Bayes and Tree Classifiers for File Classification

Dr. S. Vijayarani[1] J. Ilamathi[2]
[1]Assistant Professor [2]M.Phil. Research Scholar
[1,2]Department of Computer Science & Engineering
[1,2]Bharathiar University Coimbatore, India

Abstract— This research work primarily focuses on classifying the files which are stored in the computer system based on their extension (.pdf, .jpg, .docx, .txt, and .png). The aim of this work is to analyze the performance of two different classification algorithms. Two types of classification algorithms used and tested in this work are Naïve Bayes Updatable for Bayes modeling and REP tree (Reduced Error Pruning Tree) classification algorithm for Tree modeling. From the experimental results, it is observed that the REP tree classifier performance is better than the Naïve Bayes Updatable classifier. The performance factors used are classification accuracy and error rate. This work is carried out in WEKA data mining tool.

Key words: Classification, Bayes, Naïve Bayes Updatable, REP tree

I. INTRODUCTION

Data mining is used to "mine" or "extract" knowledge from large quantity of data. Some of the important data mining domains are sequential pattern mining, spatial mining, web mining, text mining, medical mining, multimedia mining, image mining, structure mining and graph mining. The term "data mining" is primarily used by statisticians, database researchers and business communities. The term Knowledge Discovery in Databases (KDD) is generally used to refer to the overall process of discovering useful knowledge from data, where data mining is a particular step in this process.

This work concentrated on classifying the system files which are stored in computer system based on their extension. Different types of file extensions are pdf, docx, doc, txt, png, img and jpg etc. In order to perform the classification the bayes and tree induction algorithms are used. In bayes classifier, Naïve Bayes Updateable is used and in Tree induction, REP tree classification algorithm is used.

The paper is organized as follows. Section 2 provides the literature review. Proposed methodology is given in Section 3. Section 4 described the experimental results. Conclusion is given in Section 5.

II. LITERATURE REVIEW

S.Vijayarani et al. [10] discussed the performance evaluation of tree classification algorithms, J48, Random Forest and Random Tree algorithms. The experimental results are analyzed based on the classification accuracy. From the results, authors concluded that efficiency and accuracy of J48 is better than other algorithms.

Trilok Chand Sharma et al. [9] described the data mining techniques to process a dataset and identify the relevance of classification test data. Mining tools to solve large amounts of problems such as classification, clustering, association rule, neural networks. In this research paper, decision Tree classification algorithms are analyzed. These decision tree algorithms are used in medical, banking and stock market.

Methin Zontul et al., [4] presented an analysis by examining the wind power potential of Kırklareli province which is in the west of Turkey. Statistical data between 2001 and 2007 was used in this study. The data was obtained from Kırklareli branch of State Meteorological Service. In Kırklareli region, wind speed forecasts regarding the year 2013 were made for windpower plants that are supposed to be built. WEKA tool is used for the performance analysis. REPTree algorithm is used in this work.

Menaka [5] has presented text classification is the process of classifying the text documents based on words, phrases and word combinations with respect to set of predefined categories. Text classification has many applications such as mail routing, email filtering, content classification, news monitoring and narrow-casting. Keywords are extracted from documents to classify the documents.

Kalpesh Adhatrao [3] predicted the student's performance from their previous performances using the classification data mining technique. It is analyzed by using the data set which contains the student information like student name, gender, marks scored in the board examinations of classes X and XII, marks and rank in entrance examinations and results in first year of the previous batch of students. ID3 and C4.5 classification algorithms are used by the author for this data analysis.

III. METHODOLOGY

The main aim of this research work is to find the efficient classification algorithms among Naïve Bayes Updatable, and REP tree classifier for classifying the file names based on their extension.

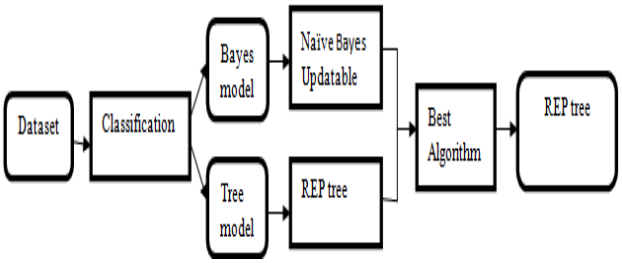

Fig. 1: System Architecture of Classification Algorithms

A. Data Set:

A synthetic file names dataset is created which contains the list of file names which are stored in the computer hard disk. Three different sizes of datasets i.e. 3000, 5000 and 7000 instances are used and seven attributes namely filename, file size, last change, last access, creation date, file extension, file path.

