International Journal of Research in Advent Technology, Vol.7, No.4S, April 2019
E-ISSN: 2321-9637
Available online at www.ijrat.org

A Comparative Performance Analysis Of Classification Algorithms Using Web Usage Mining Tool

Varun Malik[1] ,Dr Sanjay Singla[2]
[1]*Phd Scholar*,[2]*Professor*
[1]*Punjab Technical University , Jalandhar*
[2]*GGS College Of Modern Technology,Kharar,Mohali(Punjab)*

Abstract: Web usage mining is a very important task in today's world for discovering interesting patterns from web server log files. It is the application of data mining through which hidden data is discovered from log files. In this paper, a new web usage mining tool WMOT(Web Mining Optimized Tool) has been introduced. WMOT was implemented through the Java NetBeans environment and was used to analyze the performance of various classification algorithms, including Naïve Bayes, SVM, KNN, and Random Forest, to compare the results of classifying data. Out of these, the Random Forest algorithm was found to give the best results.
KEYWORDS: Web Mining; Web Usage Mining; Random Forest, Naïve Bayes.

1. INTRODUCTION

In today's era of technology, the WorldWideWeb plays a crucial role in the gathering of information and retrieval of data over the web. For a single user, it is very difficult to find a particular page or a set of pages over the web in a huge amount of data stored on web servers. To assess this problem and to solve it, users use search engines, which acts as an intermediary between users and the WWW.Websites are the main source of online services, which access and/or distribute global information in many aspects. Today,everyday use of the internet, to access data and information, is increasing rapidly in various organizations, companies,and institutions. Web usage mining is a branch of data mining techniques for using data, called weblog data, for different purposes, such as system enhancement and adaptation personalization and recommendation; advertisement. Web usage mining can extract interesting patterns from the WWW.During this process, web usage mining falls within the framework of Knowledge Discovery from data. In Web usage mining, knowledge discovery in the existing database consists of preparing appropriate data set for carrying out the mining task over the web log files. In web usage mining, data can be processed at the server side, client end or a proxy server. Each set of data obtained from servers and other sources have a different type of dataset[1].

2. WEB USAGE MINING

Web usage mining consists of applying data mining techniques in order to discover usage patterns from web usage data[1]. In web usage mining, the host identifies over the web the location and other general information of web users along with their browsing data. Web usage mining itself can be classified further on the basis of the type of usage data considered. Web Usage Mining consists of applying data mining techniques to find valuable information in the user's log files. The log files contain data related to users' activities including surfing. Web servers track and store the details of activities of these users in web log s. Organizations, in general, collect large volumes of such data and analyze these data for purposes including finding the count of individual customers, to identify the next steps in marketing products, etc. **Web Server Data** contains information including the IP address, page reference, an access time of the website and agent name. **Application Server Data** is data which-commerce applications use as part of commercial applications. Such data includes data related to various business processes and logs related to those. **Application Level Data**forms part of logs recorded at the application level.

Logs are automatically created by web servers. Due to the carrying out of data operations every day, an organization produces significant amounts of data, which are stored in these logs. Web usage mining includes extracting something novel, for interested users, from their network behaviors. When users visit any website, they leave behind some data, such as an IP address, visited pages, visiting time and so on. Web usage mining collects, analyzes and processes the log and records data. As a part of this process, mathematical methods are applied to the model and understand the behavior of users. This understanding is used to provide better services for users by enhancing the structure of websites.

In web usage mining there are mainly three types of processing, including, **preprocessing, pattern discovery and pattern analysis.** In pattern discovery, frequent pattern discovery methods are applied on the web server log data. To complete this task, firstly, data is to be cleaned using the pre-

processing phase, so that the output created as a result of this phase, can be input to various pattern analysis algorithms. Pattern analysis is defined as the understanding of the results obtained by the processing done by these algorithms and inferencing conclusions[14]. These three phases are the main steps in the process of web usage mining. Web usage mining data is preprocessed, and this data is executed using different methods of pattern discovery, where various techniques of classification or clustering are applied to find out desired results. Web usage mining is the term firstly introduced by Cooley in 1997. The resultant pattern and extraction of data from web usage mining can be used in system improvement, website restricting, use of caching, improvement of navigation and in the personalization of web search. The above-stated step of pre-processing is a difficult and complex step in web usage mining.

