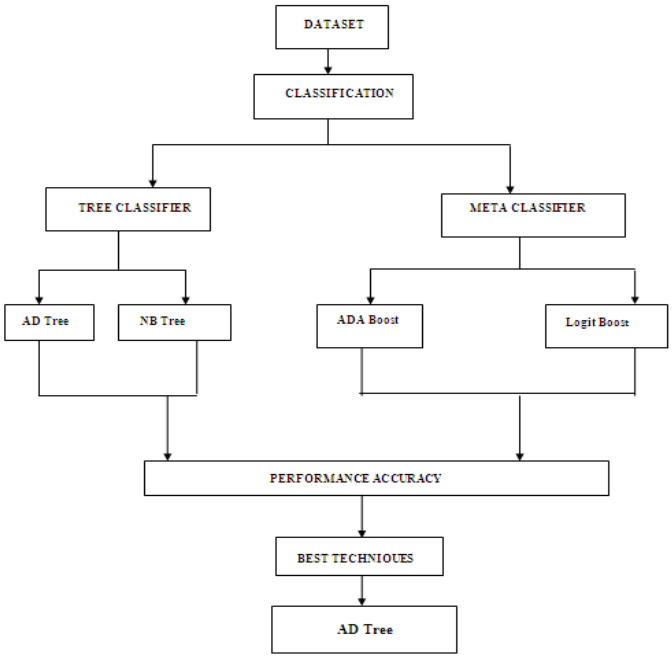

Fig 1 Overview of the Proposed Method

META CLASSIFIER

Meta classification indicates the usage of combination of multiple classifiers. This combination is carried out within three steps: In first step, multiple training subsets are constructed from a training set. In second step, each classifier is solely constructed according to both the algorithm and data training subset. In third step, the results of base classifiers are integrated and final results are obtained in a higher-level step called Meta classifier. There is also a Multiclass Classifier Meta classifier that does this for any binary class classifier. Various classification methods are proposed for the combination of classifiers (e.g. Tree-based model, Mixture model, Boosting, Bagging, etc) among which, some are briefly explained.

Adaboost : It is a machine algorithm, formulated by Yeave Freud and Robert Scapire. It is a meta-learning algorithm and used in conjunction with many other learning algorithms to improve their performance. [11]AdaBoost is an algorithm for constructing a"strong" classifier as linear combination. AdaBoost is adaptive only for this reason that subsequent classifiers built are weaked in favor of those instances misclassified by previous classifiers. AdaBoost is sensitive to noisy data and outliers. In some problems, however, it can be less susceptible to the over fitting problem than most learning algorithms. The classifiers it uses can be weak (i.e., display a substantial error rate), but as long as their

performance is not random (resulting in an error rate of 0.5 for binary classification), they will improve the final model.

Logitboost: LogitBoost is a boosting algorithm formulated by Jerome Friedmome, Trevor Hastie, and Robert Tibshirani. The original paper casts the Adaboost algorithm into a statistical framework. Specifically, if one considers AdaBoost as a generalized additive model and then applies the cost functional of logistic regression one can derive the LogitBoost algorithm.LogitBoost algorithm is an extension of Adaboost algorithm. It replaces the exponential loss of Adaboost algorithm to conditional Bernoulli likelihood loss. This Class is used for performing additive logistic regression. This classifier uses a regression scheme as the base learner, and can handle multiclass problems.

TREE CLASSIFIER

These are popular classification techniques in which at low- chart like tree structure is produced as a result in which each node denotes a test on attribute value and each branch represents an outcome of the test. They are also known as Decision Trees. The tree leaves represents the classes that are predicted.

AD Tree: An alternating decision tree (ADTree) is a machine learning method for classification. It generalizes decision trees and has connections to boosting. An alternating decision tree consists of decision nodes and prediction nodes. Decision nodes specify a predicate condition. Prediction nodes contain a single number. ADTrees always have prediction nodes as both root and leaves. An instance is classified by an ADTree by following all paths for which all decision nodes are true and summing any prediction nodes that are traversed.

NBTree: The naive Bayesian tree learner, NBTree (Kohavi 1996), combined naive Bayesian classification and decision tree learning. In an NBTree, a local naive Bayes is deployed on each leaf of a traditional decision tree, and an instance is classified using the local naive Bayes on the leaf into which it falls. The algorithm for learning an NBTree is similar to C4.5. After a tree is grown, a naive Bayes is constructed for each leaf using the data associated with that leaf. An NBTree classifies an example by sorting it to a leaf and applying the naive Bayes in that leaf to assign a class label to it. NBTree frequently achieves higher accuracy than either a naive Bayesian classifier or a decision tree learner.

