iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 **Issue:** VIII **Month of publication:** August 2019

DOI: http://doi.org/10.22214/ijraset.2019.8075

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue VIII, Aug 2019- Available at www.ijraset.com

A Novel Approach to Enhance the Performance of Web Log Classification

Sonam Singh Gurjar[1], Khushboo Agarwal[2]
[1]*M. Tech Scholar, Department of Computer Science & Engineering, MITS, Gwalior, Madhya Pradesh, India,*
[2] *Asst. Prof. Department of Computer Science & Engineering, MITS, Gwalior, Madhya Pradesh, India,*

Abstract: The World Wide Web (WWW) is expanding rapidly relating to both subtlety of websites and traffic load. It is becoming necessary to classify the usage of websites and traffic load on the basis of certain parameters. Web Usage Mining involves the technique of extricating knowledge out of weblog data taken by the web user. The main purpose of this paper is to research and analyze the performance measures of various classifier algorithms on weblog data, with or without using an ensemble approach. Ensemble technique is one of the data mining approaches which uses multiple learning algorithms for getting improved predictive outcomes. Usually, ensemble learning is one of the efficient methodologies that joins the prediction out of several base classifiers. Mostly employed ensemble techniques are Bagging and Boosting. In this paper, we are concentrating on bagging technique. The efficiency of our approach is restrained and assimilated employing web access log data taken from stamnore.co.uk website.

Keywords: Web Usage Mining, ensemble technique, base classifiers, Bagging, Boosting.

I. INTRODUCTION

World Wide Web (WWW) is emerging fastly and it is an imperative way of sharing data on the internet, because of this extraction of hidden information from web data is quite tedious. Web mining is the employment of data mining, which discovers obscure knowledge from weblog database. Web mining is categories in three ways, that is web content mining, web structure mining, and web usage mining. Where web content mining is used for mining graphs, texts, pictures from abundant web pages. Web structure mining help in discovering an association between several web pages having hyperlink connections. Web usage mining includes making interpretation and prediction of user access pattern out of web data streams. All the analysis and extraction of hidden information collected from the client's usage data on the server. By examining server access log data, we obtain several useful facts and knowledge that will assist in reorganizing a website Various web usage mining tasks are helpful in the rising performance of the website, finer governance of communicating groups of users [1]. Main important tasks of web usage mining are data gathering, data pre-processing, pattern discovery, and outlier evaluation. One of such prominently used the pattern discovery method is the classification of weblog data [1].

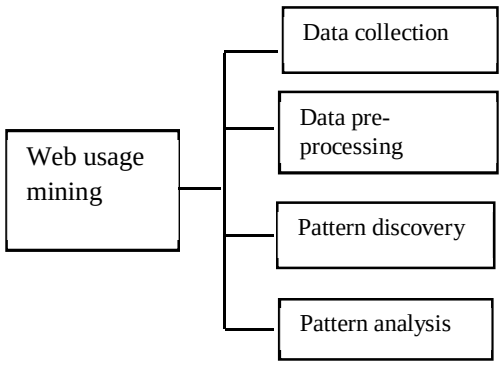


Fig.1 Important tasks of web usage mining.

The elementary resource of data for web usage mining is data stored on the webserver, proxy server, client-server. Data pre-processing includes removal of unwanted, noisy and inconsistent data from weblogs [4] [5]. Pattern discovery is the chief aspect of web usage mining, which discovers exciting configurations and useful facts from web data [3]. Pattern analysis is the ending phase of web usage mining applied to eliminate the unnecessary patterns and extricate useful patterns from the refined knowledge of the pattern discovery stage. OLAP operation, SQL (structured Query Language) is the most common method used for pattern analysis [3]. The pattern discovery phase in web usage mining performs a significant role in mining useful information. One of such

prominently used pattern discovery methods is the classification of weblog data [1]. Classification is a procedure based on the supervised learning approach since learning here is dependent on the allocation of instances to the classes in the training data. Data are mapped into various predefined classes. Many different algorithms are used for training data and classification such as naïve Bayesian classifiers, Decision tree classifier (J48), K nearest neighbor (KNN) and many more. One such most influential method used in supervised learning is merging predictions of multiple classifiers. This approach is widely known as ensemble learning, which is a meta-learning process. Inputs of the meta learner are the output for the base classifiers. The main objective of this learning method is to generate a meta-model which merges predictions of the base classifier within a single prediction by training both the base classifier and the meta classifier[16]. Firstly, we need to train the base classifier, the result of this will be used for training meta classifiers. Various ensemble machine learning techniques are bagging, boosting, random forest, voting, and stacking.

