ISSN: 2321-7782 (Online)
Volume 3, Issue 11, November 2015
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Machine Learning Algorithms for Classification of Web Users in E-Commerce Portals

B. Uma Maheswari[1]
Research Scholar in Bharathiyar University,
Coimbatore ,Tamil Nadu, India

Dr. P. Sumathi[2]
Asst. Professor, Govt. Arts College,
Coimbatore, Tamil Nadu, India

Abstract: ***The traffic on World Wide Web is increasing rapidly and huge amount of data is generated due to users' numerous interactions with web sites. Web usage mining attempts to discover useful knowledge from the secondary data obtained from the interactions of the users with the Web. Data mining is one the analytical tools for analyzing information. It allow user to analyze data from many different range, classify it, and summarize the relationships. A user profile is erratic; a method is needed to update the evolving user profiles. 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. This research compares the SVM and Evolving Agent behavior Classification based on Distributions algorithm in web usage mining. The experimental result shows the significant performance of the proposed algorithm. Classification algorithm can improve the accuracy of recommendation and also know the user behavior for improvement of design of a website.***

Keywords: Web usage mining, Evolving Agent behavior Classification based on Distributions algorithm, Support vector machine, User behaviour

I. INTRODUCTION

WUM deals with understanding user behavior in interacting with the web or with a website. One of the aims is to obtain information that may assist web site reorganization or assist site adaptation to better suit the user. Web usage mining model is a kind of mining to server logs and its aim is getting useful users' access information in logs to make sites can perfect themselves with appropriate users' requirements, serve users better and get more economy benefit. 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 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 work focused 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.

This research taken two classification algorithms namely, SVM and EvABCD algorithm and compares the efficiency of the algorithm. In web usage mining, pattern discovery is difficult because only bits of information like IP addresses and site clicks are available. But analysis of this usage data will yield the information needed for organizations to provide an effective presence to their customers. The most effective way to retrieve useful information from a database is application-dependent.

This work is organized as follows: literature survey in Section 2, the Section 3 discusses the statistical classifiers and Section 4 discusses experimental results and finally concludes in Section 5.

II. LITERATURE STUDY

Romero et al (2013) shows how web usage mining can be applied in e-learning systems in order to predict the marks that university students will obtain in the final exam of a course. They have also developed a specific Moodle mining tool oriented for the use of not only experts in data mining but also of newcomers like instructors and courseware authors. Aye (2011) mainly focus on data preprocessing stage of the first phase of Web usage mining with activities like field extraction and data cleaning algorithms. Field extraction algorithm performs the process of separating fields from the single line of the log file. Data cleaning algorithm eliminates inconsistent or unnecessary items in the analyzed data. Carmona et al (2012) presents the methodology used in an e-commerce website of extra virgin olive oil sale called www.OrOliveSur.com. They described the set of phases carried out including data collection, data preprocessing, extraction and analysis of knowledge. Santra, and Jayasudha (2012) propose to use the Naive Bayesian Classification algorithm for classifying the interested users and also we present a comparison study of using enhanced version of decision tree algorithm C4.5 and Naive Bayesian Classification algorithm for identifying interested users. Sharma et al (2011) presents study about how to extract the useful information on the web and also give the superficial knowledge and comparison about data mining. They describes the current, past and future of web mining.

Senkul and Suleyman (2012) developed a technique and a framework for integrating semantic information into Web navigation pattern generation process, where frequent navigational patterns are composed of ontology instances instead of Web page addresses. Grace et al (2011) gives a detailed discussion about these log files, their formats, their creation, access procedures, their uses, various algorithms used and the additional parameters that can be used in the log files which in turn gives way to an effective mining. Automatic weight estimation scheme is used by Sampath et al (2012). The dynamic web page weight assignment scheme uses the page request count and span time values. The proposed system is improves the weight estimation process with span time, request count and access sequence details. Tiwari et al (2011) introduce a web mining solution to business intelligence to discover hidden patterns and business strategies from their customer and web data. They propose a new framework based on web mining technology. Verbeke et al (2011) provides an extended overview of the literature on the use of data mining in customer churn prediction modeling. It is shown that only limited attention has been paid to the comprehensibility and the intuitiveness of churn prediction models.