Tool Used

In this research work, an open source tool named Weka is used. Weka is free open source data mining software which is based on a Java data mining library. Weka consists of various machine learning algorithms for different data mining applications. The algorithms are directly applied to dataset and results are generated in the form of tree. Weka contains various classifiers for classification [7], clustering, association, regression, pre-processing and visualization. Weka is also used for development of new machine learning schemes.

Data Set used

In our research work, we have used document dataset . The main focus of this research is performance and evaluation of Meta and Tree algorithms. There are many Meat and Tree algorithms in data mining but we focus mainly on AdaBoost , LogitBoost, ADTree and NBTree. This dataset contains 300 instances and 8 attributes namely Document Name, Document Author Document Published Year, Document Size, Extension ,Document Paths size, Extension and File path. Weka data mining tool is used for analyzing the performance of the classification algorithms.

IV. EXPERIMENTAL RESULTS

The main aim of this research proposal is to analyze the classification algorithms' performance for document data (output) based on the numerous input parameters as per Table 1. They are analyzed using Meta and Tree AdaBoost, LogitBoost, AD Tree and NB Tree algorithms. The WEKA application is used for the performance evaluation. Each classifier is applied for two testing beds - Cross Validation which uses 10 folds with 9 folds used for training each time and 1 fold is used for testing and Percentage Split which uses 2/3 of the dataset for training and 1/3 for testing is given as output. The screen shot of the WEKA preprocessing stage is shown in Figure 1.

Confusion Matrix : Metrics are measures that we can analysis and control without prior assumption about the data metrics relates a classifier to the process that produce the data without relying on the data itself.

Table.1 Confusion Matrix

		Predicted Class		
		Yes	No	Total
Actual Class	Yes	TP	FN	P
	No	FP	TN	N
	Total	P'	N'	P + N

In this, the classification of pre-processing is carried out based on all the values of taken Eight attributes. A comparative study of classification accuracy in AdaBoost, LogitBoost , ADTree and NBTree algorithm is carried out in this work. The TP Rate FT Rate and precision analysis is also carried out. The various formul as used for the calculation of different measures are as follows. The following formula is used to calculate the proportion of the predicted positive

cases, Precision P using TP = True Positive Rate and FP = False Positive Rate as,

$$\text{Precision (P)} = \frac{\text{No. of Relevant documents Retrieved}}{\text{Total No. of documents Retrieved}}$$

It has been defined that Recall or Sensitivity or True Positive Rate (TPR) means the proportion of positive cases that were correctly identified. It will be computed as

$$\text{Recall(R)} = \frac{\text{No. of Relevant documents Retrieved}}{\text{Total No of relevant documents}}$$

Where FN =False Negative Rate

Accuracy : It shows the proportion of the total number of instance predictions which are correctly predicted

$$\text{Accuracy} = \frac{TP + TN}{TP+TN+FP+FN}$$

Sensitivity : Sensitivity is the percentage of positive records classified correctly out of all positive records.

$$\text{Sensitivity} = \frac{TP}{(TP+FN)}$$

Specificity : Specificity is the percentage of positive records classified correctly out of all positive records.

$$\text{Specificity} = \frac{TN}{(TN+FP)}$$

F – Measure : The F – Measure can be computed as some average of the information retrieval precision and recall metrics.

$$\text{F –Measure} = \frac{2*\text{recall}*\text{precision}}{\text{precision} + \text{recall}}$$

ROC stands for Receiver Operating Characteristic. A graphical approach for displaying the trade-off between True Positive Rate (TPR) and False Positive Rate (FPR) of a classifier are given as follows.

