RESEARCH ARTICLE OPEN ACCESS

Performance Comparison of RandomForest and Hoeffding Tree classifier using WEKA data mining tool on Car reviews data

S.A.Ghogare*, Dr.S.R.Kalmegh**
**(Shri.JJT University, Research Scholar,*
*** (Associate Professor, Dept. of Computer Science, SGBAU, Amravati, M.S.,*
Corresponding Author; S.A.Ghogare

ABSTRACT
The Size of data base is increasing day by day with rapid speed. The WEKA is data processing tool contain organized collection of state of art machine learning algorithm. However, convenient interactive graphical user interfaces are provided for data exploration, for setting up large-scale experiments on distributed computing platforms, and for designing configurations for streamed data processing. This paper has been carried out to make a performance evaluation of RandomForest and Hoeffding Tree classification algorithm. The paper sets out to make comparative evaluation of two Tree classifiers from WEKA RandomForest and Hoeffding Tree in the context of dataset of car reviews to maximize true positive rate and minimize false positive rate. The WEKA tool used for result processing. The results in the paper on dataset of car reviews also show that the efficiency and accuracy of RandomForest is excellent than Hoeffding Tree**.**
Keywords- Classification, Data mining, Hoeffding Tree, RandomForest, WEKA.

Date of Submission: 28-02-2019 Date of acceptance: 25-03-2019

I. INTRODUCTION

Today the rapid growth of internet, product related word-of-mouth conversation have migrated to online markets, creating active electronic communication that provide a wealth of information. The huge amount of data is generated from various resources. Disks and online storage make it too easy to postpone decisions about what to do with all this stuff, we simply get more memory and keep it all. In data mining, the data is stored electronically and the search is computerized or at least augmented by machine. Data mining is a topic that involves learning in a practical, non theoretical sense. We are interested in techniques for finding and describing structural patterns in data, as a tool for helping to explain that data and make predictions from it. Experience shows that in many applications of machine learning to data mining, the explicit knowledge structures that are acquired and the structural descriptions are at least as important as the ability to perform well on new examples. People frequently use data mining to gain knowledge, not just predictions [8].

Nowadays, more and more e-commerce platforms offer product reviews. A product review is a textual review of a customer or expert, who describes the characteristics of a product. A product rating on the other hand represents the customer's and expert opinion on a specified scale. In the given research paper we have used car review data set. Form Comparative analysis of RandomForest and Hoeffding Tree classifier.

II. DATA MINING TOOL

Data Mining is a powerful technology with great ability to help organizations focus on the most important information in their data center. It also predict future trends, behavior and with result. It also contains variety of analytical tools that used for data analysis. It allows users to analyze the data from many different aspects, categorize it, and summarize the identified relationships. There are many Data Mining tools are available such as the WEKA, KNIME, Orange, SPSS Clemantine, MATLAB, and NeuroShell etc. These tools provide a set of Data Mining methods and algorithms that help in better implementation of data and information available to users. The available Data Mining tools can be divided into two types which are open source/non-commercial software and commercial software. These types of tools have their own strengths and weaknesses in terms of data types and the application methods. From the given set of tool in my research work we have used WEKA tool.

III. WEKA

WEKA was developed at the University of Waikato in New Zealand; the name stands for Waikato Environment for Knowledge Analysis The system is written in Java and distributed under the terms of the GNU General Public License. It runs on

almost any platform and has been tested under and even on a personal digital assistant. It provides a uniform interface to many different learning algorithms, along with methods for pre and post processing and for evaluating the result of learning schemes on any given dataset. Weka provides implementations of learning algorithms that can be easily apply to dataset. It also includes a variety of tools for transforming datasets, such as the algorithms. This tool also supports the variety file formats for mining include ARFF, CSV, LibSVM, and C4.5.The easiest way to use Weka is through a graphical user interface called Explorer as shown in Fig. I. Fig. II Shows the ARFF File Opening by WEKA and Fig. III ARFF file processing by RandomForest.

