Performance Comparison of Tree-Based Machine Learning Classifiers for Web Usage Mining

Ruchi Mittal, Varun Malik, Vikas Rattan, and Deepika Jhamb

Abstract Web usage mining plays a very important role in finding a new patterns and recognition from the web server log file. It is the subcategory of web mining that is used to mine data over the web. In this paper, the authors classified the server log file, identified new patterns of data, and analyzed the data over the tree-based classification algorithms. Authors accessed the tree algorithms (treeJ48, RandomTree, RandomForest, and REPTree) to classify the weblog data and analyzed the results. This paper summarizes the results and compared the tree-based algorithms classification data and check which algorithm gives a better result.

Keywords Web mining · Web usage mining · Machine learning classifiers

1 Introduction

The concept of web mining can be considered as storing the information related to the users' activities on computer and web usage mining is one of the categories of web mining which works on weblog data to extract patterns based on the way the user browse the information on the web [1–5]. The process of web mining consists of three independent stages, namely web content mining, web structure mining, and web usage mining. All these stages help in pointing out the uses' access patterns on the web, and such knowledge can be useful for catering to the varying needs of the organizations. The task of discovering these patterns can be accomplished using various web mining techniques such as classification, clustering, associations, and visualization. The previous research in the area of web mining has potentially identified the patterns in the log files present on the web to understand the users'

R. Mittal · V. Malik (✉) · V. Rattan
Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura, Punjab, India
e-mail: varun.malik@chitkara.edu.in

D. Jhamb
Chitkara Business School, Chitkara University, Rajpura, Punjab, India

S. K. Sabut et al. (eds.), *Proceedings of International Conference on Communication, Circuits, and Systems*, Lecture Notes in Electrical Engineering 728,
https://doi.org/10.1007/978-981-33-4866-0_47

behavior and the way they access the websites [6]. Web mining, similar to other data mining techniques, is a very effective tool used for predictive analysis [7–9].

The key aim of the paper is to perform the classification techniques on weblog data and to understand the performance of the various data mining classifiers in terms of their predictive accuracy. Our approach is mainly based on applying the tree-based classifiers such as J48, random tree, random forest, and REP tree for classifying the dataset [10].

Web usage mining is to analyze the data present on the web pages in the form of log records. These log records contain a variety of important information such as URL accessed by the user, IP address used by the user at the time of web access, and the time at which the information was accessed, and so on. Mining such information to discover hidden and novel patterns in these log files has a lot of potential in developing customers understanding to improve the quality of services offered to the customers [11]. While visiting and browsing a website, users leave a lot of information that can be processed to gain important insights about the users' preferences and behavior on the web to customize the web sites based on users' choices and preferences and to improve their experience on the web [12]. The web mining techniques can assist to collect, pre-process, analyze, and interpret the data, and the models can be built to have better website structures and improved business operations [13, 14]. The authors in this study proposed to compare the performance of various tree-based classification algorithms such as decision tree J48, random tree, random forest, and REP tree in terms of their predictive accuracy using Weka tool.

2 Related Work

According to [15], it is defined that web usage mining plays a crucial role in designing websites and web pages. In this paper, the authors discuss server log files and other relevant information on the web, which led to useful mining of data for future use. Further, authors describe a purposed architecture online MINER which led to the identification of those pages automatically which are not visited a single time, and it also analyzes the traffic volume on the website. In this paper, the authors describe that web usage mining or patterns, and data mining combined will be the basis for future work and future research which will apply data mining algorithms and the KDD process.

Authors described that web mining is a body of knowledge related to web data in the form of such techniques as web content, web structure, and web usage mining. In this paper, the authors identify the main problems in creating an intelligent tool available for mined knowledge and provide in all three categories of web mining. The paper also gives a brief overview of research in web usage mining and their architecture of usage of extracted data and their development efforts in the field of web mining. Authors state that web mining is a fast-growing area of research nowadays. The authors also suggested a starting point for identifying opportunities for future research work on web usage mining [16].

In a different study, the authors implemented a technique in web usage mining for the banking system that helped a company to handle web performance issues. The authors used the pattern discovery phase and proposed a classification technique based on the K–nearest algorithm using Euclidean distance to classify patterns, and the results aided the company to extract useful knowledge from web server logs. The authors concluded that this technique can be used to identify activity data of surfers in web server logs. This allowed the company to create relevant data that they used for analyzing web applications. In the paper, the K-nearest neighbor algorithm showed good results in comparison with Bayesian classification; therefore, it can further be used in web usage mining. Further, researchers can work on another data mining algorithm and can combine different data to get the desired results [17].