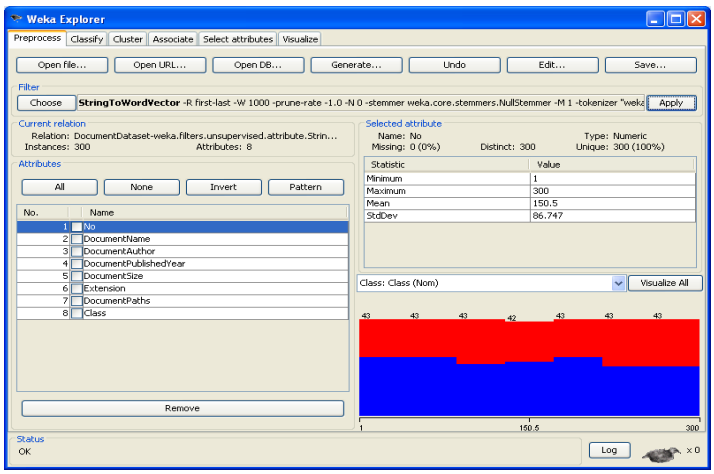

Fig 2 Preprocessing of document dataset in WEKA

The experimental results of basic classifiers are discussed in this section. To classify the correctly from the training data set, accuracy are calculated using classifiers. The accuracy of ADTree 92.67% and accuracy 91% is found in NBTree , 86%in LogitBoost algorithm and 83.33% in AdaBoost Algorithm. The confusion matrix helps us to find the various evaluation measures like accuracy, recall and precision, F-Measure. The classification accuracy of four algorithms AdaBoost, LogitBoost, AD Tree and NB Tree are observed from the Table 2 values of weighted average, which is available in the last row of each table. and Figure 3 shows the weighted average accuracy of the classification algorithm for the document data. The Figure 4 represents the comparison of the AdaBoost, LogitBoost, AD Tree and NB Tree classification algorithms based on the Table 3 values.

Table .2 Accuracy by weighted average

S.No	Parameter	TP	FP	Precision	Recall	F-measure
1	AdaBoost	0.833	0.143	0.855	0.833	0.834
2	LogitBoost	0.86	0.115	0.883	0.86	0.86
3	NBTree	0.91	0.078	0.917	0.91	0.91
4	ADTree	0.927	0.067	0.929	0.927	0.927

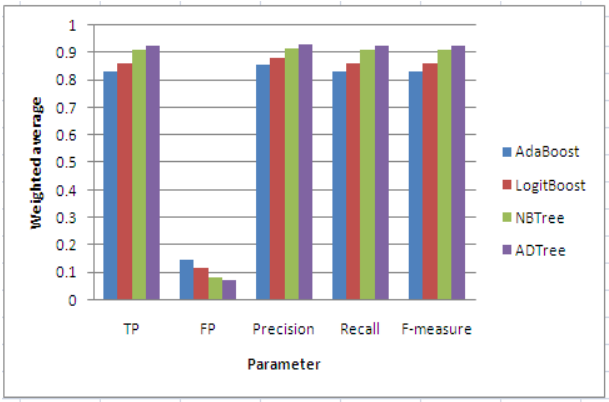

Fig 3 Weighted average of various parameters.

The following tables show the accuracy measure of classification techniques. They are the True Positive rate, F-Measure. The TP Rate is the ratio of play cases predicted correctly cases to the total of positive cases. It is a probability corrected measure of agreement between the classifications and the true classes. It is calculated by taking the agreement expected by chance away from the observed agreement and dividing by the maximum possible agreement.

F-Measure is a way of combining recall and precision scores into a single measure of performance. Recall is the ratio of relevant documents found in the search result to the total of all relevant documents. Precision is the proportion of relevant documents in the results returned.

Table.3 Performance accuracy of algorithm

Classification Algorithm	Correctly Classified Instances	Incorrectly Classified Instances
AdaBoost	83.33	16.67
LogitBoost	86	14
NBTree	91	9
ADTree	92.67	7.33

The accuracy of ADTree is 92.67%, NBTree 91% , LogitBoost 86%, AdaBoost 83.33%. The confusion matrix helps us to find the various evaluation measures like accuracy, TP,FP Recall and Precision, F-Measure.

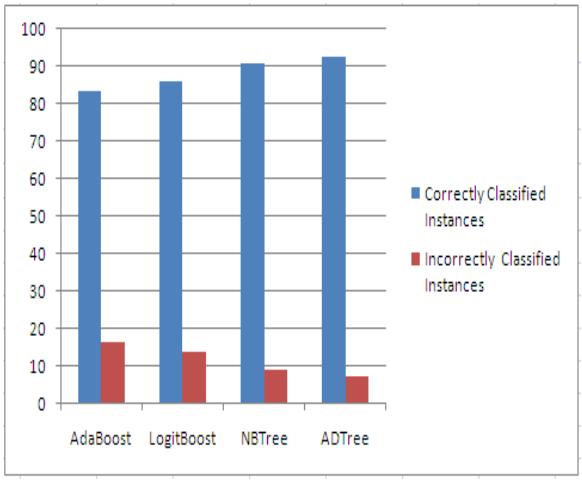

Fig 4 Performance comparison of algorithms.