Linux, Windows, and Macintosh operating systems

Fig. I: WEKA GUI Explore

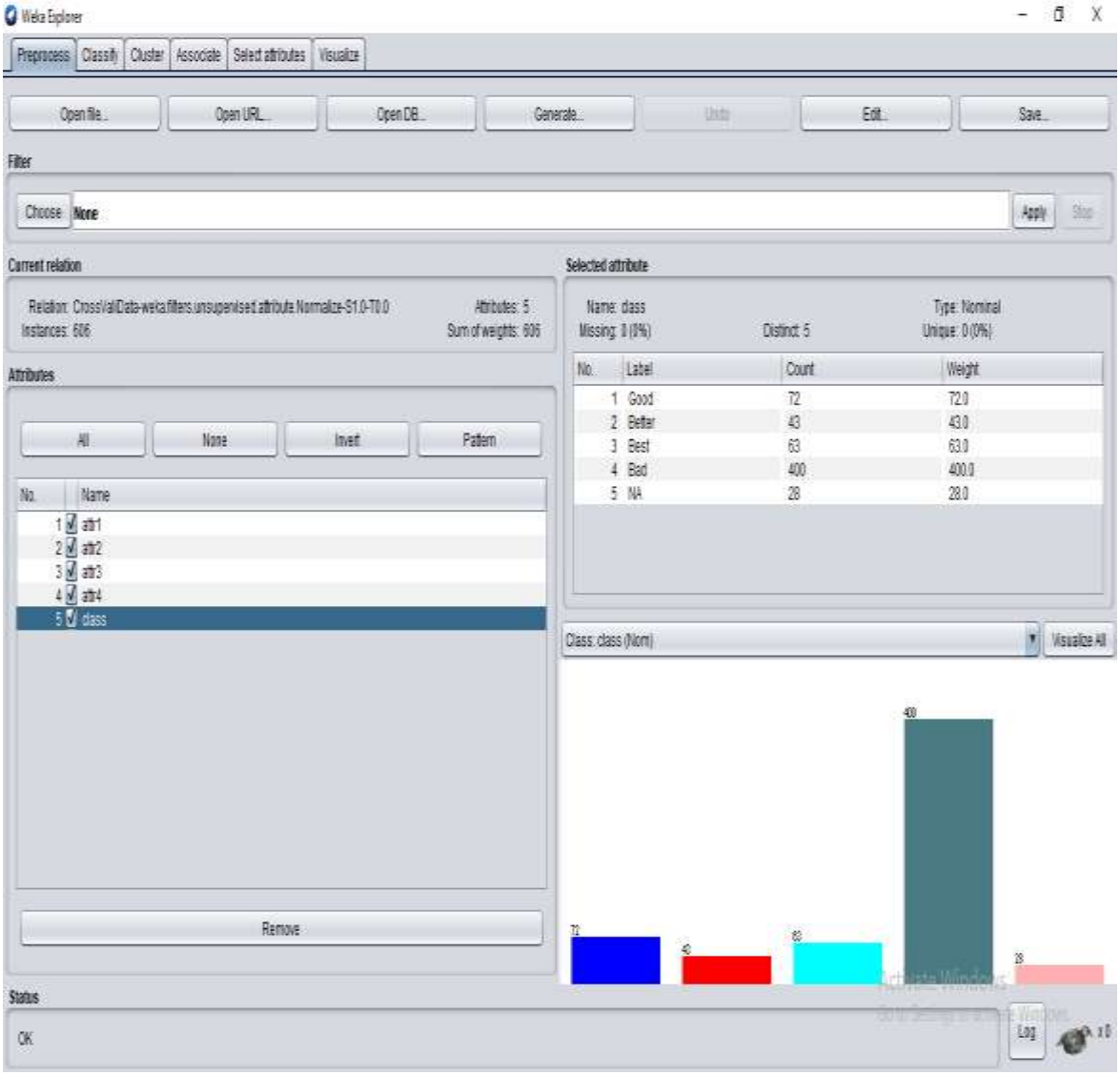

Fig. II: ARFF File Opening by WEKA

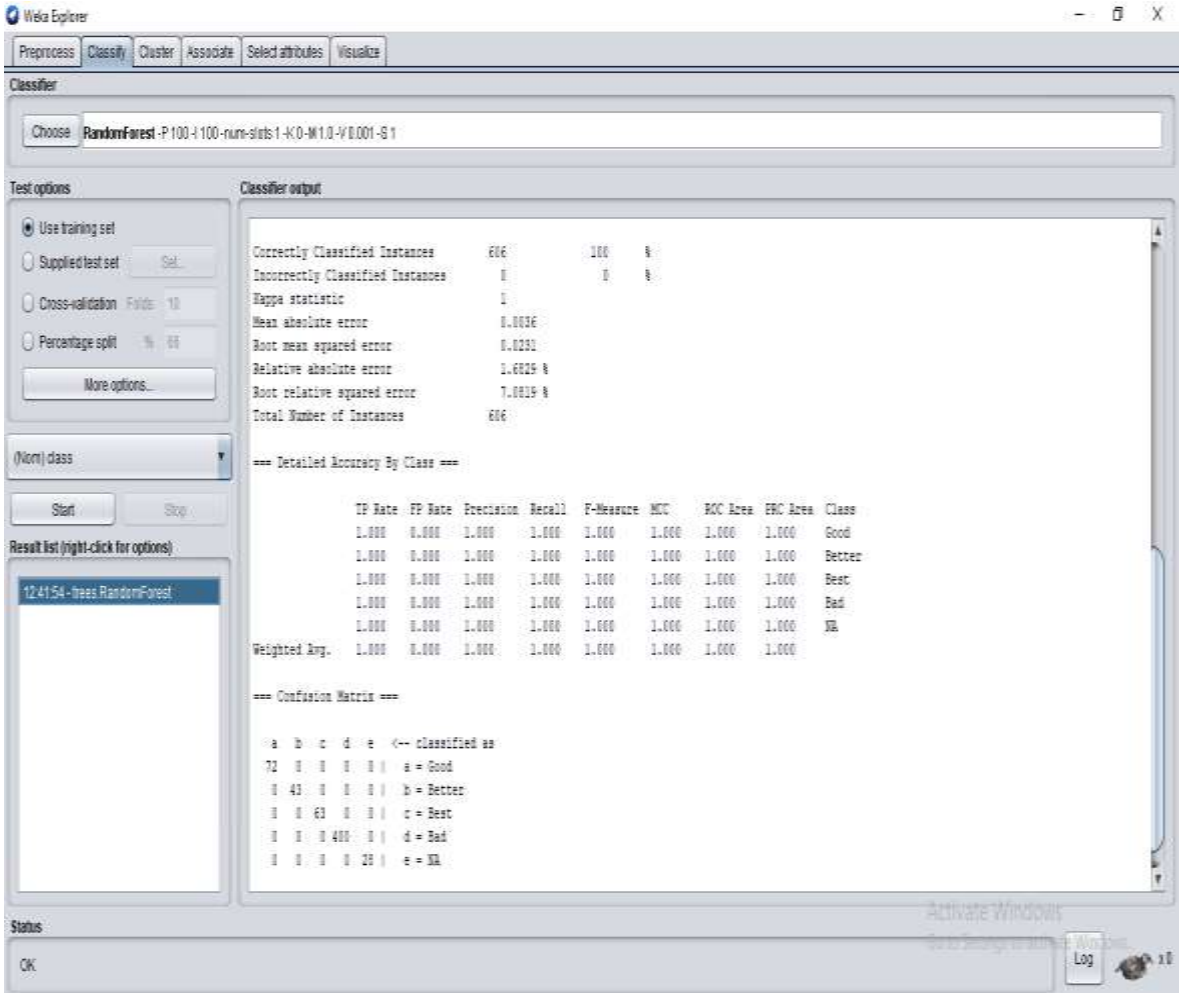

Fig. III: ARFF file processing by RandomForest

IV. CLASSIFICATION

The process in which idea and object are recognized, differentiated and understood is called as classification. Classifications also refer as Categorization. In the terminology of machine learning, classification is considered an instance of supervised learning, i.e. learning where a training set of correctly identified observation is available. The corresponding unsupervised procedure is known as clustering or cluster analysis and involved grouping data into categories based on measure of inherent similarity [11].

