INTERNATIONAL CONFERENCE ICGTES-19, ICITBDIT-19
Conference World
ISBN : 978-81-941721-9-2
OUCIP, Osmania University Campus, Hyderabad (India)
05th October 2019 www.conferenceworld.in

Performance Comparison of Tree based classifier using WEKA

Mr. Saurabh Ashok Ghogare

Research Scholar,

Department of Computer Science.

Shri. Jagdishprasad Jhabarmal Tibrewala University,

Vidyanagari, Jhunjhunu, Rajasthan (India)

ABSTRACT

The WEKA is open source free data processing tool contain organized collection of state of art machine learning algorithm. The research work carried out in this paper to make a performance evaluation of WEKA classifier. The algorithm contain for the research work is RandomForest and Hoeffding Tree. The paper sets out to make comparative evaluation of four WEKA classifiers in the context of dataset of car reviews to maximize true positive rate and minimize false positive rate. Online review provides valuable resources for potential customers, expert and developer to make decision. The WEKA tool used for result processing. The results in the paper on dataset of car reviews also show that the Performance and accuracy of RandomForest is 100 percent.

Keywords: *Hoeffding Tree, RandomForest, WEKA.*

1. Introduction

The huge amount of data is generated from various resources. In data mining, is a major advancement in the type of analytical tools. Data mining is multi disciplinary filed which is a combination of machine learning, statistics, database technology and artificial intelligence. The data is stored electronically and the search is computerized. It 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. Today, e-commerce platforms offer product reviews. A product review is a textual review of a customer, expert and developers, 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 reviews of car data set. This will be used For Comparative study of RandomForest and Hoeffding Tree classifier[1].

ISBN : 978-81-941721-9-2

2. WEKA

Data Mining is a powerful technology with great ability to, 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, classify it, and summarize the identified relationships. In the given research work the WEKA tool is used. WEKA stand for Waikato Environment for Knowledge Analysis and it was developed at the University of Waikato in New Zealand. 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[2][3]. There are various advantages of WEKA:

- Under the GNU General Public License it is freely available
- It is implemented in the java programming language and thus runs on almost any architecture.
- It contains huge collection of data preprocessing and modeling technique.
- The easiest way to use Weka is through a graphical user interface called Explorer as shown in Fig. 1.

Fig 1: WEKA GUI explorer

WEKA supports various standard data mining tasks like Data preprocessing, clustering, classification, regression, visualization and feature selection. It also support SQL databases using java database connectivity and can process the result returned by database query. This tool also supports the variety file formats for mining include ARFF, CSV, LibSVM, and C4.5.the Fig. 2 and Fig. 3 shows opening of file *.arff by WEKA explore.

ISBN : 978-81-941721-9-2

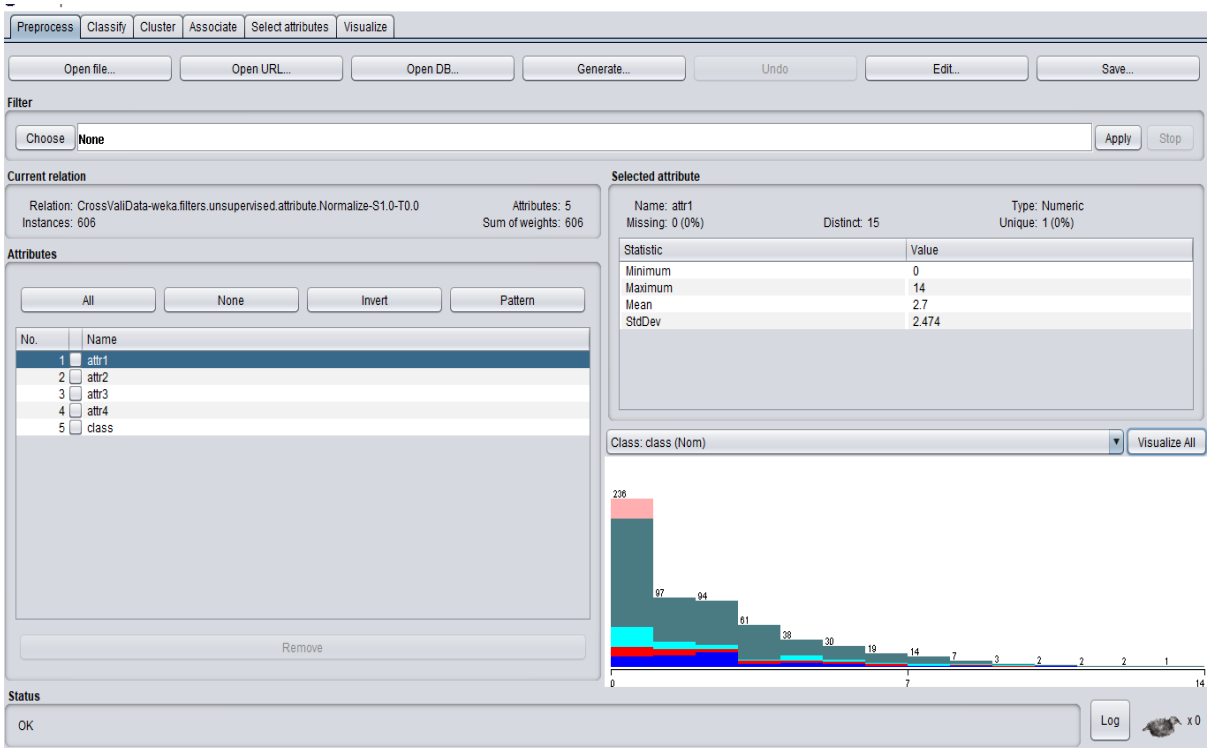

Fig 2: opening of file .arff by weka explorer

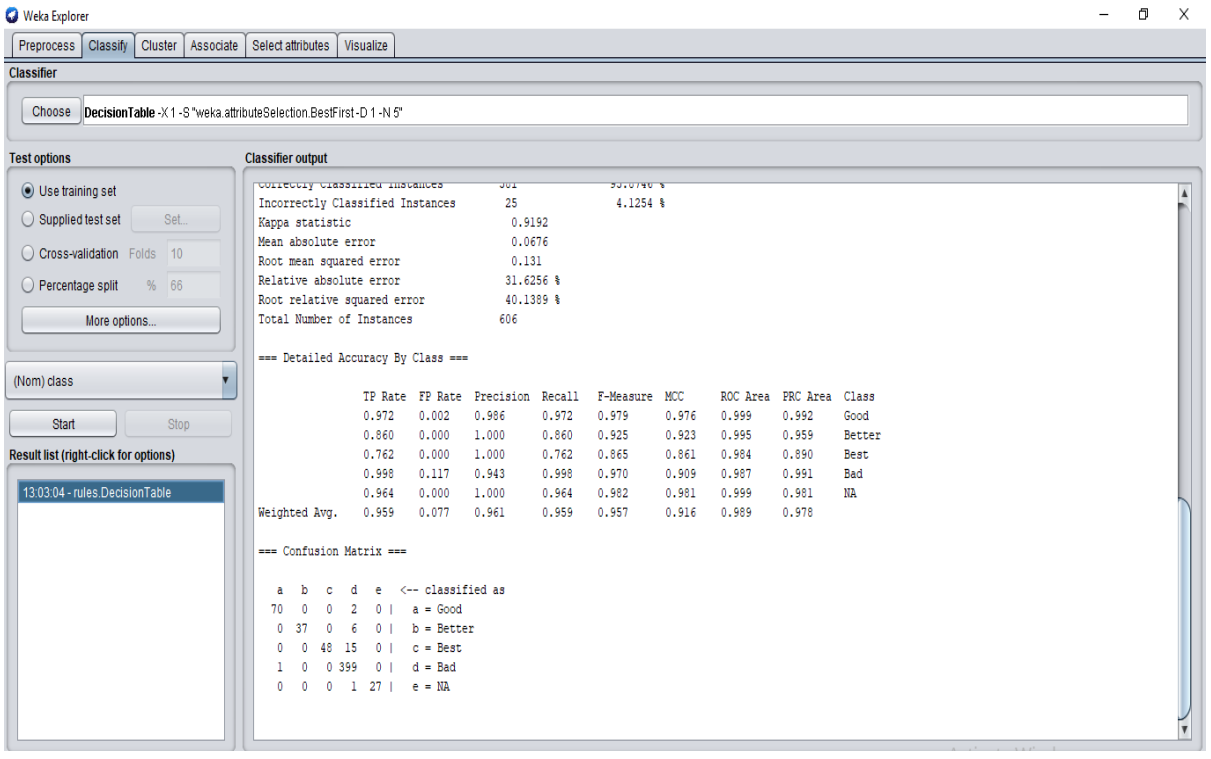

Fig 3: processing of file .ARFF by WEKA explorer

2.1 DecisionTable

Decision Table is an accurate method for numeric prediction from decision trees and it is an ordered set of If-Then rules that have the potential to be more compact and therefore more understandable than the decision trees. Selection to explore decision tables because it is a simpler, less compute intensive algorithm than the decision-tree-based approach. The algorithm, decision table, is found in the Weka classifiers under Rules[4].

2.2 RandomForest

RandomForest is 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[5].

2.3 Hoeffding Tree

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.

3. 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 as 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[6].

3.1 Data Processing

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 DecisionTable and JRip classifier were used on the data set of 606 car reviews. For processing WEKA APIs were used.

4. Result and Conclusion

Table I: Performance Analysis Table

Sr. No	Classifier Type	Classifier Name	Performance Result
1	Tree Based Classifier	RandomForest	100%
2		Hoeffding Tree	91.2541%

As per the previous discussion identification 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. Overall Performance of RandomForest algorithm is acceptable, i.e. Correctly Classified Instances are 606 out of 606 the average percentage of this is 100%, whereas another tree i.e. 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%.

REFERENCES

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

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

[3] 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.*

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

[5] Sushilkumar R. Kalmegh,(2015),"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, *Vol 5, Iss1.*

[6] Sushilkumar R. Kalmegh,(2016), ""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,*Vol 4,Iss 9.*

[7] S.A.Ghogare and Dr.S.R.Kalmegh, (2019), "Comparative analysis of J48 and LMT classifier using WEKA data mining tool on Car reviews data", ISSN 2348-743, *Special issue 110©, Iss 3,pp: 99-105.*