According to [18], the authors have defined that the process of finding valuable information and knowledge from the web data is called web usage mining. In this paper, the authors explained about web usage mining that identification of patterns is of high importance in various fields such as business intelligence, e-learning, personalization, etc. This paper described the application areas of web usage mining in-depth and also described the current issues of web usage mining. In this paper, the authors explained two major issues in web usage mining. The first issue is security for users, and the second issue is including semantics in web content. Further, the authors suggested the various areas in web usage mining which will lead to future trends in research. Further, analysis is performed on the results of mining to extract information, to help designers handle users' needs, i.e., individualizing website usage experience and better organization of websites.

Authors presented web usage mining as a data mining mechanism to extract knowledge from log data. Web usage mining is a data mining technique for pattern discovery of new patterns from web server log files for research. The web includes a large collection of data or information for academics and analyzed the data for organizations and institutes for better services in academics to improve the performance of the website by processed or accessed the log files. This paper mainly focused on data collection over web servers for academic education by using Weblog Expert lite 9.3 tools. The authors concluded that webpages are a beneficial advertisement tool for an institution and other sectors like government, etc. The huge benefit can be obtained from web advertisements, website design, and content. This is an area of data mining and pattern recognition which can lead to a good impact in research work [19].

3 Research Methodology

The data mining tool, Weka was selected to process web server log data. The data has been sourced from a leading university in India and has been used in a different study by [10]. The web server log dataset consisted of the weblog files having attributes such as session, new session, bounce rate, pages per session, and class variable having the values based on the relevant, irrelevant, and most relevant instance for a particular

row in the dataset. Initially, the data was pre-processed to handle the missing values. For this, in the pre-process tab, ReplaceMissingValue method is selected to convert the missing and zero value to NULL values for the application of classification technique on the dataset.

The web server log data is run through the J48 algorithm initially for classification. The experiment was run twice by keeping the percentage split ratio as 60:40, 70:30, 80:20, and 90:10 for specifying the training and testing data, respectively. A similar process was followed with other classification algorithms such as random forest, random tree, and REP Tree algorithms. Then, the results were observed and evaluated based on the correctly classified and incorrectly classified instances for all these algorithms.

4 Experiment and Results

This paper mainly focuses on the classification of data, where data is derived from web server log files. These log files may contain a lot of noise or unwanted data that is not relevant to the analysis according to the end-users perspective; therefore, before classification, data is pre-processed using ReplaceMissingValues filter to set the blank values to NULL values in the dataset. After the pre-processing step, various classification methods are implemented on the dataset using Weka, a data mining tool. The dataset consisting of 5022 instances, is run through various tree-based classifiers, namely treeJ48, random forest, random tree, and REP tree. The percentage split (%) method is used as a test category to classify these instances to check the performance of different algorithms on the same dataset using the WEKA tool. By default, the percentage split value is set to 66%.

The first experiment was conducted using a treeJ48 classifier by setting percentage split (%) values to 60%, 70%, 80%, and 90%, respectively, on 5022 instances of weblog files. With 60:40 split ratio on total 5022 instances, classification algorithm J48 resulted in 1968 correctly classified instances and 41 incorrectly classified instances were incorrectly instances. With 70:30 split ratio, the same algorithm correctly classified 1478 instances and 29 instances were incorrectly classified, i.e., 98.0756% correctly classified instances and 1.9244% incorrectly classified instances. In another case, 1004 instances are classified by setting the split value to 80%. Out of 1004 instances, 986 instances are correctly classified and 18 instances are incorrectly classified instances, i.e., 98.2072% correctly classified and 1.7928% incorrectly classified instances. When the percentage split is kept as 90:10, the J48 algorithm correctly classified 493 instances and 9 instances were incorrectly classified out of total 502 instances, giving an accuracy of 98.4064%. This clearly indicates that the J48 algorithm is giving the maximum accuracy on the dataset when the classification criteria in the form of percentage split ratio was kept to be 90:10 as shown in Table 1.