4.1 Decision tree or Classification tree

Classification (also known as classification trees or decision trees) is a data mining algorithm that creates a step-by-step guide for how to determine the output of a new data instance. The tree it creates is exactly that: a tree whereby each node in the tree represents a spot where a decision must be made based on the input, and to move to the next node and the next until one reach a leaf that tells the predicted output. Sounds confusing, but it's really quite straight forward [11].

4.2 RandomForest

RandomForest Build multiple decision trees and merge them together to get a more accurate and stable result. It is easy to use machine learning algorithm which is very flexible and produces great results most of the time, even without proper hyper-parameter tuning. The major advantages of random forest are that it can be used for both classification and regression problems and measure the relative importance of each feature on the prediction.

4.3 Hoeffding

Hoeffding tree uses the Hoeffding bound for construction and analysis of the decision tree. Hoeffding bounds used to decide the number of instances to be run in order to achieve a certain level of confidence. It is an incremental, anytime decision tree induction algorithm that is capable of learning from massive data streams.

V. SYSTEM DESIGN

In order to co-relate Reviews with the categories, a model based on the machine learning and XML search was designed. As an input to the model, various quality car reviews are considered which are available online from Cardekho.com, Carwale.com and other etc. Around 606 car reviews were collected on above repository. In order to extract context from the car reviews, the car reviews was process with stop word removal, stemming and tokenization on the car reviews contents. The car reviews then separated into 5 categories GOOD, BETTER, BEST, BAD, NA (not applicable) and then converted into the term frequency matrix for further analysis purpose. Due to classification in above 5 categories we are also able to find the GOOD, BETTER, BEST, BAD, NA count on every data set which help for market analysis, product rating and much more purposes. Based on this data, features (i.e. metadata) were extracted so that contextual assignment of the car reviews to the appropriate content can be done. This process is known as metadata processing [10].

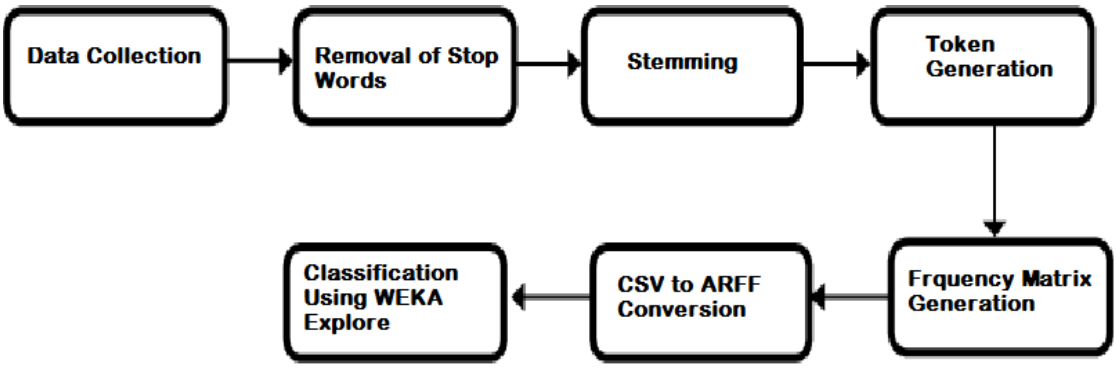

Fig. IV: Process Flow chart

VI. DATA COLLECTION

Hence it was proposed to generate car reviews data. Consequently the national and international resources were used for the research purpose. Data for the purpose of research has been collected from the various online resources .They are downloaded and after reading the car reviews they are manually classified into 12(Twelve) categories. There were 606 car reviews in total. The details are shown in following table. The attributes consider for this classification is based on GOOD, BETTER, BEST, BAD, NA count each classification having their own data dictionary and based on this they are classified, the review are made by expert and user. Hence, there will be drastic enhancement in e-Contents when we refer to the latest material available in this regards.