3. WEB MINING OPTIMIZED TOOL (WMOT)

Using NetBeans, a Java programming platform, algorithms for classification of data were implemented on the same dataset. Here experiments were carried out using theWekatool, where the dataset was uploaded using a utility namedWekaexplorer. In the project name WMOT (Web Mining Optimized Tool),NetBeans was used in combination with the Weka package for performing experiments as shown inFigure No. 1 WMOT tool was used for developing a new classification of data using the Random Forest algorithm combined with the Ant colony optimization and Genetic algorithm. WMOT implemented the Weka classification class for enhancement of the Random Forest algorithm. The main functionality of WMOT consisted of improving the Weka techniques. The main aim was to develop a new tool which enhanced the Weka tool resulting in WMOT (Web Mining Optimized Tool).

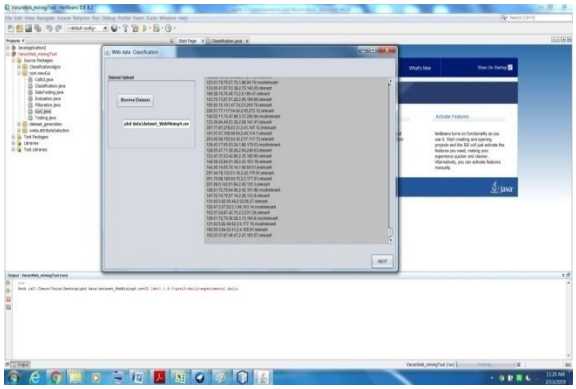

Figure 1 Pre-processing of Dataset

First of all, the dataset was uploaded into the Java IDE environment of the WMOT tool. The dataset was then filtered using*ReplaceMissingValue* filtration. As a result of this filtration, all the missing values were replaced by null values.

After the filtration step was completed in the project the next aim was the classification of the dataset usingtheJavaenvironment. The classification of the dataset using the Naïve

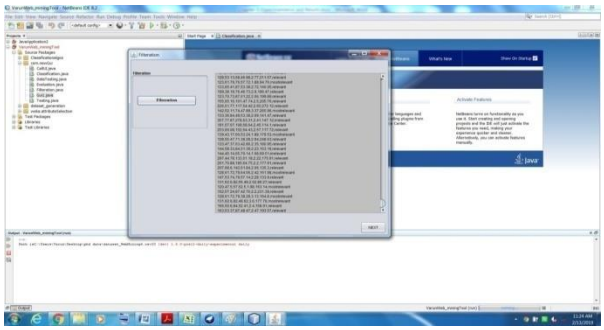

Figure 2 Filtration of Dataset

Bayes algorithm is presented in figure 3. The results of the classification were obtained by applying the NaiveBayes algorithm. The next classification algorithm applied was the SVM (support vector machine) algorithm and the results of the SVM algorithm also performed into the same WMOT web usage mining tool. SVM algorithm is also a classification algorithm used in Weka explorer by adding the additional package into Weka environment and executed the result.