The second experiment was conducted using random forest classifier by setting percentage split (%) value to 60%, 70%, 80%, and 90%, respectively, on 5022

Table 1 Classification of dataset over tree algorithms with different percentage split

Classification algorithms	Percentage split (%)	Total number of instances executed	Correctly classified instances	Incorrectly classified instances	Correctly classified instances (%)	Incorrectly classified instances (%)
treeJ48	60	2009	1968	41	97.9592	2.0408
	70	1507	1478	29	98.0756	1.9244
	80	1004	986	18	98.2072	1.7928
	90	502	494	8	98.4064	1.5936
Random forest	60	2009	1977	32	98.4072	1.5928
	70	1507	1484	23	98.4738	1.5262
	80	1004	991	13	98.7052	1.2948
	90	502	493	9	98.2072	1.7928
Random tree	60	2009	1954	55	97.2623	2.7377
	70	1507	1468	39	97.4121	2.5879
	80	1004	983	21	97.9084	2.0916
	90	502	490	12	97.6096	2.3904
REP tree	60	2009	1965	44	97.8099	2.1901
	70	1507	1471	36	97.6111	2.3889
	80	1004	977	27	97.3108	2.6892
	90	502	492	10	98.0080	1.9920

instances of weblog files. First of all percentage split ratio was kept as to 60%, and thus, total 2009 instances were used out of total 5022 instances, and out of which, 1977 instances were correctly classified, and 77 instances were incorrectly classified. With 70% split, 1507 instances out of total 5022 instances were used for classification, and in the result, 1484 instances were correctly classified, and 23 instances were incorrectly classified, i.e., 98.4738% correctly classified instances and 1.5262% incorrectly classified instances. In another case, 1004 instances are classified by setting the split value to 80%. Out of 1004 instances, 991 instances are correctly classified, and 13 instances are incorrectly classified instances, i.e., classification accuracy of 98.7052%. Finally, the percentage split is kept as 90:10, and total 502 instances were used for classification; out of which, 493 instances were correctly classified, and 9 instances were used incorrectly classified here and thus giving a classification accuracy of 98.2072%. Thus, random forest works best on the dataset giving maximum accuracy of 98.7052% by keeping the percentage split ratio to 80:20 as in Table 1.

The third experiment was conducted using random tree classifier by setting percentage split (%) value to 60%, 70%, 80%, and 90%, respectively, on 5022 instances of weblog files. Firstly, when the percentage split is 60%, total 2009 instances were classified out of a total of 5022 instances, and of these, 1954 instances were correctly classified, and 55 instances were incorrectly classified. With a 70%

split ratio, 1507 instances out of total 5022 instances were used for classification, and in the result, 1468 instances were correctly classified, and 39 instances were incorrectly classified, i.e., 97.4121% correctly classified instances and 2.5879% incorrectly classified instances. In another case, 1004 instances are classified by setting the split value to 80%. Out of 1004 instances, 983 instances are correctly classified, and 21 instances are incorrectly classified instances, i.e., 97.9084% correctly classified and 2.0916% incorrectly classified instances. When the percentage split is 90%, total instances examined are 502; out of which, 490 instances were correctly classified, and 12 instances were incorrectly classified giving the classification accuracy of 97.6096% as shown in Table 1.

The fourth experiment was conducted using REP tree classifier by setting percentage split (%) value to 60%, 70%, 80%, and 90%, respectively, on 5022 instances of weblog files. When the percentage split ratio is set to 60%, total 2009 instances were used for classification, and out of these, 1965 instances were correctly classified, and 65 instances were incorrectly classified. With 70% split ratio, 1507 instances out of total 5022 instances were used for classification, and in the result, 1471 instances were correctly classified, and 36 instances were incorrectly classified, i.e., 97.6111% correctly classified instances and 2.3889% incorrectly classified instances. In another case, 1004 instances are classified by setting the split value to 80%. Out of 1004 instances, 977 instances are correctly classified, and 27 instances are incorrectly classified instances, i.e., 97.3108% correctly classified and 2.6892% incorrectly classified instances. When the percentage split ratio is set to 90:10, total 502 instances were used for classification; out of which, correctly classified instances were 492, and incorrectly classified instances were 10 giving classification accuracy of 98.0080% as shown in Table 1 (Figs. 1, 2, 3, and 4).