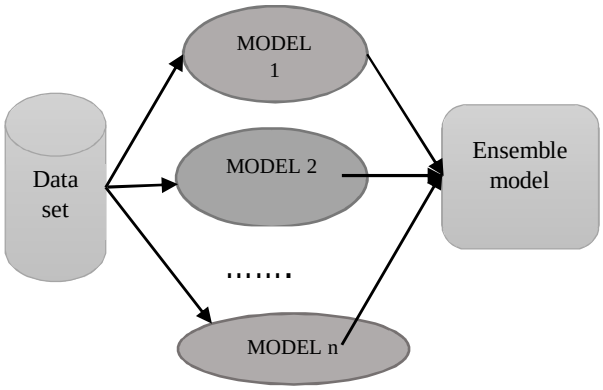


Fig.2 Basic ensemble machine learning process.

II. LITERATURE SURVEY

Mohammed Hamed Ahmed Elhebir et al [1] This paper resembles the reliability of various ensemble models. Performance evaluation was done on several classifiers. All the findings reveal that ensemble learning considerably enhance classification accuracy. Viswanathan K et al. [9] Give Performance Comparability of C4.5 and K NN Classification techniques. The accuracy of classification algorithm validates by using terms like error rate and computation time. This paper considered several classification algorithms with a comparison of performance accuracy. It presents that C4.5 is the best classification algorithm which imparts better results. Bina Kotiyal et al. [ ] Presented classification on weblog data using Naïve Bayesian classifier which helps in discovering user access pattern. The concept of multi-classification is also possible with naïve Bayesian classification. It provides higher efficiency as well as productiveness of the administration by minimizing search time of web user. There is one weakness in this classification algorithm is that it cannot classify undefined classes.

Yash Singhal, Ayushi Jain et al. [10] Examined several machine learning techniques and performed bagging and boosting using a decision tree. They offered an examination of decision tree algorithms.

Fauzia Yasmeen Tani et al. [8] This paper provides an assemblage of decision tree classifiers They utilized ID3 classifier in order to excavate web data. The approach used in this paper can enhance the classification accuracy of log data.

R. Suneetha, R. Krishnamoorthi [7] Improved decision tree C4.5 is used to detect attentive users through weblog data. The results provide advancement in time and memory consumption hence able to present improvement in the accuracy.

Saha et al. [11] Give classification illustration using ensemble machine learning techniques through joining rough set attributes, alleviation and rule production for classifying web pages. A decision table is constructed by taking outputs of separate base learners after this complex set theory was introduced regarding decision table to produce meta classifier.

Zhong and Zou [12] This paper presents an ensemble model by combining support vector machine classifiers. Principle component analysis technique was utilized for the classification of web pages and feature reduction was also exercised.

Choudhary and Raikwal [13] Created a model for web page classification by using naïve Bayes and K-nearest neighbor. On pre-processed log data those algorithms were employed for measuring the uniformity between training and test data.

Thendral Puyalnithi et al. [14] Study the influence of different classification algorithms in the prediction. In this paper, various classification algorithms accustomed to perform the experiment, such as naïve Bayes, random forest, decision tree as a base learner. Comparative analysis was done using bagging and boosting ensemble methods. The results show that naïve Bayes gives better efficiency on the smaller dataset and decision tree is good for bigger data set. Random Forest gives an average performance. At the end of the experiment, it was concluded that the bagging technique with a larger data set gives better results.

Kulkarni & Kelkar [15] Carried out research based on Ensemble Techniques using bagging, boosting, Ada-Boost. The experimental results gave the performance of ensemble classifiers. The research concluded that Bagging classifier is more effective than individual classifiers. It was also showing that Bagging classifier gives better performance accuracy.