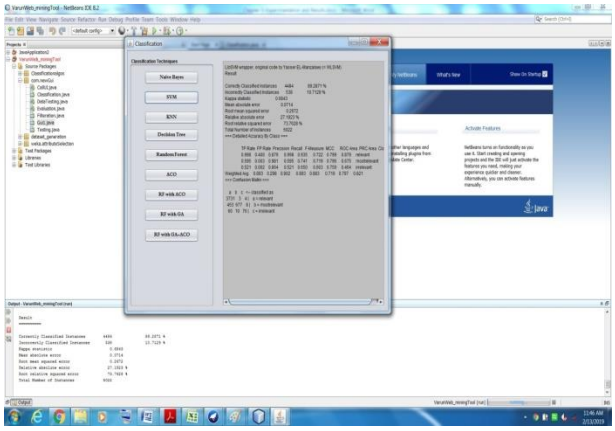

Figure 3 Classification of naive Bayes algorithms

In the SVM algorithm, the correctly classified instances are 4484 and incorrectly classified instances are 538 out of 5022 total instances. In SVM algorithm correctly classified instances are 89.2871% while incorrectly classified instances are 10.7129%. After the SVM algorithm, experiments were carried out using the KNN algorithm, i.e., the nearest neighbor algorithm for classification of the dataset. In the KNNalgorithm, the correctly classified instances

are 4737 and incorrectly classified instances are 285 out of 5022 total instances. In KNN algorithm correctly classified instances are 94.325% while incorrectly classified instances are 5.675%.

4. IMPLEMENTATION OF RANDOM FOREST IN WMOT

In the random forest algorithm, the correctly classified instances are 4915 and incorrectly classified instances are 107 out of 5022 total instances. In a random forest algorithm, correctly classified instances are 97.8694 % while incorrectly classified instances are 2.1306 %.

In the classification of the dataset in Weka explorer, there are many other parameters also through which the performance of the algorithm is calculated, like kappa statistics value, mean absolute error, root mean square error, relative absolute error, root relative absolute error. These are the error rate performance parameters in Weka explorer.

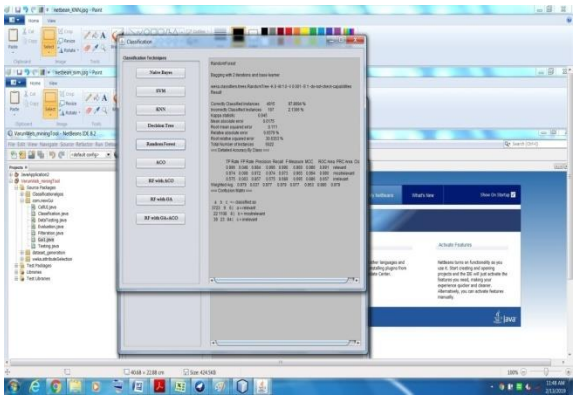

Figure 4 Classification of Random Forest

On the basis of the experimental results stated above and the calculation of performance parameters, the results are summarized in Table 1.

TABLE 1 Classification Using Naïve Bayes, SVM, KNN, Random Forest(RF) in Weka

PERFORMANCE PARAMETERS	NAÏVE BAYES	SVM	KNN	RF
Correctly classified Instances	4092	4484	4737	**4915**
Incorrectly Classified Instances	930	538	285	**107**
Kappa Statistics	0.4868	0.6843	0.8548	**0.945**
Mean absolute error	0.1548	0.0714	0.0383	**0.0175**
Root mean square error	0.2972	0.2672	0.1931	**0.0111**
Related absolute error	58.9383%	27.1923%	14.5953%	**6.6579%**
Root relative square error	82.0194%	73.763%	53.2953%	**30.6353%**
Total no of Instances	5022	5022	5022	**5022**
TP rate	0.815	0.893	0.943	**0.979**
FP rate	0.343	0.298	0.087	**0.037**
Precision	0.804	0.902	0.943	**0.977**
Recall	0.815	0.893	0.943	**0.979**
F-Measure	0.805	0.883	0.943	**0.977**

MCC	0.525	0.719	0.862	**0.953**
ROC Area class	0.904	0.797	0.929	**0.985**
PRC Area	0.889	0.821	0.921	**0.979**
Correctly classified Instances(%)	81.4815%	89.2871%	94.3250%	**97.8694%**
Incorrectly Classified Instances(%)	18.5185%	10.7129%	5.6750%	**2.1306%**

The above table summarizes the results of the various algorithms over the web server log dataset and it can be clearly concluded that the random forest algorithm gave the optimized result of existing algorithms with the maximum number of correctly classified instances i.e. 4915 out of 5022 total instances of the dataset.

