Volume 4, Issue 7, July 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Comparative Study of Classification Algorithms for Web Usage Mining

Supreet Dhillon
Computer Science, SGGSWU,
Fatehgarh Sahib, Punjab, India

Kamaljit Kaur
Computer Science, SGGSWU,
Fatehgarh Sahib, Punjab, India

Abstract— Web usage mining is the process of extracting useful information from server web logs. Information regarding interested web users provides valuable information for web designer to quickly respond to their individual needs. Classification Algorithms can be used for classifying the interested users. In this paper we are analysing the performance of classification algorithms on the bases of some factors like accuracy, precision, session based timing, recall. In section (I) and (III) introduction about web usage mining is provide. In (II) section literature review is provided. In section (IV) classification algorithms are discussed and in section (V) comparison between these algorithms are analyzed. And lastly section (VI) provides conclusion.

Keywords— Web Mining, Web Usage Mining, Web Log, Classification

I. INTRODUCTION

Web mining is the application of data mining techniques to extract knowledge from web data including web documents, hyperlinks between documents, usage logs of websites, etc. Web Usage Mining is a part of Web Mining which in turn, is a part of Data Mining. As data mining is the process of extracting meaningful and valuable information from large volume of data. Web usage mining [1] is the process of mining useful information from server logs. Web usage mining is the process of finding out what users are looking for on internet. This information can then be used in a variety of ways such as, improvement of websites, e-commerce, website personalization, user future request prediction etc. The use of this type of web mining helps to gather the important information from customers visiting the site. This paper focus on the web usage mining and identification of user's behavior on the web. The behavior of users on the web can be analyzed by extracting useful information from web log data. Web log file is automatically created and manipulated by every hit to the website. Log files usually contain noisy and irrelevant data. Preprocessing is done to remove unnecessary data from log file. After then pattern discovery and pattern analysis can be performed for extracting useful patterns. Such interested patterns can be generated using several techniques like classification, clustering, association rule mining. In this paper we deal with classification algorithms for studying the user/client behavior and for the generation of interested user patterns. Consideration of interested web users can be done on the basis of probability of relevant and irrelevant links. Relevant links are the most visited links that can be identified on the basis of time spend on a webpage or number of hits done to a particular link.

II. RELATED WORK

Recently, several web usage mining systems and classification algorithms have been proposed. Here, in the following we review some of the most significant ones. Gupta et al. [17] discusses the current, past and future of web mining. In this they introduce online resources for retrieval Information on the web i.e. web content mining, and the discovery of user access patterns from web servers, i.e. web usage mining that improve the data mining drawback. Mishra et al. [5] describes FP-Growth algorithm for obtaining frequent access patterns from the web log data and providing valuable information about the user's interest. Bhattacharjee et al. [9] proposed a framework of web log mining to implicit measures of user interests through Country and predicting user's future requests in WWW. Dhawan et al. [18] gives an overview of web log file, describes its various types and gives a detail of the process of Web usage mining. Zhong et al. [4] present an n-gram based model to utilize path profiles of users from very large data sets to predict the user's future requests. Their work shows that using simple n-gram models for n greater than two will result in significant gain in prediction accuracy while maintaining reasonable applicability. Vallamkondu et al. [6] describe a new approach to predict user behavior in e-commerce sites and presents the problem of decreasing the latency experienced by the users while traversing e-commerce sites. Kulkarni et al. [8] propose architecture of on line recommendation in web usage mining for enhancing accuracy of classification by interaction between classifications, evaluation, and current user activates and user profile in online phase of this architecture. Tanasa et al. [12] proposed a complete methodology for preprocessing the Web logs and a divisive general methodology with three approaches as well as associated concrete methods for the discovery of sequential patterns with a low support. Nithya et al. [16] continues the line of research on Web access log analysis to analyze the patterns of web site usage and the features of users behavior and novel pre-processing technique is proposed by removing local and global noise and web robots.

III. WEB USAGE MINING

Web Usage Mining involves with the application of data mining methods to discover user access patterns from web data. The main task of web usage data is to capture web browsing behavior of users from a specified web site. Web usage mining can be classified according to kinds of usage data examined. In our context, the usage data is web log data, which maintains the information regarding the user navigation. As our work concentrates on web usage mining, it is the application of data mining techniques to discover usage patterns from web data. Data is usually collected from user's interaction with the web, like web/proxy server logs. Usage mining tools discover and predict user behavior, in order to help the designer to improve the web site, to attract visitors, or to give regular users a personalized and adaptive service.
The major problem with Web Usage Mining is the nature of the data they deal with. With the growth of internet, Web Data has become huge in nature and a lot of transactions are taking place in seconds. Apart from the volume of data, the data is not completely structured. It is in a semi structured format so that it needs a lot of preprocessing before the actual extraction of the required information.