III. PROPOSED METHODOLOGY

A. Data Set

In this paper for research work, web access log data is used. Weblog data collected from stanmoreltd.co.uk website having 4605 total access log entries from 20 July 2014 to 17 April 2015. After pre-processing and removing unwanted data total 4529 entries were left, which will be used for mining.

B. Classification Model

The effectiveness of ensemble techniques in web usage mining can be gauged by introducing classification accuracy experiment. This classification accuracy-test compares the performance of ensemble in respect of base classifier. At the initial stage, the performance of the base classifier is tested and then accuracy is calculated. After that, the experiment is carried out on the classifier with bagging. Measuring the usefulness of ensemble technique on the log data we possibly anticipate a greater level of classification precision[13]. If the employed ensemble classifier is not able to give rise of classification accuracy then it will surpass the data processing overhead. Since the analyzation capacity of ensemble method is highly expressive in compare with a single classifier. Ensemble methods have become a trending topic over the past few years. Through combining classifiers, our purpose is to improve the efficiency or performance of the classifier. There are various methods used for combining classifiers[1]. In this paper, we are proposing a Bagging Ensemble Technique, J48 and Naïve Bayesian classification algorithm used as a base classifier for the speciation or categorization of log data set. Mapping and classification of the class label were done on user-based browsing history.

C. Proposed Bagging Technique

Bagging for "bootstrap aggregating", Bagging is a procedure for enhancing performance of machine learning classification algorithms. In data mining Bootstrap aggregating (Bagging) was introduced by Leo Breiman ameliorate the classification by associating classifications of erratically developed training sets[16]. A bagging classifier is an ensemble meta estimator that need samples of data and then classifier is trained on each data sample. The classifier aggregates their personalized predictions to model a final prediction. It works out on several models and averages them to produce a final ensemble model[17].

Bagging procedure is shown in figure 3. Having original data set D. In the first step of Bagging several bootstrap samples are generated with replacement. In the second step, each base model is then trained by using a learning algorithm as a base classifier. In the third step, bagging employed decision by a majority or massed majority voting to integrate outcomes of base mo

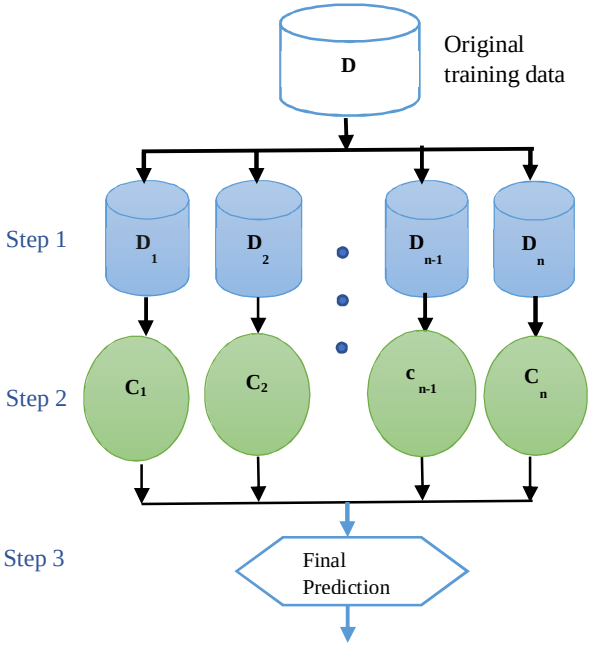


Fig.3 General structure of bagging ensemble model

D. Decision Tree

C4.5 algorithm is associated with the decision tree algorithm set. This algorithm is a progression of the ID3 algorithm which is a very simple decision tree algorithm[2]. C4.5 takes input by means of examining URLs, experimental dataset in terms of sample URLs dataset that would be applied to construct a tree would be exemplified. C4.5 practices tree structure formation of data values, where leaf node points out the class of the instances[11]. From the root node to leaf node sorting is done for the classified instances. C4.5 able to perform error-based pruning and can handle missing values.C4.5 is capable to deal with continuous data. It helps in preventing overfitting. It is an efficacious way of building data representation[7].

In our research, we use an upgrade version of the decision tree C4.5 algorithm (J48). J48 is freely accessible Java execution of the C4.5 algorithm implemented on Weka tool.