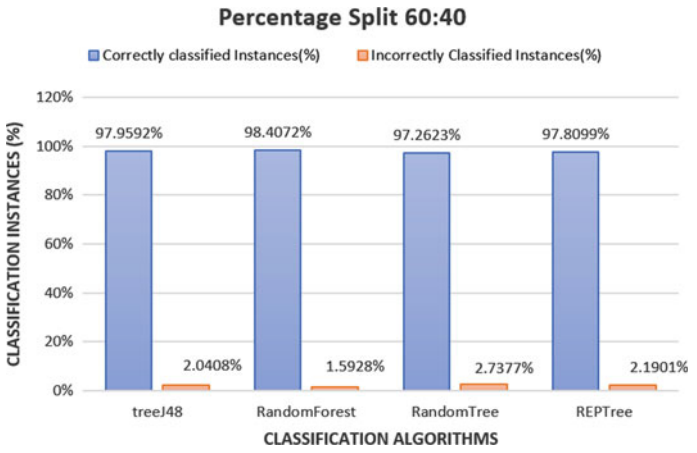

Fig. 1 Classification based on 60:40 percent split

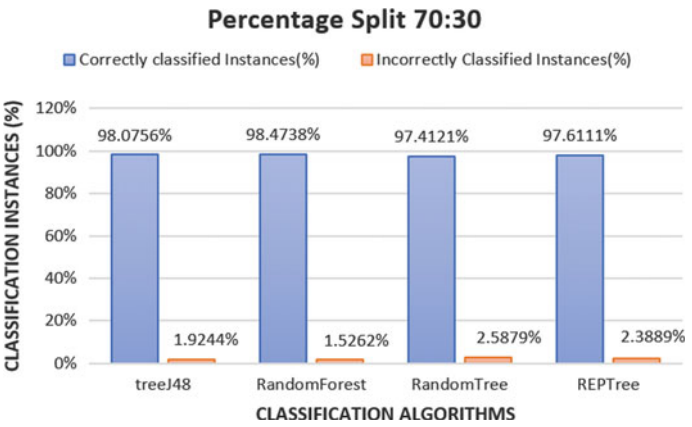

Fig. 2 Classification based on 70:30 percent split

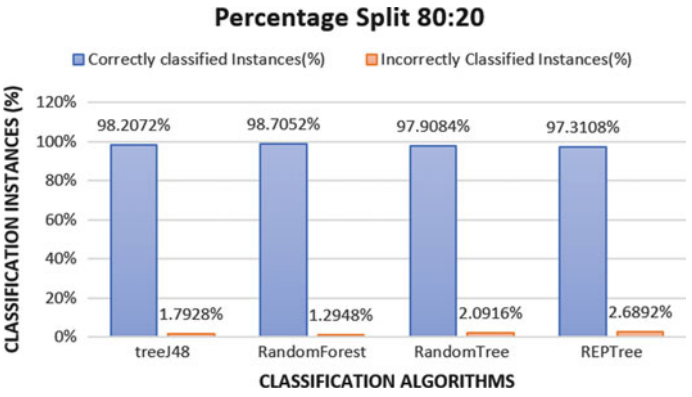

Fig. 3 Classification based on 80:20 percent split

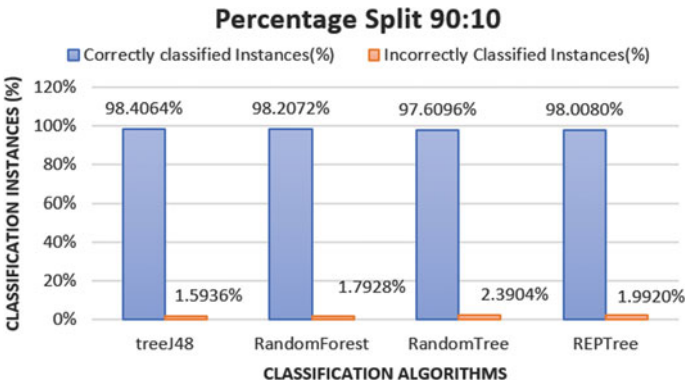

Fig. 4 Classification based on 90:10 percent split

5 Conclusion

In this paper, the authors concluded the results of tree-based algorithms under the classification category of web mining using the data mining tool WEKA. The authors have summarized the result of different tree-based algorithms to check these for maximum efficiency on the web server log dataset. Authors have performed the analysis using various algorithms such as treeJ48, random tree, random forest, and REP tree and observed the results. Tree-based J48 gives maximum accuracy of 98.2072% with 90:10 split ratio, random tree gives maximum accuracy of 97.9084% with 80:20 split ratio, random forest gives maximum accuracy of 98.7052% with 80:20 split ratio, and REP tree algorithm gives maximum accuracy of 98.0080% with 90:10 split ratio. With this experimentation of classifying the data, the authors have concluded that random forest algorithm gives the best result with a classification accuracy of 98.7052% correctly classified instances when the percentage split ratio is kept at 80% while REPTree gives minimum accuracy by correctly classifying 97.3108% instances at the same percentage split ratio, i.e., 80%. In conclusion, the authors discovered that random forest gives better results between tree algorithms section while using the WEKA tool while REP tree gives the lowest classified instances when we compared over different percentage split ratio.