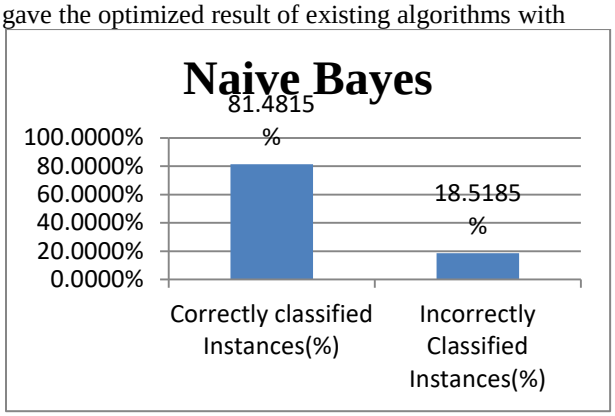

Figure 5 Naïve Bayes Classification

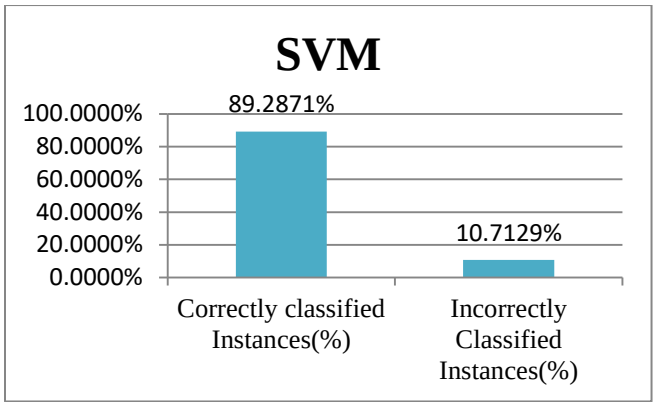

Figure 6 SVM Classification

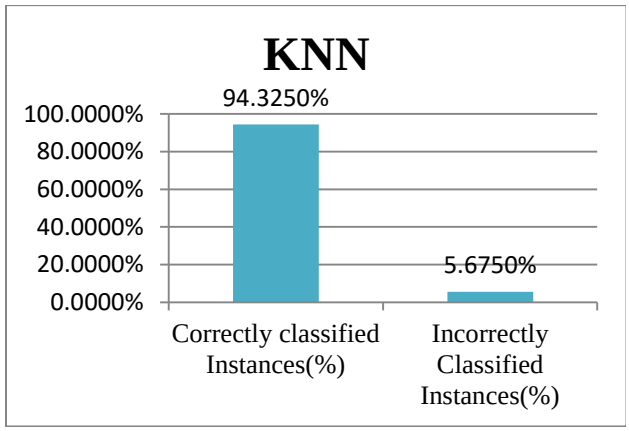

Figure 7 KNN classification

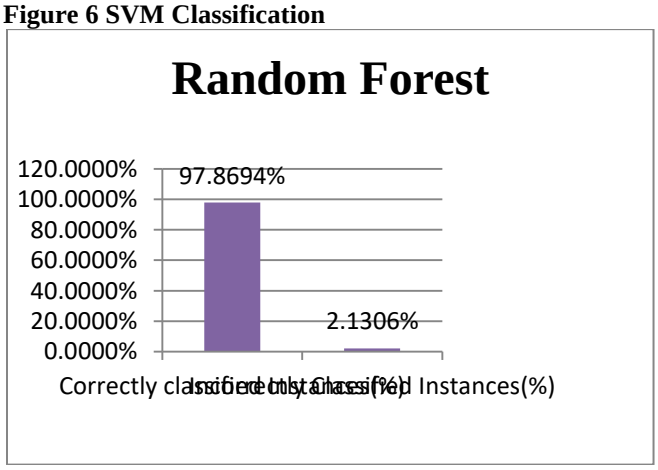

Figure 8Random Forest Classification

Conclusions

In this paper, web usage mining was discussed as being a very important area in today's era of technology, which is useful in organizations, institutions and other areas of the world wide web. To enhance the performance of classification of web usage mining a new tool WMOT was built using the Java NetBeans environment, in combination with the Weka package. Various classification algorithms were applied to data, including naïve Bayes, SVM, KNN, and Random Forest algorithms. The results of experiments led us to observe that the correctly classified instances of Random Forest are higher as compared with other classification algorithms. In our experiments using the WMOT tool, the correctly classified instances in Random Forest are 97.86% which is best when compared with others. Thus, the performance of Random Forest is higher in the classification of data using the WMOT Tool. From the results, it is thus concluded that Random Forest