B. Classification Algorithms:

Classification is a data mining (machine learning) technique used to predict group membership for data instances. It is

an ordered set of related categories used to group data according to its similarities. There are many types of classification algorithms in data mining, this paper discussed about two types of classification algorithms, they are, Bayes classification algorithm (Naïve Bayes Updatable), and Tree classification algorithm (Reduced Error Pruning Tree (REP tree).

1) Rep Tree (Reduced Error Pruning Tree):

REP Tree algorithm is based on the principle of calculating the information gain with entropy and reducing the error arising from variance [7]. This method is firstly suggested by Quinlan [8]. With the help of this method, complexity of decision tree model is decreased by "reduced error pruning method" and the error arising from variance is reduced [7, 8]. Reduced Error Pruning is a technique in machine learning that reduces the size of decision trees by removing sections of the tree that provide little power to classify instances. The dual goal of pruning reduces complexity of the final classifier as well as better predictive accuracy by the reduction of over fitting and removal of sections of a classifier that may be based on noisy or erroneous data.

a) Algorithm for Rep Tree (Reduced Error Pruning Tree):

REP Tree (Reduced Error Pruning Tree) Algorithm

- Create a root node for the tree
- Check for the base case
- Apply Feature Selection using Genetic Search
- Best Tree = Construct a DT using training data
- Perform Cross validation
  1. Divide all examples into N disjoint subsets, E = E1, E2, ... EN
  2. For each i = 1, ..., N do
     2.1 Test set = Ei
     2.2 Training set = E - Ei
     2.3 Compute decision tree using Training set
     2.4 Determine performance accuracy Pi using Test set
  3. Compute N-fold cross-validation estimate of performance = (P1 + P2 + ... + PN)/N

Perform Reduced Error Pruning technique

- Find the attribute with the highest info gain (A_Best)

Perform Model complexity

- Partition S into S1,S2,S3... according to the value of A_Best
- Repeat the steps for S1, S2, S3
- Classification :
- For each tie D, apply the DT to determine its class

Fig. 2: Algorithm for Rep Tree (Reduced Error Pruning Tree)

2) Naïve Bayes Updateable:

The bayesian classification represents a supervised learning method as well as a statistical method for classification. Assumes an underlying probabilistic model and it allows us to capture uncertainty about the model in a just way by determining probabilities of the outcomes. Bayesian classification provides practical learning algorithms and prior knowledge and observed data can be combined. Bayesian Classification provides a useful perspective for understanding and evaluating many learning algorithms such as naïve bayes, naïve bayes kernel, and naïve bayes updateable.

The Naive update calculates the prior probabilities of a class, and the conditional prior probabilities for a set of features given a class.

Probability of class membership:

$$P_C = P(C)$$

Probability of feature F given a class C:

$$P_{F|C} = P(F|C)$$

There are two tables that are updated, one containing the class frequencies (probability of a class) and the other containing the probability of a feature given a class. The update reaches its final value in just one step.

IV. EXPERIMENTAL RESULTS

A. Accuracy Measure:

Table 1 shows the accuracy measures for classification tree techniques. The accuracy measures are Correctly Classified Instance, Incorrectly Classified Instance, Precision, True positive rate, F Measure, Receiver Operating Characteristics (ROC) Area and Kappa Statistics.

Algorithms and Instances		Parameter						
Algorithms	Instances	Correctly classified	In correctly classified	TP Rate	Precision	F-Measure	ROC Area	Kappa Statistics
Rep Tree classifier	3000	100	0	99.99	99.78	99.33	99.45	99.98
	5000	100	0	99.99	99.78	99.33	99.45	99.98
	7000	100	0	99.99	99.78	99.33	99.45	99.98
Naive Bayes Updatable classifier	3000	82.93	17.60	82.93	79.44	79.13	99.11	80.57
	5000	86.13	13.44	86.11	84.51	83.95	83.98	84.39
	7000	85.48	14.51	83.71	85.13	82.97	82.88	83.86