References

1. M.F. Arlitt, C.L. Williamson, Internet web servers: workload characterization and performance implications. IEEE/ACM Trans. Netw. **5**(5), 631–645 (1997)
2. S. Miller, M. Crystal, H. Fox, L. Ramshaw, R. Schwartz, R. Stone, R. Weischedel, BBN: Description of the SIFT system as used for MUC-7, in *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax*, Virginia, April 29-May 1, 1998
3. B. Jarboui, M. Cheikh, P. Siarry, A. Rebai, Combinatorial particle swarm optimization (CPSO) for partitional clustering problem. Appl. Math. Comput. **192**(2), 337–345 (2007)
4. C.H. Lee, Y.H. Fu, Two levels of prediction model for user's browsing behavior, in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, vol. 1 (2008)
5. T. Hussain, S. Asghar, S. Fong, A hierarchical cluster based preprocessing methodology for Web Usage Mining, in *2010 6th International Conference on Advanced Information Management and Service (IMS)* (IEEE, 2010), pp. 472–477
6. R. Khanchana, M. Punithavalli, Web usage mining for predicting users' browsing behaviors by using FPCM clustering. Int. J. Eng. Technol. **3**(5), 491 (2011)
7. R. Mittal, Multivariate regression predictive modelling in analysing student performance: a data mining approach. J. Comput. Theor. Nanosci. **16**(10), 4362–4366 (2019)
8. R. Mittal, Prediction of heart diseases using Hayes process macro serial mediation model 6. Indian J. Public Health Res. Dev. **10**(10), 538–543 (2019)
9. R. Mittal, Identification of salient attributes in social network: a data mining approach, in *International Conference on Recent Developments in Science, Engineering and Technology* (Springer, Singapore, 2019), pp. 173–185
10. V. Shrivastava, N. Gupta, Performance improvement of web usage mining by using learning based k-mean clustering. Int. J. Comput. Sci. Its Appl. ISSN 2250-3765

11. T.P. Hong, M.J. Chiang, S.L. Wang, Mining weighted browsing patterns with linguistic minimum supports, in *IEEE International Conference on Systems, Man and Cybernetics*, vol. 4 (IEEE, 2002), 5 p
12. R. Cooley, B. Mobasher, J. Srivastava, Data preparation for mining World Wide Web browsing patterns. Knowl. Inf. Syst. **1**(1), 5–32 (1999)
13. H. Chen, M. Chau, Web mining: machine learning for web applications. Ann. Rev. Inf. Sci. Technol. **38**(1), 289–329 (2004)
14. G. Stumme, A. Hotho, B. Berendt, Semantic web mining: state of the art and future directions. J. Web Semant. **4**(2), 124–143 (2006)
15. D.S. Babu, S.A. Nabi, M.A. Ali, Y. Raju, Web usage mining: a research concept of webmining. Int. J. Comput. Sci. Inf. Technol. **2**(5) (2011)
16. R.K. Malviya, M.C. Malviya, V.K. Soni, R. Joshi, P. Purohit, Survey of web usage mining. Int. J. Comput. Sci. Technol. (IJCST) **2**(3) (2011). ISSN:2229–4333(Print)| ISSN: 0976–841(Online)
17. S. Suharjito, D. Diana, H. Herianto, Implementation of classification technique in web usage mining of banking company, in *2016 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, pp. 211–218 (2016)
18. S. Jain, R. Rawat, B. Bhandari, A survey paper on techniques and applications of web usage mining, in *2017 International Conference on Emerging Trends in Computing and Communication Technologies (ICETCCT)* (IEEE, 2017), pp. 1–6
19. S.P. Singh, Analysis of web site using web log expert tool based on web data mining, in *2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)* (IEEE, 2017), pp. 1–5