VII.PERFORMANCE INVESTIGATION

The car reviews so collected needed a processing. Hence as given in the design phase, all the review were processed for stop word removal, stemming, tokenization and ultimately generated the Frequency matrix based on GOOD, BETTER, BEST, BAD, NA count. Stemming is used as many times when review is printed, for a same there can be many variants depending on the tense used or whether it is singular or plural. Such words when processed for stemming, generates a unique word. Stop words needs to be removed as they do not contribute much in the decision making process. Frequency matrix thus generated can be processed for generating a model and the model so generated was used in further decision process. With the model discussed above, two tree classifier RandomForest, Hoeffding Tree, Were used on the data set of 606 car reviews. For processing WEKA APIs were used. In following tables showing figures for true positive rate and false positive rate. The 1.0 represent the best, whereas the worst is 0.0. The result after processing is given in the form of confusion matrix which is shown in Table II, IV. And table no III and V showing true positive and false positive rate of RandomForest and Hoeffding Tree [10].

TABLE I: Car data set Classification.

Sr. No	Car Companies	Numbers of reviews
1	Chevrolet	38
2	Fiat	27
3	Ford	36
4	Honda	47
5	Hyundai	59
6	Mahindra R Mahindra	63
7	Maruti Suzuki	95
8	Renault	53
9	Skoda	23
10	Tata Motors	90
11	Toyota	41
12	Volkswagan	34
	Total	**606**

TABLE II
CONFUSION MATRIX FOR RANDOMFOREST

CLASSIFIED AS	GOOD	BETTER	BEST	BAD	NA
GOOD	72	0	0	0	0
BETTER	0	43	0	0	0
BEST	0	0	63	0	0
BAD	0	0	0	400	0
NA	0	0	0	0	28

TABLE III
TABLE SHOWING TRUE POSITIVE AND FALSE POSITIVE RATE OF RANDOMFOREST

CLASS	TP RATE	FP RATE	PRECISION	RECALL	F-MEASURE	ROC AREA
GOOD	1.000	0.000	1.000	1.000	1.000	1.00
BETTER	1.000	0.000	1.000	1.000	1.000	1.00
BEST	1.000	0.000	1.000	1.000	1.000	1.00
BAD	1.000	0.000	1.000	1.000	1.000	1.00
NA	1.000	0.000	1.000	1.000	1.000	1.00
Weighted Avg.	1.000	0.000	1.000	1.000	1.000	1.00

TABLE IV
CONFUSION MATRIX FOR HOEFFDING TREE

CLASSIFIED AS	GOOD	BETTER	BEST	BAD	NA
GOOD	70	0	0	2	0
BETTER	0	39	1	3	0
BEST	0	0	27	36	0
BAD	0	7	3	390	0
NA	0	0	0	1	27

TABLE V
TABLE SHOWING TRUE POSITIVE AND FALSE POSITIVE RATE OF HOEFFDING TREE

CLASS	TP RATE	FP RATE	PRECISION	RECALL	F-MEASURE	ROC AREA
GOOD	0.972	0.000	1.000	0.972	0.986	0.989
BETTER	0.907	0.012	0.848	0.907	0.876	0.978
BEST	0.429	0.007	0.871	0.429	0.574	0.990
BAD	0.975	0.204	0.903	0.975	0.938	0.977
NA	0.964	0.000	1.000	0.964	0.982	0.994
Weighted Avg.	0.913	0.136	0.912	0.913	0.903	0.980