Table 1: Accuracy Measure

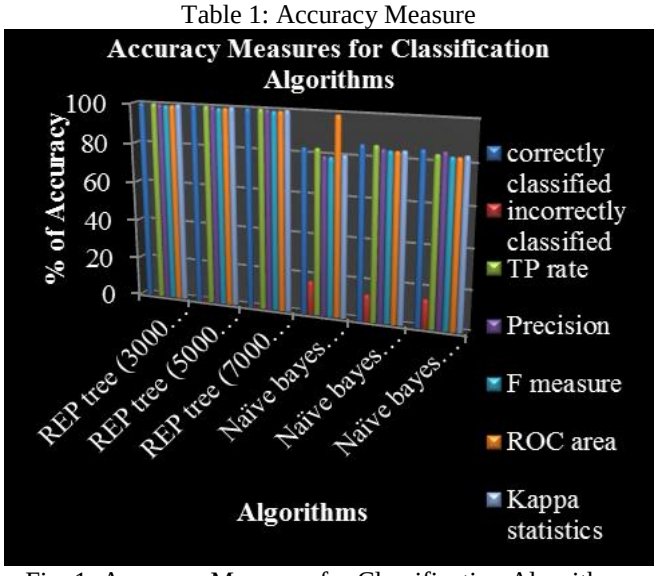

Fig. 1: Accuracy Measures for Classification Algorithms

From the Figure 1 it is observed that REP Tree (Reduced Error Pruning Tree) performance is better than Naïve Bayes Updateable algorithm.

Algorithms		Classification Accuracy
REP Tree	3000	100
	5000	100
	7000	100
Naïve Bayes Updatable	3000	82.93
	5000	86.11
	7000	83.71

Table 2: Overall Accuracy

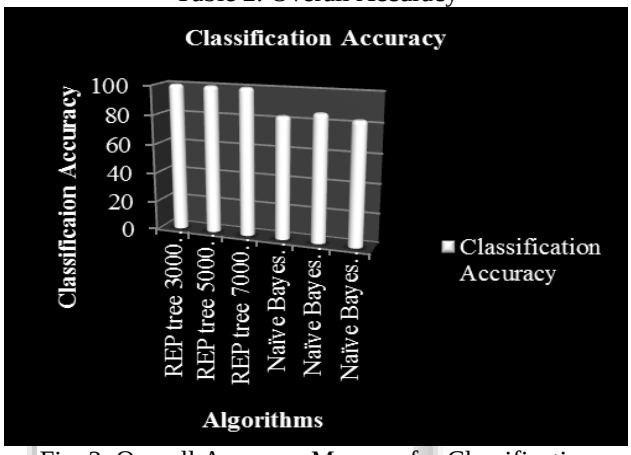

Fig. 2: Overall Accuracy Measure for Classification Algorithms

Figure 2 shows the overall accuracy of REP and Naïve Bayes Updateable algorithm.

B. Error Rate;

The different types of error rate factors are the Root Mean Square Error (R.M.S.E), Mean Absolute Error (M.A.E), Root Relative Squared Error (R.R.S.R), and Relative Absolute Error (R.A.E).

1) Root Mean-Squared Error (RMS):

It is the average of the squared differences between the each computed value of its corresponding actual value [8]. The RMS is nothing but the square root of the mean squared error. The important process is to square all the errors and if the negative values becomes positive.

The RMSE of a model prediction with respect to the estimated variable X_{model} is defined as the square root of the mean squared error:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(X_{obsi} - X_{model,i})^2}{n}}$$

Where X_{obs} is observed values and X_{model} is modeled values at time/place *i*.

2) Mean Absolute Error (MAE):

It is the average of errors between the actual values and the predicted values where all errors are considered as positive value [8].

$$\mathrm{MAE} = \frac{\mathrm{SAE}}{N} = \frac{\sum_{i=1}^{N}|x_i - \hat{x}_i|}{N}$$

Where:

- $\{x_i\}$ is the actual observations time series
- $\{\hat{x}_i\}$ is the estimated or forecasted time series
- SAE is the sum of the absolute errors (or deviations)
- N is the number of non-missing data points

3) Root Relative Squared Error (RRSE):

It is the sum of the squared errors to the sum of average errors of the actual value [8]. The mathematical expression is given as

Mathematically, the root relative squared error E_i of an individual program *i* is evaluated by the equation:

$$E_i = \sqrt{\frac{\sum_{j=1}^{n}(P_{(ij)} - T_j)^2}{\sum_{j=1}^{n}(T_j - \overline{T})^2}}$$

Where $P_{(ij)}$ is the value predicted by the individual program *i* for sample case *j* (out of *n* sample cases); T_j is the target value for sample case *j*; and $\overline{T}$ is given by the formula:

$$\overline{T} = \frac{1}{n}\sum_{j=1}^{n} T_j$$

For a perfect fit, the numerator is equal to 0 and $E_i = 0$. So, the E_i index ranges from 0 to infinity, with 0 corresponding to the ideal.