E. Naïve Bayesian Algorithm

The Naïve Bayesian classifier algorithm follows the concept of Bayes' Theorem. This classifier algorithm has no parent and every attribute possesses the class as its distinct parent. Naïve Bayesian is a powerful and easy to employed algorithm having lesser succession, therefore works great on a larger dataset. It takes predetermined dataset values presumptions are computed for every class through enumerating association over the values. Highest possibly class is having maximum expectancy[15].

The advantage of Naïve Bayesian algorithm is that it involves just a limited numbers of training data to envisage the parameters, which are beneficial for classification. It identified variance in the variables for every class.

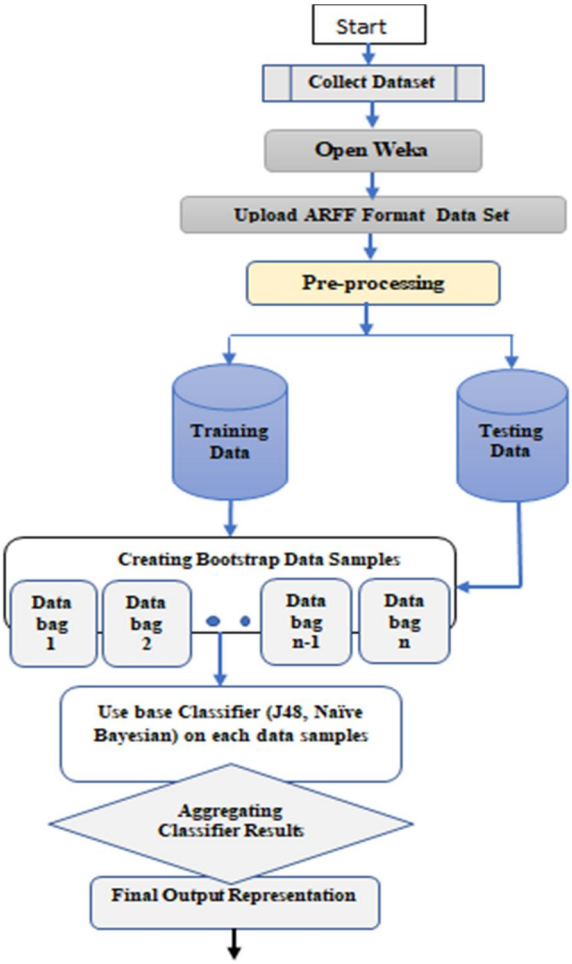


Fig.4 Data Flow diagram for Bagging

The above flow diagram in Figure 4 portrays numerous steps engaged in the working process of Meta classifier. The phases of Bagging algorithm are considered in the flow chart designed, where every single step has been emphasized. The flow chart describes the steps taken for the working of Bagging technique.

IV. EXPERIMENTAL RESULTS

User access log data has been taken from a website for classification. Figure 5 shows the raw log access file. Total 4529 entries were classified after pre-processing in which inappropriate data are uninvolved. The filtered file is converted into CSV format. We split log data into 70% of training and 30% of testing data. Training dataset includes 7 attributes and 3170 instances on the other hand testing data includes 1358 instances

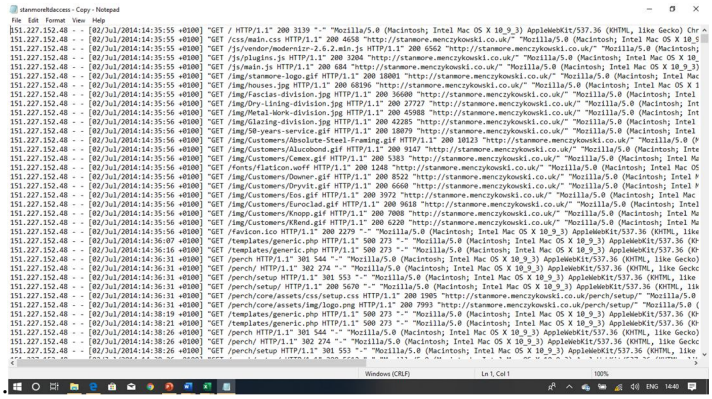


Fig.5 Sample raw log data

We compared and analyse the performance of Decision Tree Classifier (J48), Naive Bayesian Classifier (NB) on different metrics. All the results are represented in the form of tables. Comparative outcomes of different classifiers in the form of accuracy, time and kappa statistic are presented in Table1.