From the above graph, it is analyzed that the ADTree algorithms performs better than the other algorithms. Therefore the ADTree classification algorithm performs well because it contains highest accuracy when compared to AdaBoost, LogitBoost, and NBTree.

V. 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. This research work evaluate the performances in terms of classification accuracy of AD Tree, NBTree and AdaBoost and LogitBoost algorithms using various accuracy measures like FP rate, TP rate, Recall, Precision and F-measure. The experimental results shows that the highest accuracy is Found in ADTree 92.67% and accuracy 91% is found in NBTree , 86%in LogitBoost algorithm and 83.33% in AdaBoost Algorithm. Based on the classification results of all the four algorithms, the performance of ADTree is better than the other three algorithms for the chosen data set.

REFERENCES

[1] Annan Naidu Paidi, " Data Mining: Future Trends and Applications", International Journal of Modern Engineering Research (IJMER), Vol.2, ISSN: 2249-6645 , Nov-Dec. 2012 .

[2] Manika Verma and Dr. Devarshi Mehta, "A Comparative study of Techniques in Data Mining " International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, Volume 4, Issue 4, April 2014.

[3] Monika D.Khatri and S.Dhande, " History and current and future trends of Data mining Techniques", International Journal of advance Research in computer science and management studies, vol.2 , ISSN: 2321-7782,March 2014.

[4] Anuradha Patra, "A Survey Report on Text Classification with Different Term Weighing Methods and Comparison between Classification Algorithms" International Journal of Computer Applications, (0975 – 8887) Volume 75– No.7, August 2013.

[5] Aaditya Jain," Text Classification by Combining Text Classifiers to Improve the Efficiency of Classification", International Journal of Computer Application, ISSN:2250-1797, Volume 6, March-April 2016.

[6] Femi Joseph, "Text Categorization Using Improved K Nearest Neighbor Algorithm**",** International Journal For Trends In Engineering & Technology, Volume 4 , ISSN: 2349 – 9303,April 2015.

[7] Mohammad Behrouzian Nejad, Iman Attarzadeh and Mehdi Hosseinzadeh , " An Efficient Method for Automatic Text Categorization", International Journal of Mechatronics, Electrical and Computer Technology ,Vol. 3(9), ISSN: 2305-0543 ,Oct 2013.

[8] Abdullah H. Wahbeh , "A Comparison Study between Data Mining Tools over some Classification Methods " (IJACSA) International Journal of Advanced Computer Science and Applications **,** vol 2, April 2014.

[9] Shweta Srivastava," Weka: A Tool for Data preprocessing, Classification, Ensemble, Clustering and Association Rule Mining "International Journal of Computer Applications, Volume 88 , ISSN: 0975 – 8887, February 2014.

[10] Trilok Chand Sharma and Manoj Jain, "WEKA Approach for Comparative Study of Classification Algorithm ",International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, ISSN : 2278-1021, April 2013.

[11] Aurangzeb Khan and Baharum Baharudin, "A Review of Machine Learning Algorithms for Text-Documents Classification", Journal of Advances In Information Technology, VOL. 1, February 2010.

[12] B.H.ChandraShekar and Dr.G.Shoba, "Classification Of Documents Using Kohonen's Self-Organizing Map", International Journal of Computer Theory and Engineering, Vol. 1, ISNN: 1793-8201, December:2009.

[13] Hanumanthappa, " India Language Text Documents

Categorization and Keyword Extgraction",IJCTA,vol.9(3), 2016.

[14] Pooja Bolaj, " A survey on text categorization techniques for India regional languages", International Journal of computer science and Information Technologies,vol.7,ISSN: 0975-9646,2016.

[15] Narayana Swamy ,"A Detailed study on Indian Languages Text Mining",International Journal of Computer Science and mobile computing, vol.3, ISSN: 2320-088X, Nov.2014.

ACKNOWLEDGMENT

Dr.K.Meenakshi Sundaram has completed MCA, M.Phil and PhD in Computer Science. he is working as Assistant Professor in the Department of Computer Science, Erode Arts and Science College, Erode. he fields of research interest are data mining, privacy, security and data streams.

T.Sathya Devi is currently Pursuing M.Phil from Erode Arts and Science College, Erode Her area of Interest is Natural Language Processing, Text Mining and Data Mining.