The main purpose of Yadav et al (2012) is to study the customer's behavior using the Web mining techniques and its application in e-commerce to mine customer behavior. Varghese et al (2012) proposes a cluster optimization methodology based on fuzzy logic and is used for eliminating the redundancies occur in data after clustering done by web usage mining methods. A novel approach Growing Neural Gas is introduced by Sharma (2012) kind of neural network, in the process of Web Usage Mining to detect user's patterns. The process details the transformations necessaries to modify the data storage in the Web Servers Log files to an input of GNG. Park et al (2012) provides information about trends in recommender systems research by examining the publication years of the articles, and provides practitioners and researchers with insight and future direction on recommender systems. Jovanovic et al (2012) applied classification models for prediction of students' performance, and cluster models for grouping students based on their cognitive styles in e-learning environment. Additionally they propose a Moodle module that allows automatic extraction of data needed for educational data mining analysis and deploys models developed in their study.

III. RESEARCH METHODOLOGY

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.

a) Support Vector Machine

Support Vector Machines (SVM) is supervised learning models with associated learning algorithms that analyze data and recognize patterns used for classification and regression analysis. A SVM is a discriminative classifier formally defined by a separating hyperplane. SVM 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. SVM in its basic form implement two class classifications. It has been used in recent years as an alternative to popular methods such as neural network. The advantage of SVM, is that it takes into account both experimental data and structural behavior for better generalization capability based on the principle of structural risk minimization (SRM). Its formulation approximates SRM principle by maximizing the margin of class separation, the reason for it to be known also as large margin classifier.

The basic SVM formulation is for linearly separable datasets. It can be used for nonlinear datasets by indirectly mapping the nonlinear inputs into to linear feature space where the maximum Margin decision function is approximated. The mapping is done by using a kernel function. Multi class classification can be performed by modifying the 2 class scheme. The objective of recognition is to interpret a sequence of numerals taken from the test set. The architecture of proposed system is given in fig. 3.The SVM (binary classifier) is applied to multi class numeral recognition problem by using one-versus-rest type method. The SVM is trained with the training samples using linear kernel. Classifier performs its function in two phases; Training and Testing. After pre-processing and Feature Extraction process, Training is performed by considering the feature vectors which are stored in the form of matrices. Result of training is used for testing the numerals. Algorithm for Training is given in algorithm.

1) Disadvantages:

» SVM is a binary classifier. To do a multi-class classification, pair-wise classifications can be used (one class against all others, for all classes).

» Computationally expensive, thus runs slow.

» A common disadvantage of non-parametric techniques such as SVMs is the lack of transparency of results. SVMs cannot represent the score of all companies as a simple parametric function of the financial ratios, since its dimension may be very high. It is neither a linear combination of single financial ratios nor has it another simple functional form. The weights of the financial ratios are not constant. Thus the marginal contribution of each financial ratio to the score is variable.

3.2 Evolving Agent behavior Classification based on Distributions of relevant events (EVABCD)

» The EVABCD approach for automatic classifier design of the behaviour profiles of users. Original evolving user behaviour classifier is based on Evolving Fuzzy Systems and it takes into account the fact that the behaviour of any user is not fixed, but is rather changing. In this paper, propose an adaptive approach for creating behavior profiles of the customer. It is based on representing the observed behavior of interesting visitors as an adaptive distribution of

her/his relevant atomic behaviors (events). Once the model has been created, EVABCD presents an evolving method for updating and evolving the customer's interest.

» Thus the goal of EVABCD can divide into two phases:

1. Creating and updating user profiles from the records.
2. Classifying a new sequence of records into the predefined profiles.

» This action involves in itself two sub actions:

1. Creating the user behaviour profiles. This sub action will study the interesting visitor's records sequences typed by different users online and generate the related profiles.
2. Classifier Evolution. The sub action consists of online purchasing and classifier update, also includes the possible behaviour to be a model and stored in EPLIB.
3. Classification of user. The created web page visitor's profiles in the preceding section are related with the one of the prototype from the EPLIN and classify into classes created by the model.