Table 1:The classification performance of each classification algorithm in terms of TP, FP rate, Precision, Recall, F Measure, ROC area, PRC area.

Parameters / algorithm	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC area	PRC area
J48	0.82	0.04	0.95	0.822	0.88	0.95	0.80
J48 with Bagging	**0.88**	**0.02**	**0.97**	**0.88**	**0.92**	**0.98**	**0.93**
Naïve Bayes	0.85	0.03	0.96	0.859	0.90	0.98	0.93
Naïve Bayes with Bagging	0.86	0.03	0.96	0.86	0.90	0.98	0.93

Table 2: Comparison of different classifiers using Accuracy, time and kappa statistic for individual Base and Meta Classifiers. Best results are shown in bold.

Algorithm	Correctly classified instances (% values)	Incorrectly classified instances (% values)	Time is taken to build a model (in seconds)	Kappa statistics
J48	1116 (82.1797%)	242(17.8203%)	0.04	0.7761
J48 with Bagging	1203 (88.5862%)	155(11.4138%)	0.11	0.858
NB	116685 (85.8616%)	192(14.1384%)	0.08	0.8235
NB with Bagging	1169 (86.0825%)	189(13.9175%)	0.46	0.8263

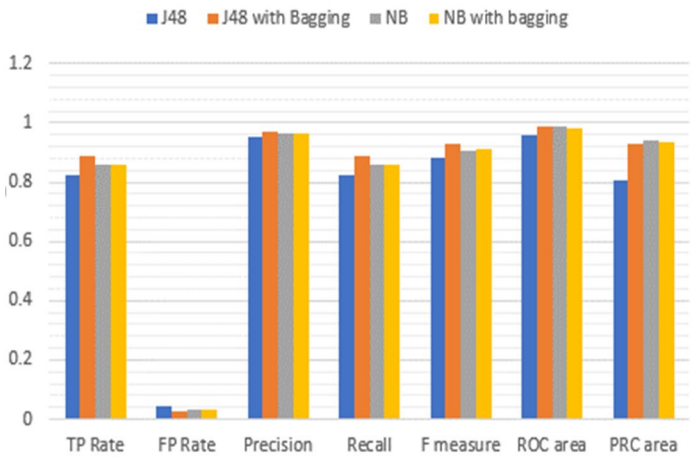


Fig.6 Performance metrics on balance chart.

Table 3 The mean absolute errors (MAE) and root mean squared error (RMSE) for each Classifier algorithms.

Classifier algorithms	MAE	RMSE
J48	0.021	0.1017
J48 with Bagging	0.0172	0.0872
Naïve Bayes	0.156	0.0912
Naïve Bayes with Bagging	0.0169	0.093

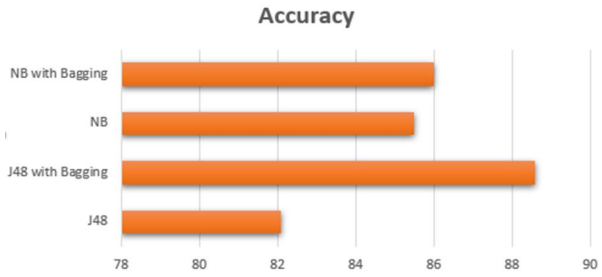


Fig.7 The accuracy derived by using different classification techniques.

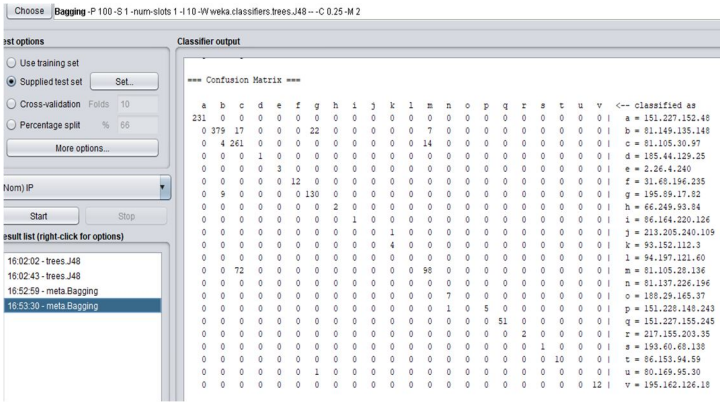


Fig.8 Confusion matrix of Meta Bagging(J48) classifier algorithm.