has maximum correctly classified instances when compared with naïve Bayes, SVM, KNN algorithms and gave maximum correct output.

REFERENCES

[1] Srivastava T., Desikan P., Kumar V. (2000) Web Mining – Concepts, Applications, and Research Directions. In: Chu W., Young Lin T. (eds) Foundations and Advances in Data Mining. Studies in Fuzziness and Soft Computing, vol 180.Springer, Berlin, Heidelberg.

[2] Facca F.M., Lanzi P.L. (2003) Recent Developments in Web Usage Mining Research. In: Kambayashi Y., Mohania M., Wöß W. (eds) Data Warehousing and Knowledge Discovery. DaWaK 2003. Lecture Notes in Computer Science, vol 2737. Springer, Berlin, Heidelberg.

[3] FadouaOuamani, ZeinaJrad, Marie-Aude Aufaure, Hager BaazaouiZghal, and Henda Ben Ghezala (2007) PWUM: A Web Usage Mining Multi-Agent Architecture for Web Personalization, IADIS International Conference WWW/Internet 2007.

[4] M. Castellano, G. Mastronardi, A. Aprile, and G. Tarricone (2007), "A Web Text Mining Flexible Architecture" International Journal of Electrical and Electronics Engineering 1:8 2007

[5] Cooley R., Mobasher, B., & Srivastava, J. (1997) Web mining: Information and pattern discovery on the World Wide Web. Proceedings of the International Conference on Tools with Artificial Intelligence, 558-567.

[6] Ranjitha Kumar, Arvind Satyanarayan, Cesar Torres, Maxine Lim, Salman Ahmad, Scott R. Klemmer, Jerry O. Talton (2013) Webzeitgeist: Design Mining the WebACM Human Factors in Computing Systems (CHI), 2013

[7] L. Yaffi, M. Lesot and N. Labroche (2007) "A New Web Usage Mining and Visualization Tool," 19th IEEE International Conference on Tools with Artificial Intelligence(ICTAI 2007) (ICTAI), Paris, France, 2007, pp. 321-328.

[8] Niu Y et.al.,(2003), WebKIV: Visualizing Structure and Navigation for Web Mining Applications by YonghеNiu, Tong Zheng, Jiyang Chen, Randy Goebel, IEEE / WIC International Conference on Web Intelligence, (WI 2003), 13-17 October 2003, Halifax, Canada

[9] Qingtian Han, XiaoyanGao, Wenguo Wu (2008) "Study on Web Mining Algorithm Based on Usage Mining", 9th International Conference on Computer-Aided Industrial Design and Conceptual Design, 22-25 Nov. 2008

[10] Ramakrishna, M.T Gowdar, L.K., Havanur, M.S., &Swamy, B.P. (2010). Web Mining: Key Accomplishments, Applications, and Future Directions. International Conference on Data Storage and Data Engineering, 187-191,2010

[11]Chowdhury Farhan Ahmed, Syed KhairuzzamanTanbeer, Byeong-Soo Jeong (2010), "Mining High UtilityWeb Access Sequences in DynamicWeb Log Data", Proceedings of the 2010 11th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, Pages 76-81,2010

[12] Singh B. & Singh, H.K. (2010), "Web Data Mining research: A survey" IEEE International Conference on Computational Intelligence and Computing Research, 1-10,2010

[13] DimitriosPierrakos, Georgios Paliouras, Christos Papatheodorou, Constantine D. Spyropoulos,(2011) "KOINOTITES: A Web Usage Mining Tool for Personalization",2011