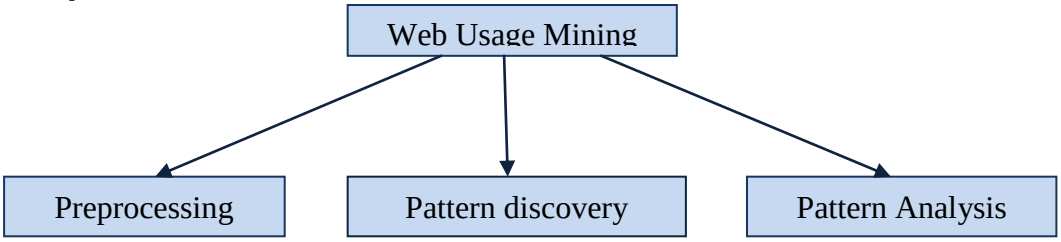

Figure 3.1 Web Usage Mining Process

Preprocessing of web log data [7] has to be deal in such a way that irrelevant data and noise fields are removed completely. Pattern discovery means to extract patterns of web usage from server logs. After that pattern analysis means interesting patterns can be extracted. Classification algorithms can be used for classifying interested users.

IV. CLASSIFICATION ALGORITHMS

The Classification algorithms are discussed under this section. The need and requirement of the user's of the websites to analyze the user preference become essential due to massive internet usage. Classification techniques are to be applied on the web log data and the performance of these algorithms can be measured. Here, in the following several classifiers are being discussed.

A. Decision Tree Classifier

Decision Tree Classifier (DTC) is a simple and widely used classification technique. It is a classifier in the form of a tree structure. In which there is decision node that specifies a test on a single attribute and leaf node that indicates the value of the target attribute. Arc/edge is there for split of one attribute. Path is a disjunction of test to make the final decision. It applies a straight forward idea to solve the classification problem. Decision trees [3] classify instances or examples by starting at the root of the tree and moving through it until a leaf node. It poses a series of carefully crafted questions about the attributes of the test record. Each time it receives an answer, a follow-up question is asked until a conclusion about the class label of the record is reached. If in practice decisions have to be taken online with no recall under incomplete knowledge, a decision tree should be paralleled by a probability model as a best choice model or online selection model algorithm. Decision tree classifier has limitation as it is computationally expensive because at each node, each candidate splitting field must be sorted before its best split can be found.

B. Naive Bayes Classifier

Naive Bayes classifier (NBC) is a simple probabilistic classifier based on applying Bayes' theorem with strong independence assumptions. It can predict class membership probabilities. Naïve Bayes probabilistic classifiers are commonly studied in machine learning. The basic idea in Naive Bayes approaches is to use the joint probabilities of words and categories to estimate the probabilities of categories given a document. The naive part of Naive Bayes methods is the assumption of word independence, i.e. the conditional probability of a word given a category is assumed to be independent from the conditional probabilities of other words given that category. This assumption makes the computation of the NB classifiers [13] far more efficient than the exponential complexity of non-naive bayes approaches because it does not use word combinations as predictors.

C. Support Vector Machine

Support Vector Machines (SVM) are supervised learning models with associated learning algorithms that analyze data and recognize patterns used for classification and regression analysis. A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. SVM [10] constructs a hyperplane or set of hyperplanes in a high-dimensional space, which can be used for classification, regression, or other tasks. A good separation is achieved by the hyperplane that has the largest distance to the nearest training data point of any class, since in general the larger the margin the lower the generalization error of the classifier. A special property is that they simultaneously minimize the empirical classification error and maximize the geometric margin; hence they are also known as maximum margin classifiers.

D. Neural Networks

Neural Networks (NNs) are models for classification and prediction. The idea behind neural networks [11] is to combine the input information in a very flexible way that captures complicated relationships among these variables and between them and the response variable. For instance, recall that in linear regression models the form of the relationship between the response and the predictors is assumed to be linear. In many cases the exact form of the relationship is much more complicated or is generally unknown. In linear regression modeling we might try different transformations of the predictors, interactions between predictors, and so on. In comparison, in neural networks the user is not required to specify the correct form. Instead, the network tries to learn about such relationships from the data. In fact, linear regression and logistic regression can be thought of as special cases of very simple neural networks that have only input and output layers and no hidden layers.