Algorithms and Instances		Parameters			
Algorithms	Instances	MAE	RMSE	RAE	RRSE
REP Tree Classifier	3000	0	0	0	0
	5000	0	0	0	0
	7000	0	0	0	0
Naïve Bayes Updatable Classifier	3000	0.03	0.04	24.84	55.24
	5000	0.02	0.03	19.09	48.56
	7000	0.01	0.03	20.70	5016

Table 3: Error Rate for Classification Algorithms

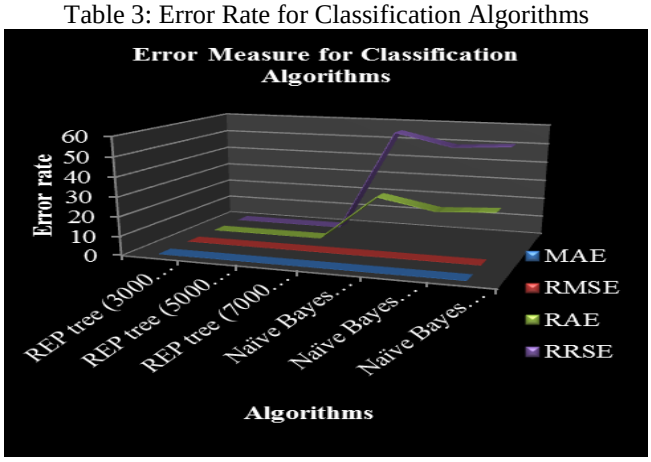

Fig. 3: Error rate Measure for Classification Algorithms

Figure 3 is analyze and compared REP Tree (Reduced Error Pruning Tree) and Naïve Bayes Updatable classifiers, the REP tree classifier is providing low error rate.

V. CONCLUSION

Data mining can be used to extract the useful knowledge from large data. In this paper, Bayes classifier and Tree classifier is used for classifying system files which are stored in the computer system. The classification algorithms include two techniques namely Bayes classifier for Naïve Bayes Updateable and Tree classifier for REP Tree (Reduced Error Pruning Tree). By analyzing the experimental results it is observed that the REP Tree (Reduced Error Pruning Tree) classification technique has produced better result than Naïve Bayes Updateable techniques.

REFERENCES

[1] Abdullah Wahbeh H, Mohammed A1-Kabi, 2012: "Comparative Assessment of the performance of three WEKA Text Classifiers Applied to Arabic Text", 2009 International Conference on Machine Learning and Computing IPCSIT vol.3 (2011) © (2011) IACSIT Press, SingaporeVol.21, No. 1, 15-28.
[2] Berry Michael W., (2004), "Automatic Discovery of Similar Words", in "Survey of Text Mining: Clustering, Classification and Retrieval", Springer Verlag, New York, LLC, 24-43.
[3] Kalpesh Adhatrao, Aditya Gaykar, "Predicting Students Performance Using ID3 and C4.5 Classification Algorithm", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.5, September 2013, page no: 39 - 52.
[4] Metin Zontul et al., "Wind Speed Forecasting Using Reptree and Bagging Methods In Kirklareli-Turkey", Journal of Theoretical and Applied Information Technology.
[5] Menaka.S, Radha.N, "Text Classification using Keyword Extraction Technique", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 12, December 2013, ISSN: 2277 128X, page no: 734 - 740
[6] Navathe, Shamkant B., and Elmasri Ramez, (2000), "Data Warehousing And Data Mining", in "Fundamentals of Database Systems", Pearson Education pvt Inc, Singapore, 841-872.
[7] Quinlan J (1987) "Simplifying decision trees", International Journal of ManMachine Studies, 27(3), pp. 221–234.
[8] Suguna .N, and Dr. K. Thanushkodi, "An Improved k-Nearest Neighbor Classification using Genetic Algorithm". IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 4, No 2, July 2010.
[9] Trilok Chand Sharma, Manoj Jain, 2013: 'WEKA Approach for Comparative Study of Classification Algorithm', Vol. 2, Issue 4. 4] Vishal Gupta and Gurpreet S. Lehal "A Survey of Text Mining Techniques and Applications" journal of emerging technologies in web intelligence, vol. 1, no. 1, AUGUST 2009
[10] Vijayarani S, M.Muthulakshmi "Performance Evaluation of classification Tree Algorithms for File Categorization", IFRSA's International Journal Of Computing |Vol 4|issue 1|Jan 2014,Page No: 430 - 435
[11] Witten IH, Frank E (2005) "Data mining: practical machine learning tools and techniques" – 2nd ed.. the United States of America, Morgan Kaufmann series in data management systems.
[12] Yaniv Gurwicz, Boaz Lerner, "Bayesian network classification using spline-approximated kernel density estimation", Pattern Recognition Letters 26 (2005) 1761–1771, www.elsevier.com/locate/patrec.