[14] Ivancsy, Renata and IstvánVajk (2006) "Frequent Pattern Mining in Web Log Data." ,ActaPolytechnicaHungarica, Vol. 3, No. 1, 2006

[15] L.K. Joshila Grace, V.Maheswari, DhinaharanNagamalai (2011), "Analysis of Web Logs and Web User in Web Mining", International Journal of Network Security &Its Applications (IJNSA), Vol.3, No.1, January 2011

[16] D. Gracia-Saiz and M.E. ZorrilaPantaleon(2011), "E-learning Web Miner: A data mining application to help instructors involved in virtual courses"Educational Data Mining, 2011

[17] Dr. D. Suresh Babu, SK Abdul Nabi, Mohd Anwar Ali, Y Raju (2011), "Web Usage Mining: A Research Concept of Web Mining" International Journal of Computer Science and Information Technologies, Vol 2(5),2011

[18] Rakesh Kumar Malviya, Mahesh Chandra Malviya, Vinay Kumar Soni, Ritesh Joshi, PreeteshPurohit(2011), "Survey of Web Usage Mining" International Journal of Computer Science and TechnologyIJCST Vol. 2, Issue 3, September 2011 ISSN:2229-4333(Print)| ISSN: 0976-841(Online)

[19] Saloni Aggarwal and VeenuMangat (2015), "Application Areas of Web Usage Mining" Fifth International Conference on Advanced Computing & Communication Technologies, 2015

[20] Neha Goel, Dr. C. K. Jha (2015), "Preprocessing Weblogs: A Critical phase in Web Usage

Mining" International Conference on Advances in Computer Engineering and Applications (ICACEA),2015

[21] Ravi Khatri, Daya Gupta (2015), "An Efficient Periodic Web Content Recommendation Based on Web Usage Mining" IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS),2015

[22] Dr. Sanjay Kumar Dwivedi, BhupeshRawat (2015), "A Review Paper on Data Preprocessing: A Critical Phase in Web Usage Mining Process" 978-1-4673-7910-6/15/$31.00_c 2015 IEEE

[23] Nandita Agrawal, Prof.AnandJawdekar (2016), "User-Based Approach For Finding Various Results In Web Usage Mining" Symposium on Colossal Data Analysis and Networking (CDAN), 2016

[24] Monika Dhand I, Rajesh Kumar Chakrawarti(2016), "A Comprehensive Study of Web Usage Mining" Symposium on Colossal Data Analysis and Networking (CDAN), 2016

[25] V.Anitha, Dr.P.Isakki(2016), "A Survey on Predicting User Behavior Based on Web Server Log Files in a Web Usage Mining" 978-1-4673-8437-7/16/$31.00 ©2016 IEEE

[26] P. Sukumar, L.Robert, S. Yuvaraj (2016), "Review on Modern Data Preprocessing Techniques in Web Usage Mining" 978-1-5090-1022-6/16/$31.00 ©2016 IEEE

[27] Sunena, Kamaljit Kaur (2016), "Web Usage Mining-Current Trends and Future Challenges" International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT),2016

[28] Suharjito, Diana, Herianto(2016), "Implementation of Classification Technique in Web Usage Mining of Banking Company" International Seminar on Intelligent Technology and Its Application, 2016

[29] C. Ramesh, K.V. Chalapati Rao, A. Govardhan (2017), "Ontology Based Web Usage Mining Model" International Conference on Inventive Communication and Computational Technologies(ICICCT 2017)

[30] Vellingiri, J., S. Kaliraj, S. Satheeshkumar and T. Parthiban (2015), "A Novel Approach for User Navigation Pattern Discovery and Analysis for Web Usage Mining" Journal of Computer Science 2015, 11 (2): 372.382DOI:10.3844/jcssp.2015.372.382

[31] Wang Jicheng, Huang Yuan, Wu Gangshan and Zhang Fuyanat (1999), "Web Mining: Knowledge Discovery on the Web"0-7803-5731-0/99/$10.000 1999 IEEE