E. Rule Based Classifier

Rule-based classifier (RBC) [14] makes use of set of IF-THEN rules for classification. We can express the rule in the following from: IF condition THEN conclusion. The IF part of the rule is called rule antecedent or precondition. The THEN part of the rule is called rule consequent. In the antecedent part the condition consists of one or more attribute. The consequent part consist class prediction. It is easy to interpret and generate.

F. K-Nearest Neighbor Classifier

K-Nearest Neighbors (K-NN) algorithm is a non-parametric method used for classification and regression. In both cases, the input consists of the K closest training examples in the feature space. K is a user-defined constant, and an unlabeled vector or test point is classified by assigning the label which is most frequent among the k training samples nearest to that query point. In k-NN classification [15], the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

V. COMPARISON BETWEEN CLASSIFICATION ALGORITHMS

We took into account log file with approximately 700 entries from a particular server. Here, in the following we compare results of Decision Tree Classifier (DTC), Naïve Bayesian Classifier (NBC) and Support Vector Machine (SVM), Neural Networks (NNs), Rule Based Classifier (RBC) and K-Nearest Neighbor Classifier (K_NN). Results are displayed in the form of graphs as comparison for recall is present in Figure 5.1 and time, accuracy, precision in Figure 5.2, 5.3 simultaneously.

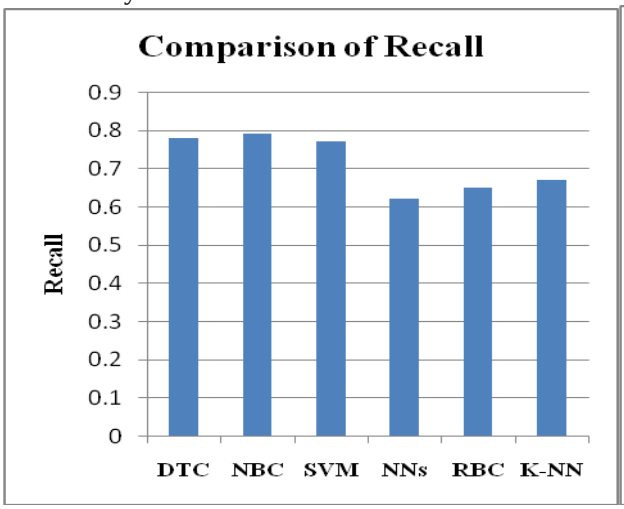

Figure 5.1 Comparative study for Recall

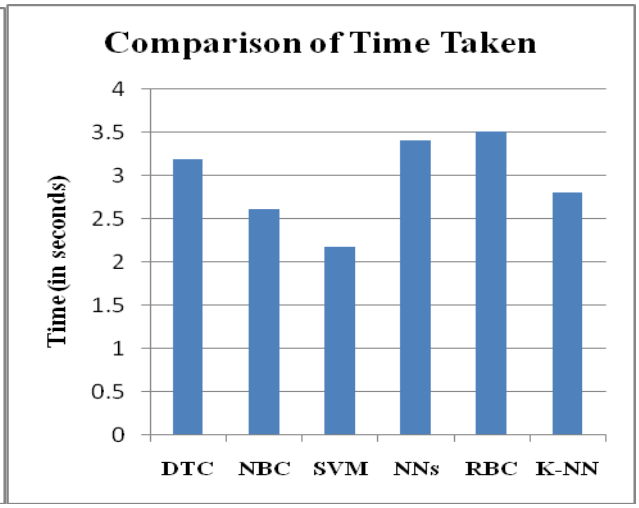

Figure 5.2 Comparative study for time taken

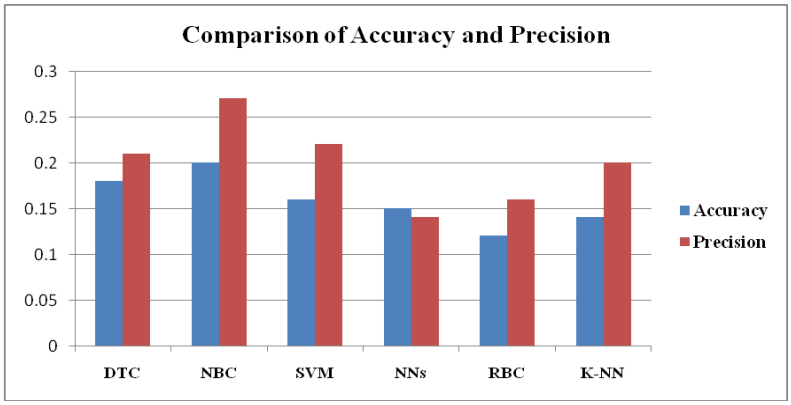

Figure 5.3 Comparative study for accuracy and Precision

VI. CONCLUSION

Web usage mining and Classification algorithms are discussed above. This paper presents a frame for web usage mining based on classification algorithms including their features and limitations. We observe that Naive Bayesian performed well with respect to all the factors. Decision Tree classifier and SVM also perform well as compared to others. We can further work on web usage mining with the combination of these algorithms. Based upon the respective features classification can be performed for web usage mining as a future work.

