Volume 5, Issue 4, 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Comparative Analysis of Model based Techniques

J. Haweliya*
Computer Engineering,
IET, DAVV, India

M. Sharma
Computer Engineering,
IET, DAVV, India

B. Nigam
Information Technology,
IET, DAVV, India

Abstract— ***World Wide Web consists of huge amount of information and this amount is also increases day-by-day. Prediction of web user behavior is becoming a challenge of today competitive edge of World Wide Web. Only prediction of the next web page is not sufficient, evaluation of prediction models is important due to the advantages and disadvantages of every model. If high prediction accuracy is achieved with minimum complexity, then the result of prediction is helpful. In this paper we have evaluated and compare two probabilistic models Markov Model and Bayes Network on the basis of three parameter elapsed time during preprocessing, training time and prediction accuracy. Experiments are conducted on two different real datasets. To model the user navigation on the web and predict the next action taken by the user widely used model is traditional Markov Model. There are some serious limitations on Markov Model. Bayes network is another alternative to predict the next accessed web page but further they have some limitations due to which we can't achieve high prediction accuracy. In this paper we use the Back Propagation neural network algorithm with the combination of either Markov model or Bayes network so that the prediction accuracy can be improved till 95%.***

Keywords— Markov Model; Bayes Network; Back Propagation Neural Network;

I. INTRODUCTION

With the growth of technology at a faster rate, World Wide Web (in short WWW) has also grown exponentially. In day-to- day life it has spread along the whole world and being used in every field. WWW became universal and play a vital role as a tool to collect, share and propagate information along the world. As the popularity of web increased, its complexity also increases due to the existence of bulk amount of data. Web is now turns into an enormous treasury of information and it become very complex for the user to retrieve the information of his/her interest from such a huge nugget. Web mining is an interesting research area that combines the both Data Mining and World Wide Web. Oren Etzioni was the first person who coined the term Web mining in his paper in 1996. Etzioni starts by making a postulate that the information on the Web is adequately structured and also outlines the subtasks of Web mining [1]. In this paper he describes the Web mining processes. We can define web mining as the discovering and analyzing process to retrieve the useful/meaningful information from the www. Although there is so much research have been done on data mining on the web but still researcher have the scope in this area. Mining the data may vary from structured to unstructured. The main focus of Data mining is on structured data which is organized in a database while the text mining focuses/handles unstructured data. Web mining deals with the semi-structured and/or unstructured data. The Web data is actually in unstructured form so it triggers more complexity in the process of Web mining. Getting the information from web has become a very challenging task. The techniques of Web mining behaves like a device to carry out this challenge. These techniques are very helpful in automatic discovery and retrieval of information from the internet [2].

In the Data mining communities it is categorized into three: Web Content mining, Web Structural mining and Web Usage mining [3]. Extraction of useful content from the structured or unstructured web document is explained in Web Content Mining [4]. The addressing problems of Web search and automatic community detection has been solved by two algorithms that work on the Web graph [5]. A complete framework and observations to retrieve the useful patterns from log files of a real Web site has been described in [6].

The structure of this paper is as follows: Section 2 describes the literature survey where brief description of previous research is given. Section 3 explains the architecture of the system that gives the complete idea about how this system will act. Section 4 deals with the experimental results using MATLAB. Section 5 gives the conclusion and future enhancement of this research. Section 6 is about the references used.