VIII. CONCLUSION

As per the previous performance investigation of Car review from dynamic resources can be done with the propose model, we used two classifier i.e. RandomForest and Hoeffding Tree to analyze the data sets. As a result it is found that RandomForest algorithm performs well in categorizing 606 instances all the car review. Overall Performance of RandomForest algorithm is acceptable, i.e. Correctly Classified Instances are 606 out of 606 the average percentage of this is 100%, Incorrectly Classified Instances is 0, the average percentage of this is 0.000%, whereas Hoeffding Tree algorithm works little bit less i.e. Correctly Classified Instances are 553 out of 606 the average percentage of this is 91.2541%, Incorrectly Classified Instances are only 53 the average percentage of this is 8.7459 %.

REFERENCES

[1]. Dr.Sushilkumar Rameshpant Kalmegh, Effective classification of indian news using classifier hyperpipes and naivebayes from weka, International Journal of Pure and Applied Eesearch in Engineering and Technology, (2016), ISSN: 2319-507X, Vol4, Iss 9.

[2]. S. R. Kalmegh, Comparative analysis of WEKA Data Mining Algorithm RandomForest, RandomTree and LADTree for Classification of Indigenous News Data, International Journal of Emerging Technology and Advanced Engineering, (2015), ISSN 250-2459, Vol 5, Iss 1.

[3]. King,M, A,and Elder,J, F, Evaluation of Fourteen Desktop Data Mining Tools, IEEE International Conference on Systems, Man and Cybernetics , (1998), ISSN: 1062-922X.

[4]. Wei Peng, Juhua Chen and Haiping Zhou, An Implementation of ID3: Decision Tree, Learning Algorithm Project of Comp 9417: Machine Learning University of New South Wales, School of Computer Science & Engineering, Sydney, NSW 2032, and Australia

[5]. Uzair Bashir & Manzoor Chachoo, Performance evaluation of j48 and bayes algorithms for intrusion detection system, International Journal of Network Security & Its Applications (IJNSA), (2017), Vol.9, Iss.4.

[6]. N.Landwehr,M.Hall & E.Frank, Logistic model trees,For Machine Learning, (2005), Vol.59, Iss.12.

[7]. Mahendra Tiwari, Manu Bhai Jha and OmPrakash Yadav, Performance analysis of Data Mining algorithms in Weka, IOSR Journal of Computer Engineering(IOSRJCE),2012,ISSN 2278-0661,Vol 6,Iss 3.

[8]. Sushilkumar Rameshpant Kalmegh, Comparative Analysis of WEKA Data Mining Algorithm RandomForest, RandomTree and LADTree for Classification of Indigenous News Data, International Journal of Emerging Technology and Advanced Engineering, ISSN 2250-2459, (2015), Vol 5, Iss1.

[9]. Sushilkumar Rameshpant Kalmegh, Effective classification of indian news using classifier hyperpipes and naivebayes from WEKA, International journal of pure and applied research in engineering and technology, ISSN 2319-507X, (2016), Vol 4, Iss 9.

[10]. S.A.Ghogare and Dr.S.R.Kalmegh, Comparative analysis of J48 and LMT classifier using WEKA data mining tool on car Review Data, Research Journey International E- Research Journal, (2019), ISSN:2348-7143, Special Issue 110 (C).

[11]. S.R.Kalmegh & S.N.Deshmukh, Categorical Identification of Indian News Using J48 and Ridor Algorithm, International Refereed Journal of Engineering and Science (IRJES), (2014), ISSN:2319-183X, Vol3, Iss6.

[12]. Ian H. Witten, Eibe Frank & Mark A. Hall, Data Mining Practical Machine Learning Tools and Techniques,(Third Edition, Morgan Kaufmann Publishers is an imprint of Elsevier, 2016).

S.A.Ghogare" Performance Comparison of RandomForest and Hoeffding Tree classifier using WEKA data mining tool on Car reviews data" International Journal of Engineering Research and Applications (IJERA), Vol. 09, No.03, 2019, pp. 37-42