REFERENCES

[1] Q Zhang, R S Segall, "Web Mining: a survey of current research, techniques and software", Journal of Information Technology and Decision, 2008.

[2] Benjamin Lenz and Bernd Barak, "Data Mining and Support Vector Regression Machine Learning in Semiconductor Manufacturing to Improve Virtual Metrology", International Conference on System Sciences, pp. 3447-3456, 2013.

[3] Yang Hang and S Fong, "An experimental comparison of decision trees in traditional data mining and data stream mining", IEEE 6th International Conference on Advanced Information Management and Service (IMS), pp.442-447, 2010.

[4] Zong Su, Qiang Yand and Hongjiang Zhang, "A Prediction System for Web Requests using N-gram Sequence Models", In Proceedings of the First International Conference on Web Information Systems Engineering, Vol.1, pp.214-221, 2000.

[5] Rahul Mishra and Abha Choubey, "Discovery of Frequent Patterns from Web Log Data by using FP-Growth algorithm for Web Usage Mining", International Jouurnal of Advanced Research in Computer Science and Software Engineering, Issue 9, Vol.2, 2012.

[6] Sudhir Vallamkondu and Le Gruenwald, "Integrating Purchase Patterns and Traversal Patterns to Predict HTTP Requests in E-Commerce Sites", IEEE International Conference on E-Commerece, pp.256-263, 2003.

[7] Vijayashri Losarwar, Dr. Madhuri Joshi, "Data Preprocessing in Web Usage Mining", International Conference on Artificial Intelligence and embedded Systems, July 2012.

[8] Subhash K.Shinde and Dr.U.V.Kulkarni, "A New Approach For On Line Recommender System in Web Usage Mining" International Conference on Advanced Computer Theory and Engineering, pp.973-977, 2008.

[9] Pragya Rajput, Joy Bhattacharjee, Roopali Soni, "A Proposed Framework to Implicit Measures of user Interests through Country and Predicting User's Future Requests in WWW", International Journal of Soft Computing and Engineering, Issue 1, Vol.3, 2013.

[10] Tsyurmasto, Peter, Michael Zabarankin, and Stan Uryasev. "Value-at-risk support vector machine: stability to outliers." Journal of Combinatorial Optimization, pp. 218-232, 2014.

[11] A Ghaffari, H Abdollahib, M R Khoshayanda, I Soltani Bozchalooic, A Dadgara, "Performance comparison of neural network training algorithms in modeling of bimodal drug delivery", International Journal of Pharmaceutics, pp. 126–138, 2006.

[12] Doru Tanasa and Brigitte Trousse, "Advanced Data Preprocessing for Intersites Web Usage Mining", Published by the IEEE Computer Society, pp. 59-65, March 2004.

[13] A. K. Santra and S. Jayasudha, "Classification of Web Log Data to Identify Interested Users Using Naïve Bayesian Classification", International Journal of Computer Science Issues, Vol.9, Issue 1, January 2012.

[14] Mohammad, Rami M., Fadi Thabtah, and Lee McCluskey. "Intelligent rule-based phishing websites classification." Information Security, IET 8.3, pp. 153-160, 2014.

[15] Tomasev, Nenad, Milos Radovanovic, Dunja Mladenic, and Mirjana Ivanovc, "Hubness-based fuzzy measures for high-dimensional k-nearest neighbor classification." International Journal of Machine Learning and Cybernetics 5, Vol.3, 2014.

[16] P.Nithya and P.Sumathi, "Novel Pre-Processing Technique for Web Log Mining by Removing Global Noise, Cookies and Web Robots", International Journal of Computer Applications (0975 – 8887), Vol.53, September 2012.

[17] Aishwarya Rastogi, Smita Gupta, Srishti Agarwal, Nimisha Agarwal, "Web Mining: A Comparative Study", International Journal Of Computational Engineering Research , ISSN: 2250–3005, Issue 2, Vol.2, 2012.

[18] Sanjeev Dhawan and Swati Goel, "Web Usage Mining: Finding Usage Patterns from Web Logs", American International Journal of Research in Science, Technology, Engineering and Mathematics, pp. 203-207, 2013.