II. LITERATURE SURVEY

So many researches have been done in the area of Web mining to predict the next accessed Web page using model based techniques as well as by neural network algorithms. A clustering algorithm SOM (Self organized Map) has been proposed to discover hidden relationships among the Web server data and access patterns. It is an unsupervised learning algorithm. In this algorithm they uses three features as the input for the algorithm number of request, page volume and time index [7]. Emphasize on ART (Adaptive Resonance Theory) Model mechanism of Neural Network. Huge amount of Web logs can easily been classified using ART. It can classify and cluster any type of complex log data [8]. They compare the results of SOM and K-means algorithm for two Website one for music and other for gastronomic. They identify common patterns in Web and conclude that the SOM is better than k-means. But there is a constraint on performance of SOM when the quantity of session is increased k-means will outperform than SOM [9]. Here they uses GNG (Growing Neural Gas) algorithm to identify common patterns in Web. It was introduced in 1991 by Thomas Martinetz and Klous. This algorithm uses only those parameters that are constant in time. The GNG algorithm is better than k-means and SOM. It has a better group of users with k-means and SOM [10]. They give the comparison among all the variants of LQV (Learning vector quantization) algorithm. These variants are OLVQ (Optimized learning vector quantization), MLVQ (Multipass vector quantization), HLQV (Hierarchical learning vector quantization) and LVQ itself. All the mentioned algorithms have been compared on these parameters: Version, Accuracy, Time, Efficiency and Capacity [11]. The performance of the Back propagation Algorithm in predicting the next possible Web page is above 90% [12]. They shown that HMM (Hidden Markov Model) can better exploit information extracted from ink pattern than MM (Markov Model) and NB (Naïve Bayes). It is an optimal inference technique to encode useful information for musical notation representation. HMM uses three observed variables ink direction, spatial information and stroke information [13]. Various types of tools and techniques are described in this paper [14].

III. SYSTEM ARCHITECTURE

The whole process of system is divided into three modules. They are as follows:-

1. Generation of Training and Testing Dataset module: - This is the first module of our system which basically deals with original dataset. Input for this module is original dataset and output will be some training and testing data set generated from this original dataset. The output will be in the form of indexed data.

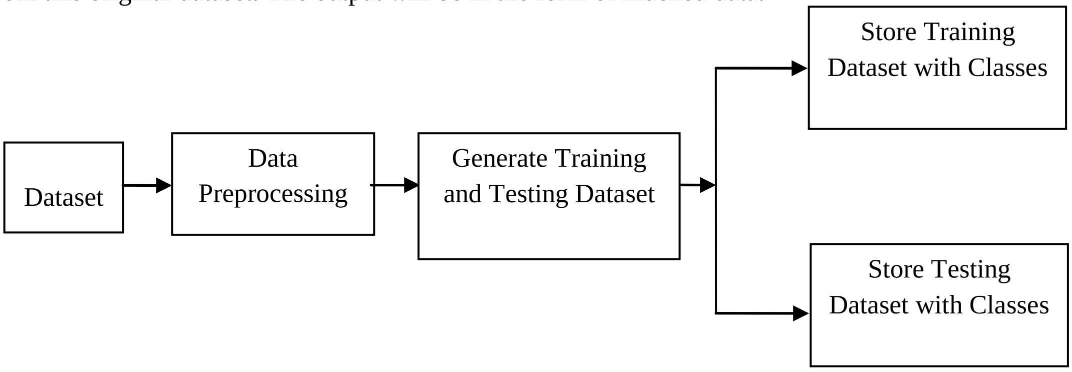
Dataset
Data Preprocessing
Generate Training and Testing Dataset
Store Training Dataset with Classes
Store Testing Dataset with Classes

Figure 3.1: Generation of Training and Testing Dataset

2. Storage steps for Neural Network Classifier module: - This module used to give the different steps to store the neural network classifier. For this module the input is trained dataset with classes which loaded from memory than use either Markov Model or Bayesian Network to extract the feature, after feature extraction train the neural network by using feed forward back propagation algorithm and finally we store the neural network classifier into memory.

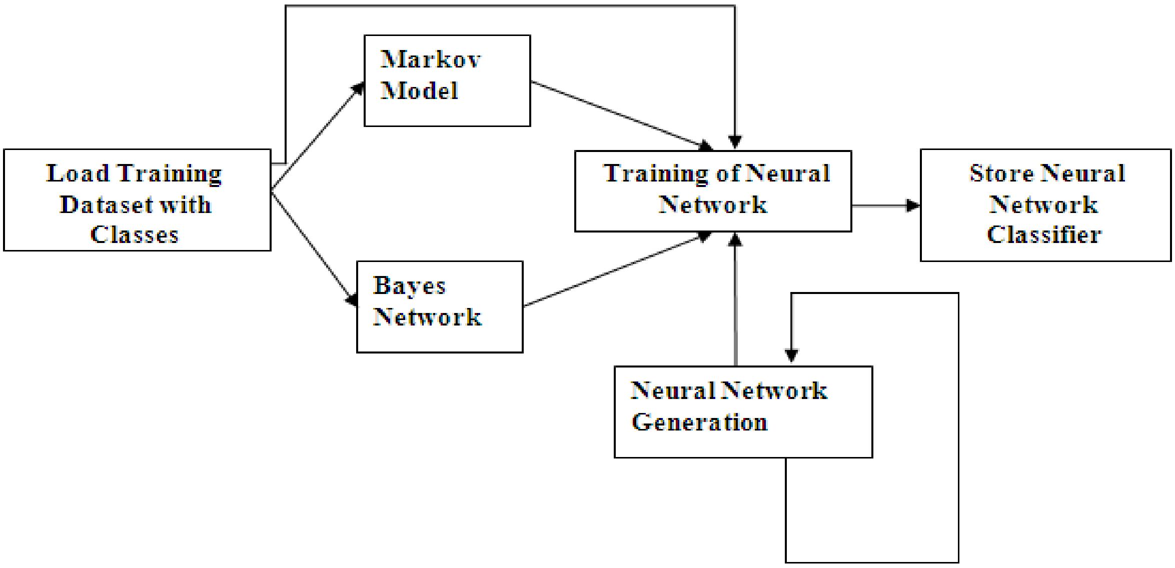
Markov Model
Load Training Dataset with Classes
Training of Neural Network
Store Neural Network Classifier
Bayes Network
Neural Network Generation

Figure 3.2: Storage steps for Neural Network Classifier

3. Calculation of Prediction Accuracy : - In this module we load test dataset and classes as input then use either Markov Model or Bayesian Network to generate feature, after feature extraction generate pre-fetch data index and do classification by loading the neural network classifier. In this module the output is compared with original classes and calculates the prediction accuracy as output.

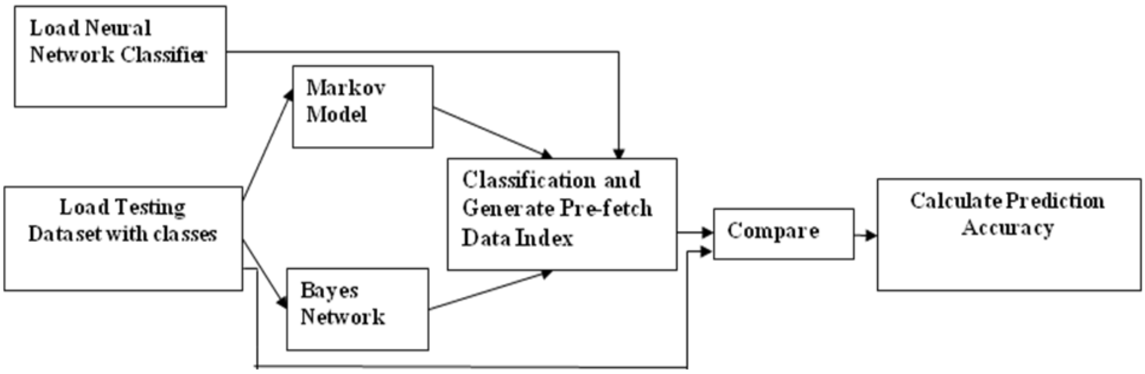
Load Neural Network Classifier
Markov Model
Load Testing Dataset with classes
Classification and Generate Pre-fetch Data Index
Compare
Calculate Prediction Accuracy
Bayes Network

Figure 3.3: Calculation of Prediction Accuracy

IV. EXPERIMENTAL RESULTS

In the analysis of result we have taken three parameters elapsed time in preprocessing, training time and prediction accuracy. Here we take two website access logs to generate the result one is http://studentsolutions.in and other is http://www.ijrndes.org.

4.1 Elapsed Time in Preprocessing

Elapsed time is the time taken by the software to preprocess the data. In other words it's the total time required to generate indexed data. The below shown figure 4.1 and figure 4.2 represent the variation on elapsed time for each dataset as the size of dataset fluctuate.

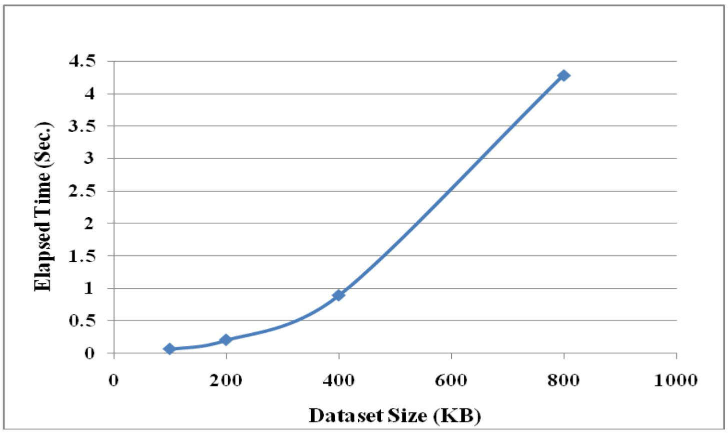

Figure 4.1: Elapsed Time for Student Solution Web Log file

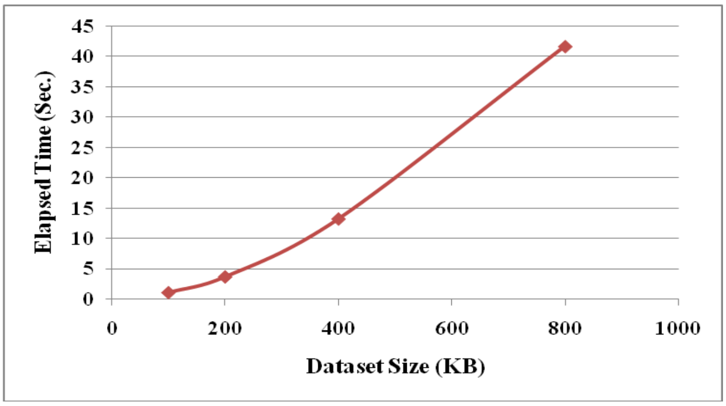

Figure 4.2: Elapsed Time for IJRNDES Web Log File

For both the case we conclude that the elapsed time will increase approximately two to three times as the data size is double (like 100, 200, 400, 800….).

4.2 Training Time

In this software system two probability based model is used one is Markov Model and other one is Bayes Network. Each one takes some time to train the network. So, how the training time for Markov Model and Bayes Network will change as the dataset size increases is evaluated. Two Web Log files are taken for training purpose. Thereafter compare the results of training time. Below graph in figure 4.3 and 4.4 shows that the training time for each Web Log file.

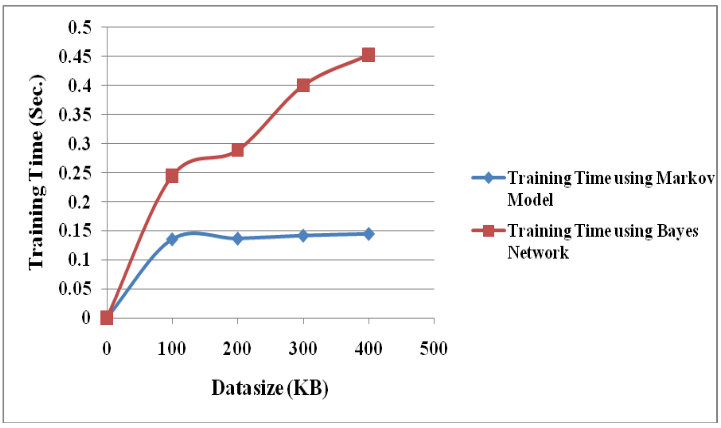

Figure 4.3: Training Time for Student solution Web Log File

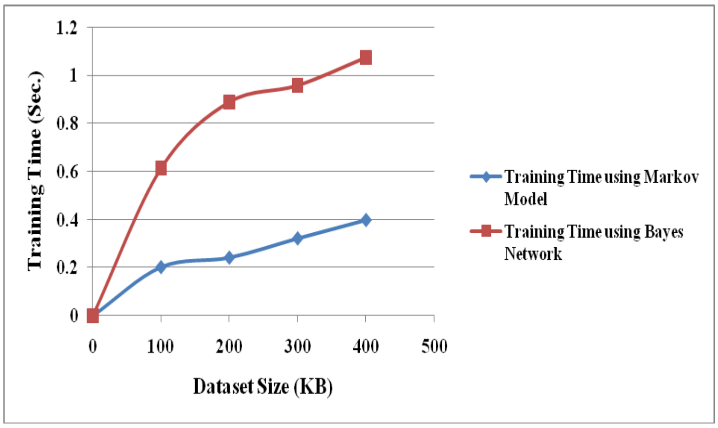

Figure 4.4: Training Time for IJRNDES Web Log File

In both the graph the X-axis represents the Dataset Size in KB and Y-axis represents the training time using each model in seconds. We conclude that for both the model as the dataset size increases the training time will also increases. For both the Web Log files, using Bayes Network the dataset will take more time to train as compared to Markov Model.

4.3 Prediction Accuracy

The amount of correctly identified instances is known as the accuracy of the system. As we are using the Back Propagation Neural Network algorithm in our software system so the accuracy will depends on the mean square error of the algorithm. We adjust the weights of the neural network to minimize the mean square error (MSE) on training set. Until all training examples produce correct output or MSE ceases to decrease. For all training examples, do

Begin Epoch For each training example do

1. Compute the network output
2. Compute the error
3. Back propagate this error from layer to layer and adjust weights to decrease this error End Epoch

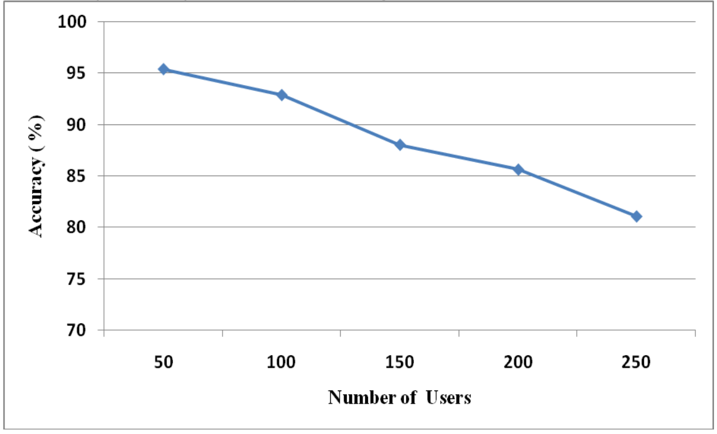

Figure 4.4: Accuracy using Markov Model and Bayes Network for Student Solution Web Log

Similarly in Bayes Network the accuracy of the system will decreases as the amount of Web access data increases due to the noise in the datasets. The below Figure 4.5 represent the same.

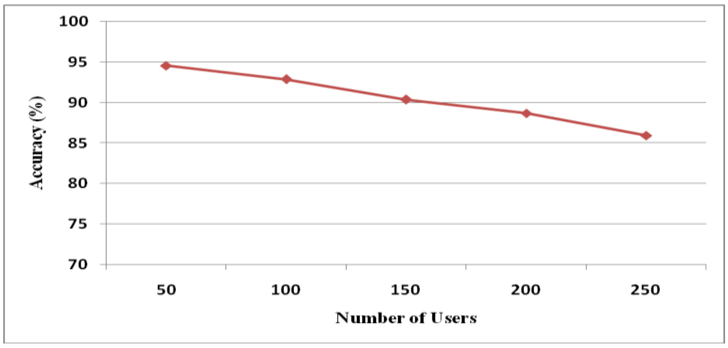

Figure 4.5: Accuracy using Markov Model and Bayes Network for IJRNDES Web Log

V. CONCLUSION AND FUTURE ENHANCEMENT

This paper proposes prediction using either combination of Markov Model with Back Propagation Neural Network or Bayes Network with the Back Propagation Neural Network algorithm. Prediction of the next accessed Web page can also be done by individual model based technique or any one of the neural network algorithm. But combined form give us better coverage and prediction accuracy instead of single one. We simulate the result using MATLAB. This paper also gives the elapsed time during the preprocessing as well as training time for each model based technique with respect to dataset size. We achieve the prediction accuracy of 95% for both the real dataset.

The proposed system mainly improves only accuracy of prediction and investigate the training time, in future we can work on some other parameters like search time, memory usage etc.. Another enhancement is that we can use selective order Markov Model so that state space complexity can be decreases.

REFERENCES

[1] O. Etzioni, “The World-Wide Web: quagmire or gold mine”, Communications of the ACM, 39(11), pp 65-68, 1996.

[2] M. Eirinaki and M. Vazirgiannis.,”Web Mining for web Personalization.”, ACM Trans. Inter. Tech. Vol. 3, No.1, pp 1-27, 2003.

[3] B. Singh, H.K. Singh, “WEB DATA MINING RESEARCH: A SURVEY”, Computational Intelligence and Computing Research (ICCIC), IEEE International Conference, pp. 1 - 10, 2010.

[4] N. Anwat, V. Patil, “Survey Paper on Web Usage Mining for Web Personalization”, International Journal Of Innovative Research & Development, vol. 3, Issue 7, pp 127-132, 2014.

[5] M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, Andrew S. Tomkins, “The Web as a Graph: Measurements, Models, and Methods”, SpringerLink, Vol. 1627, pp 1-17, 1999.

[6] O. Nasraoui, M. Soliman, E. Saka, A. Badia, R. Jermain, ” A Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites “, IEEE, pp. 202 – 215, 2008.

[7] Farhad F. Yusifov, “Web Traffic Mining using Neural Network”, International Scholarly and Scientific Research and innovation, Vol. 2, pp. 862-864, Nov.2008.

[8] J. Jagani and K. Patel, “A survey on Web Usage Mining with Neural Network and Proposed Solution on Several Issues”, Journal of Information, knowledge and Research in Computer Engineering, pp. 330-333, 2013.

[9] P. Britos, D. Martineli, H. Merlind and R. Gracia Martinez, “Web usage mining using self organized Maps”, International Journal of Computer Science and Network Security, pp. 45-50, 2007.

[10] S. Muddalwar, S. Kawar, “Applying Neural Network in Web Usage Mining”, International Journal of Computer Science and Management Research, 2012.

[11] J. Jagani and K. Patel, “An enhanced algorithm for Classification of Web data for Web Usage Mining using Supervised Neural Network Algorithms”, published in International Journal of Computer Engineering & Technology, Volume 90 - Number 17, pp 25-30, 2014.

[12] S. Santhi and Dr. F. Srinivasan “An improved Usage Mining using Back Propagation Algorithm With Functional Updates”, IEEE, pp 1465 – 1468, 2009.

[13] K. Chin Lee, S. Phon-Amnuaisuk and C. Yee Ting, “A Comparison of HMM, Naïve Bayesian, and Markov Model in exploiting knowledge content in Digital ink: A Case Study on Handwritten music notation recognition”, IEEE, pp. 292-297, 2010.

[14] K. Chaudhary, S. Kumar Gupta “Web Usage Mining Tools & Techniques: A Survey”, International Journal of Scientific & Engineering Research, Volume 4, Issue 6, pp. 1762-1768, June-2013.