» EVABCD have following structure for evaluating the interesting visitor's profiles.

1. Classify the new Sample: It was describe the sample prototype for different users for maintaining their histories.
2. Calculate Potentials: Every sample data set can be maintaining newly search pages details with new prototypes.
3. Update: If any modifications are present in the web page. We are also maintaining new prototype for storing that information with automatic consistency.
4. Remove: Remove the unnecessary results from old prototype for storing newly coming datasets. Supervised and Unsupervised Learning: In this requirement, assigning the prepare dataset for storing relevant information from relational dataset. In data sets representation is formed the training data for preparing new customer profiles based on their behaviors.

Thus, EVABCD is computationally more simple and efficient as it is recursive and one pass. In fact, since the number of attributes is very large in the proposed environment and it changes frequently, EVABCD is the most suitable alternative. Finally, the EVABCD structure is simple and interpretable to identify the interested visitors.

2) Advantages

EVABCD had following advantage, It is recursive,

» It can be used in an interactive mode;

» It is computationally efficient and fast in updating the profile of the user.

» In addition, its structure is simple as well as interpretable. This personalization technique can also be used to monitor and also to detect abnormalities based on a time-varying behaviour of same users and to detect masquerades.

IV. EXPERIMENTAL RESULTS

In this section describe the efficient results for our time varying queries present in the approach EVABCD. The evaluation is carried out using classification accuracy of the algorithm.

TABLE 1: ATTRIBUTES USED FOR E COMMERCE

Attributes	Description
A1	Look for product offers
A2	Price details
A3	Read sub pages
A4	Product benefits
A5	Visit all pages for few minutes
A6	visit web page regularly

The characteristics of the attributes are discussed in table 1. It contains six attributes to find the interested and non interested buyers.

TABLE 2: RATE FOR INTERESTED AND NON-INTERESTED CUSTOMER

Customer Class	Attributes	Rate
With purchase interest	A1	0,0,1,0,1,1
	A2	1,1,1,1,1,0
	A3	0,2,1,3,0,1
	A4	1,0,1,0,1,1
Without purchase interest	A5	0,1,0,0,1,0
	A6	0,0,0,0,1,0

Table 2 shows the values for with purchasing and without purchasing interest. The interested buyers may have maximum value of above one for all the attributes, but non interest buyers have maximum value as zero. Table 2 taken a six customer from huge amount of customers and show the result.

TABLE 3: CLASSIFICATION RATE (PERCENT) OF DIFFERENT CLASSIFIERS IN THE CUSTOMER READING INTEREST USING DIFFERENT SUBSEQUENCE LENGTHS

Number of records for training	**Classifier Rate (%)**				
	Subsequences length	**With purchase interest**		**Without purchase interest**	
		SVM	**EvABCD**	**SVM**	**EvABCD**
100	2	32.6	20.2	19.3	20.1
	3	35.8	33.5	21.7	28.6
	4	36.3	65.8	33.5	40.9
	5	32.8	68.1	41.1	52.8
	6	37.7	70.4	45.6	65.7
500	2	38.2	73.7	43.8	65.9
	3	35.7	71.3	45.2	66.2
	4	39.8	73.8	45.8	67.4
	5	41.3	75.6	46.5	70.5
	6	43.8	75.9	49.3	71.3
1000	2	44.6	69.2	50.8	73.8
	3	44.9	75.0	51.2	74.6
	4	45.6	79.5	55.3	77.5
	5	46.2	82.7	56.8	79.3
	6	46.8	85.6	57.4	81.6

According to these data, EVABCD perform slightly better than the SVM classifiers in terms of accuracy. The percentages of users correctly classified by EVABCD are higher to the results obtained and lower than the percentages obtained by SVM. In general, the difference between EVABCD and the SVM is considerable for small subsequence lengths, but this difference decreases when this length is longer. These results show that using an appropriate subsequence length, the proposed classifier can compete well with offline approaches. Nevertheless, the proposed environment needs a classifier able to process streaming data in online and in real time. In addition, the learning in EVABCD is performed in single pass and a significantly smaller memory is used. Spending too much time for training is clearly not adequate for this purpose. Then the number of attributes is very large in the proposed environment and it changes frequently, EVABCD is the most suitable alternative. Finally, the EVABCD structure is simple and interpretable.

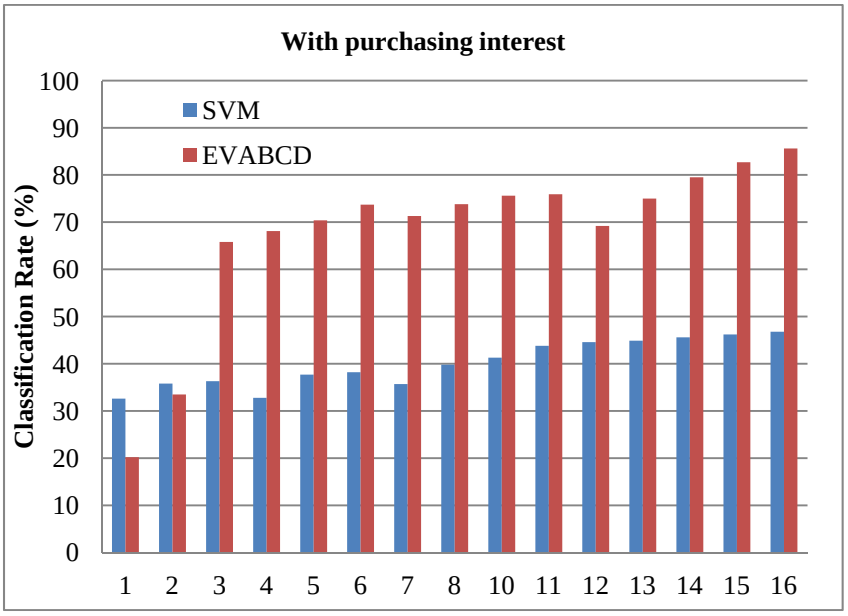

Figure 1: Classification rate for interested customer

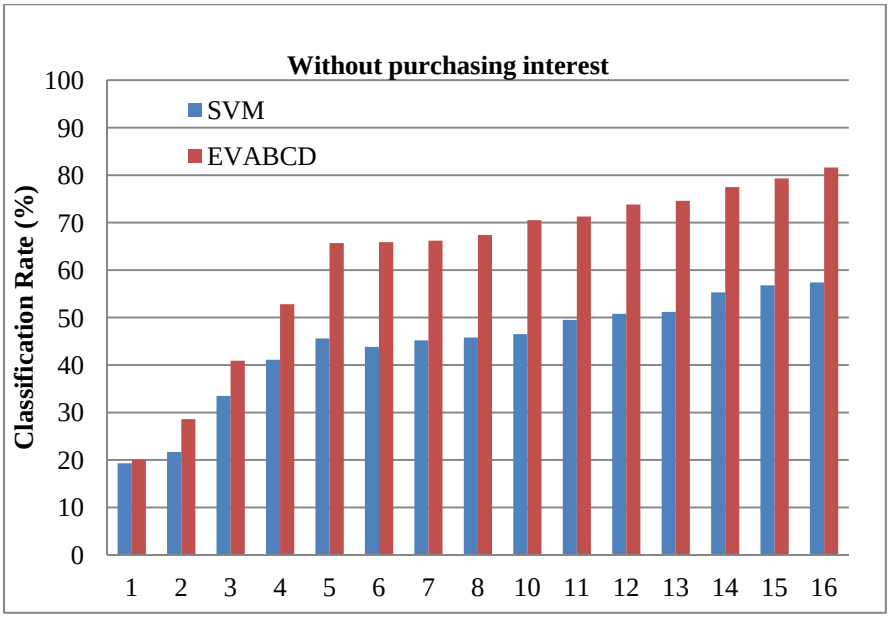

Figure 2: Classification rate for non-interested customer

Figure 1 and Figure 2 show the interested and non-interested customers for purchasing in e-commerce. Thus the proposed method of EVABCD has better performance in classification rate.

V. CONCLUSION

Web usage mining and classification algorithms are discussed above. This work presents a frame for web usage mining based on classification algorithms including their features and limitations. We observe that EvABCD performed well with respect to all the factors and compared with SVM classifier. Based upon the respective features classification can be performed for web usage mining as a future work. Main objective is that classification of user habit in more & more accurate with session base divide data after data cleaning concept for the use of more dynamic web site & web pages in future for business improvement, marketing, government agency put security. In future improve web site or make dynamic web pages so use more large data sets to find more accurate classification

References

1. Romero, Cristobal, Pedro G. Espejo, Amelia Zafra, Jose Raul Romero, and Sebastian Ventura. "Web usage mining for predicting final marks of students that use Moodle courses." Computer Applications in Engineering Education 21, no. 1 (2013): 135-146.
2. Aye, Theint Theint. "Web log cleaning for mining of web usage patterns." InComputer Research and Development (ICCRD), 2011 3rd International Conference on, vol. 2, pp. 490-494. IEEE, 2011.
3. Carmona, Cristóbal J., S. Ramírez-Gallego, F. Torres, E. Bernal, M. Jose del Jesus, and Salvador García. "Web usage mining to improve the design of an e-commerce website: OrOliveSur. com." Expert Systems with Applications 39, no. 12 (2012): 11243-11249.
4. Santra, A. K., and S. Jayasudha. "Classification of web log data to identify interested users using Naïve Bayesian classification." International Journal of Computer Science Issues 9, no. 1 (2012): 381-387.
5. Sharma, Kavita, Gulshan Shrivastava, and Vikas Kumar. "Web mining: Today and tomorrow." In Electronics Computer Technology (ICECT), 2011 3rd International Conference on, vol. 1, pp. 399-403. IEEE, 2011.
6. Senkul, Pinar, and Suleyman Salin. "Improving pattern quality in web usage mining by using semantic information." Knowledge and information systems 30, no. 3 (2012): 527-541.
7. Grace, L. K., V. Maheswari, and Dhinaharan Nagamalai. "Analysis of web logs and web user in web mining." arXiv preprint arXiv:1101.5668 (2011).
8. Sampath, P., C. Ramesh, T. Kalaiyarasi, S. Sumaiya Banu, and G. Arul Selvan. "An efficient weighted rule mining for web logs using systolic tree." InAdvances in Engineering, Science and Management (ICAESM), 2012 International Conference on, pp. 432-436. IEEE, 2012.
9. Tiwari, Sonal, Deepti Razdan, Prashant Richariya, and Shivkumar Tomar. "A web usage mining framework for business intelligence." International Journal of Electronics Communication and Computer Technology (IJECCT) Volume 1 (2011).
10. Verbeke, Wouter, David Martens, Christophe Mues, and Bart Baesens. "Building comprehensible customer churn prediction models with advanced rule induction techniques." Expert Systems with Applications 38, no. 3 (2011): 2354-2364.
11. Yadav, Mahendra Pratap, Mhd Feeroz, and Vinod Kumar Yadav. "Mining the customer behavior using web usage mining in e-commerce." In Computing Communication & Networking Technologies (ICCCNT), 2012 Third International Conference on, pp. 1-5. IEEE, 2012.
12. Varghese, Nayana Mariya, and Jomy John. "Cluster optimization for enhanced web usage mining using fuzzy logic." In Information and Communication Technologies (WICT), 2012 World Congress on, pp. 948-952. IEEE, 2012.
13. Sharma, Anshuman. "Web usage mining using neural network." International Journal of Reviews in Computing 9 (2012).
14. Park, Deuk Hee, Hyea Kyeong Kim, Il Young Choi, and Jae Kyeong Kim. "A literature review and classification of recommender systems research." Expert Systems with Applications 39, no. 11 (2012): 10059-10072.