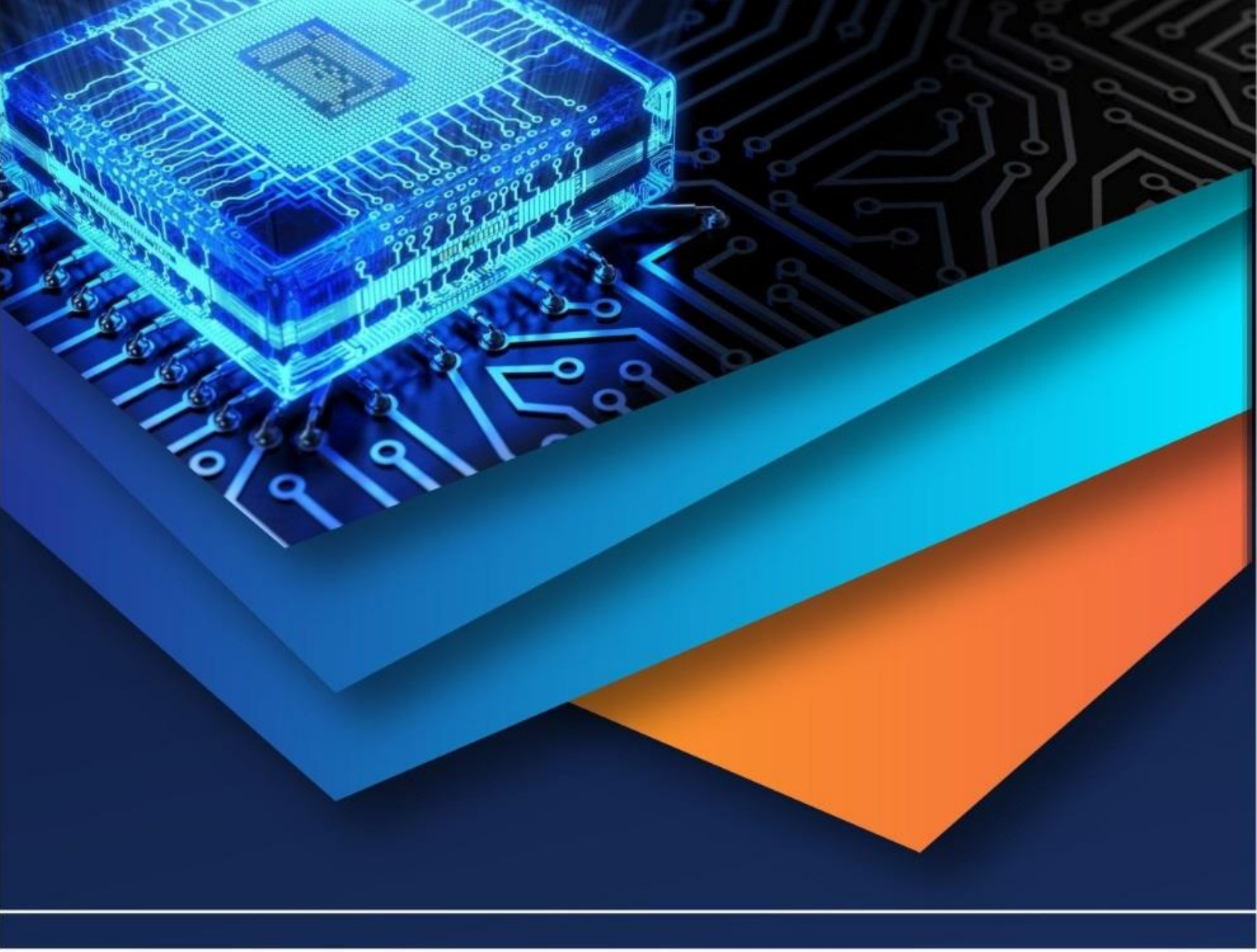
V. CONCLUSION AND FUTURE WORK

In terms of accuracy, precision, recall J48 with Bagging works great on log data. It has the minimum error rate and gives higher performance accuracy compared to well-known classifier algorithm NB. Hence, Bagged J48 classifier algorithm is more appropriate on larger log dataset. The operation has emphasized on a class of the recurrently accessed patterns of attentive users.

It assists website engineers to make the website perform better. The proposed method helps in reducing variance and therefore error. In the future, we will discourse the matter of class inequality with changed data sampling techniques and estimate the outcome of numerous feature ranking methods to improve the accuracy of web mining techniques.

REFERENCES

[1] Mohammed Hamed Ahmed Elhebir, "A Novel Ensemble Approach to Enhance the Performance of Web Server Logs Classification", International Journal of Computer Information Systems and Industrial Management Applications ISSN 2150-7988 Volume 7 (2015) pp. 189-195.
[2] R. Sandrilla e," A Study on Data Pre-processing Methods on Web Log Data in Web Usage Mining", International Journal of Computer Sciences and Engineering, Vol. 6, Issue-7, July 2018 E-ISSN: 2347-2693
[3] ERRITALI, Mohammed, "Pre-treatment of weblog files." Journal of Information Sciences and Computing Technologies 2.1 (2015): 108-121.
[4] Mugali, Chaitra L., "Pre-Processing and Analysis of Web Server Logs." (2014).
[5] Chitraa, V., and A. Selvadoss Thanamani. "A novel technique for sessions identification in web usage mining preprocessing." International Journal of Computer Applications 34.9 (2011): 23-27.
[6] Bina Kotiyal, Ankit Kumar, Bhaskar pant, "Classification Technique for Improving User Access on Web Log Data", Intelligent Computing, Networking, and Informatics, Advances in Intelligent Systems and Computing 243, DOI: 10.1007/978-81-322-1665-0_111, Springer India 2014.
[7] K. R. Suneetha, R. Krishnamoorthi," Classificatio of Weblog Data To Identify Interested Users Using Decision Trees", https://www.researchgate.net/publication/
[8] Fauzia Yasmeen Tani, "Ensemble of Decision Tree Classifiers for Mining Web Data Streams", Communications on Applied Electronics (CAE) Foundation of Computer Science FCS Volume 1– No.1, December 2014.
[9] Viswanathan K, Mayilvahanan K, and R. Christy Pushpaleela, "Performance Comparison of SVM and C4.5, Algorithms for Heart Disease in Diabetic", International Journal of Control Theory and Applications, ISSN: 0974-5572, Volume 10, Number 25, 2017.
[10] Yash Singhal, Ayushi Jain," Review of Bagging and Boosting Classification Performance on Unbalanced Binary Classification", 8th International Advance Computing Conference (IACC), 2018, IEEE.
[11] Saha S, Murthy CA and Pal SK. Rough set-based ensemble classifier for web page classification. Fundamental Informatic 2007; 76(1): 171–187.
[12] Zhong S and Zou D. Web page classification using an ensemble of support vector machine classifiers. Journal of Networks 2011; 76(1): 1625–1630.
[13] Choudhary R and Raikwal J. An ensemble approach to enhance the performance of webpage classification. International Journal of Computer Science and Information Technologies 2014; 5(4): 5614–5619.
[14] Thendral Puyalnithi , "Comparison of Performance of Various Data Classification Algorithms with Ensemble Method Using RAPIDMINER", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 5, May 2016, ISSN: 2277 128X.
[15] S. Kulkarni and V. Kelkar, "Classification of Multispectral Satellite Images Using Ensemble Techniques of Bagging, Boosting and Ada- Boost," pp. 253–258, 2014.
[16] Ankita Singh Tomar, Rajendra Kumar Gupta, Khushboo Agrawal,"A Review Of DM Approaches For Predicting Student's Perfromance", International Jouranal Of Technical Innovation in Modern Engineering & Science, (IJITMES),e ISSN:2455–2585,Volume 4,Issue 10,October 2018.
[17] Ankita singh tomar, Rajendra Agrawal, Khushboo agrawal"A Data mining Approach for identifying the reason behind the alchohol consuming students",JETIR,vol 6 , Issue 1, January 2019.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY