Supervised Learning Techniques in Web Usage Mining: Comparison and Analysis

Parth Suthar[1]
[1]*Information Technology Department, Parul Institute of Engineering and Technology, Parul University*

Abstract- The World Wide Web contains enormous amount of information about almost every imaginable subject. Web mining is a special area of data mining, which deals with identifying interesting patterns and useful information from the web. This knowledge is useful in improving the quality of services provided by web. The information about user access is stored in the form of web access logs at web servers and proxies. Web usage mining is a discipline that deals with extracting users' information regarding user interests and behaviour profiles by processing those web access logs. This knowledge is useful in improving the areas like web personalization, recommendation systems, business intelligence, market segmentation etc. In this paper, we review, analyse and compare various existing supervised learning techniques utilized for web usage mining. We also present methods to compare and test the performance of those techniques

Index Terms- supervised learning, web usage mining, web access logs, classification, KNN, SVM, Naïve Bayes classifier, decision tree classifier, rule based classifier, k-fold cross validation, bootstrapping, confusion matrix

I. INTRODUCTION

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 behaviour 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 behaviour, 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 many 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 pre-processing before the actual extraction of the required information.

II. STEPS OF WEB USAGE MINING

The steps involved in web usage mining are as shown in figure 1.

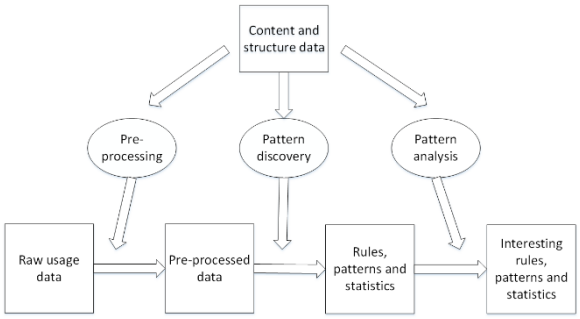

Figure 1 Web Usage Mining Process

A. Pre-processing

Data generated by servers is very bulky and noisy. It contains information about every resource accesses like images, videos, scripts, stylesheets, web pages etc. Identifying relevant data from this and organizing it in terms of users as well as sessions is what pre-processing performs. Steps of pre-processing are as follows. [9][10]

1) Data cleaning
2) User identification
3) Session identification
4) Path completion

B. Pattern Discovery

This step is performed to identify frequent patterns from server generated data. User accesses many resources through clicking the hyperlinks. By identifying the sequence of those click streams pattern

about user's interests can be realized. These patterns if constrained by time threshold, sessions of requests can be identified. Mining those sessions, behaviour & interestingness of users can be identified. These tasks are performed in pattern discovery step. Pattern discovery draws upon methods and algorithms developed from several fields such as statistics, data mining, machine learning and pattern recognition. [8] Various techniques involved in this step are frequent itemset mining, clustering, statistical analysis, classification and sequential analysis.

C. Pattern analysis

The need behind pattern analysis is to filter out uninteresting rules or patterns from the set. The common techniques used for pattern analysis are visualization techniques, OLAP techniques, Data & Knowledge Querying, and Usability Analysis. Visualization techniques are useful to help application domains expert analyse the discovered patterns. [11]

III. SUPERVISED LEARNING

Supervised learning is a data mining (machine learning) technique used to predict group membership for data instances. For example, you may aspiration to use categorization to predict whether the weather on a particular day will be "sunlit", "wet" or "hazy". [7] In the Web domain, one is interested in developing a profile of users belonging to a particular class or category. This requires extraction and selection of features that best describe the properties of given the class or category.

Supervised learning techniques play an important role in Web analytics applications for modelling the users according to various predefined metrics. Given a set of user transactions, the sum of purchases made by each user within a specified period of time can be computed. A classification model can then be built based on this enriched data in order to classify users into those who have a high propensity to buy and those who do not, taking into account features such as users' demographic attributes, as well as their navigational activities. The need and requirement of the users of the websites to analyse the user preference become essential due to massive internet usage. Supervised learning techniques are to be applied on the web log data and the performance of these algorithms can be measured. Following section describes popular Supervised learning techniques available.

IV. LITERATURE REVIEW

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 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. [1]

Complex and extensive web sites are becoming more and more popular. Web related data analysis is the way of providing the statistics about the investments. Commercial web mining packages do not answer all questions which may be interesting to the data analyst. In this paper author proposes several issues and what could help to improve web site's retention. The author also proposes decision trees for web user behaviour analysis. This includes prediction of user future actions and the typical pages leading to browsing termination. Decision tree package C4.5 was used in this study. Decision trees showed reasonable computational performance and accuracy. Experiments showed that it is possible to predict future user actions with reasonable misclassification error as well as to find combinations of sequential pages resulting in browsing termination. In addition to this, decision trees generated human understandable rules which can be used to analyse further for web site improvement. [21]

B. Naïve 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 far more efficient than the exponential complexity of non-naive Bayes approaches because it does not use word combinations as predictors. [13][1]

The need and requirements of the admins of the websites are to analyse the user preference become essential, due to massive internet usage. Retrieving the decisive information about the user preferences is achieved, using Naive Bayesian Classification algorithm with quicker time and lesser memory, by means of constructive Naïve Bayes function. [20]

The constructed model of web data classification uses Naïve Bayes classifier. It works on parameters like time spent and total accessed pages during that time. Ex. If a user spends more than 5 minutes on the web site total accessed pages by him are greater than 10 than the user is classified as interested user. [20]

C. Support Vector Machine

Support Vector Machines (SVM) are supervised learning models with associated learning algorithms that analyse 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 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. SVMs have been successfully applied to a variety of real-world problems like particle identification, face recognition, text categorization, bioinformatics, civil engineering and electrical engineering etc.

An ensemble of SVM classifiers is a collection of SVM classifiers, each trained on a subset of the training set in order to get better results. The prediction of the ensemble of SVMs is computed from the prediction of the individual SVM classifier, that is, during classification, for a new unlabeled input test x_{test} the j -th SVM classifier in the collection returns a probability $P_j(y=1|x_{test})$ of x_{test} belonging to the positive class, where j = 1,2,…,m and m is the number of SVM classifiers in the collection. The ensemble estimated probability, $P_{ens}(y=1|x_{test})$, is obtained by [22]

$$P_{ens}(y = 1|x_{test}) = \frac{1}{m}\sum_{j=1}^{m} P_j(y = 1|x_{test})$$

The approach above is implemented on web pages from yahoo sports. Multiple pages of 6 categories are used as the dataset. 80% of the data is used as the training set and rest as the test set. [22]

D. Neural Networks

The intention of Neural Network (NN) is to mimic the human ability to acclimatize to varying circumstances and the current environment. Neural Networks are models for classification and prediction. The idea behind neural networks 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 modelling 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. [14][1]

Proposed work in paper [19] pre-process, discovers and analyses the Web Log Data of Dr. T.M.A.PAI polytechnic website. A neuro-fuzzy based hybrid

model is employed for Knowledge Discovery from web logs. First, a sample set of pre-processed web log data is clustered using Fuzzy C Means Clustering algorithm. Then, the input of clustering algorithm is given as input to multilayer feed forward neural network with back propagation learning capability and the output of clustering algorithm is given as target output and the training of neural network is done. Mean Square Error – the average squared error between the network outputs and the target outputs is used as the performance measure. Experimental results show that the neural network attained the best validation performance = 0.0080918 at epoch 18. The neuro – fuzzy model combines the neural networks and the fuzzy set theory. Clustering is a subjective process, which means that the same set of data items repeatedly need to be partitioned differently for various applications. It makes clustering difficult as a single algorithm or approach will be inadequate to solve all the clustering problems. This problem is taken care of by the neuro – fuzzy system as it is a self-learning system and generates patterns and rules automatically. The neuro – fuzzy clustering grouped the users having similar browsing patterns into clusters.

E. Rule based Classifier

Rule-based classifier (RBC) 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.[1]

F. K – Nearest Neighbor Classifier

K-Nearest Neighbours (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 unlabelled 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, the output is a class membership. An object is classified by a majority vote of its Neighbours, with the object being assigned to the class most common among its k nearest Neighbours (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest Neighbour. The distance matrices utilized with K-NN are Euclidean distance or cosine similarity. Usually the value of K parameter is taken as odd so as to avoid any ties while classification. [3][1]

Proposed work in [3] uses RSS reader website's data to generate recommendations for unknown users. It first extracts the data, cleans it and groups it into appropriate sessions and generates the data mart. Then for each new user request, the K- Nearest Neighbour classifier with k = 1, 3 or 5 predicts the class label for it and based on the predicted class label, the recommendations are generated. Euclidean distance measure is used to find the distance between test user and training user set. The classifier is easy, simple to understand and it takes into account the real time recommendations rather than showing recommendations based on user's history.

V. COMPARISION AND ANALYSIS

The binary tree decision tree classifier suffers from data fragmentation as the data is partitioned frequently. Naïve Bayes has fixed numbers of parameters and doesn't asymptotically approach an optimal classifier with the increase in the number of training samples increases. The mean accuracy is higher for KNN classifier than decision tree and Naïve Bayes. [6] KNN requires careful choice of K value suitable to particular dataset. The accuracy in the recommendations of KNN is approximately 70%. [3] This implies that for 70% cases the KNN classifies the new instance of data correctly. KNN works well in the absence of any priori knowledge about distribution of the data. KNN is more scalable than Decision tree and Naïve Bayes classifiers. [3] For small datasets, the time taken by SVM is lower than other classifiers. KNN gives better accuracy than Naïve Bayes classifier. Classification of web log data using Naïve Bayesian method is one of the well-known approaches that improve the overall performance of the web server. [7] Naïve Bayes classifier is better in terms of memory utilization. The Naïve Bayesian has low time complexity but is not very efficient as per the error rate and the classified instances of the attribute values have been concerned. The experiment results of [17] suggests that with appropriate K value, KNN is better

than Neural Network with Back Propagation in terms of accuracy. Neural network is sensitive to noise but KNN is not. KNN yields smoother decision regions and provides probabilistic information. Neural Network with random sub-windows performs better than KNN and Simple Neural Network. [17] SVM has slower convergence rate in the testing phase. [18]

Following table summarizes the comparison of classifiers in terms of their advantages and disadvantages over each other.

Table 1 Comparision of supervised learning techniques

Classifier	**Advantages**	**Disadvantages**
Decision tree classifier	-Nonlinear relationships between parameters do not affect tree performance -Easy to interpret -Robust in the presence of noise	-Efficient only for small data sets -Requires the training set to be entirely in main memory -Scalability issue for larger data sets -Techniques developed to overcome the issues: SLIQ SPRINT, Rainforest, BOAT -Doesn't handle nonnumeric data well
Naïve Bayesian	-Scalable -Liner learning rate -Needs small training set -Higher accuracy and speed with large data sets -Works better with text classifiers	-Fails to produce a good estimate for the correct class probabilities -High error rate -No of irrelevant attributes are more
KNN	-Transparent -Easy to implement than most other machine learning techniques	-Slow while classifying test tuple -Lazy learner
	specifically when there is little or no prior knowledge about data distribution [6] -Scalable -Outperforms the cross-validated paired-differences t test over Naïve Bayesian and binary decision tree classifiers [5]	
SVM	-Most optimal classifier -Works well on higher dimensions -Efficient utilization of memory -Easier to train -Better than KNN for small training set	-kernel models can be quite sensitive to overfitting the model selection criterion -Selection of kernel is crucial for precise classification -Memory efficiency is very low [18]
Back propagation	-Efficient -Can approximate any function reasonably well	-Convergence time is high -Sensitive to the value of learning rate -Sensitive to the number of hidden layers and neurons -Sensitive to noise
Rule based classifier	-As highly expressive as decision trees -Can classify new instances rapidly	-Possibility of contradictions -Inefficiency -Needs data to be in refined form

VI. CONCLUSION

From the above comparison, it can be deduced that for classifying the web usage data the performance of K- Nearest Neighbour classifier is comparable with that of Naïve Bayesian and SVM classifier. Though it is a lazy learner and it typically does not process the training data, it has some fair advantages over others like scalability and ability to work when prior knowledge about dataset is very little or unavailable. It has been statistically proven with cross-validated paired-differences t test that it outperforms both Naïve Bayesian and SVM classifier for large datasets. Therefore, K-Nearest Neighbour is a suitable choice among all for classifying the web usage data.

REFERENCES

[1] Jiawei Han, Michaline Kamber, Jian Pei, Data mining concepts and techniques, 3rd ed., Elsevier Inc., 2012.

[2] [Online]. Available:http://www.slideshare.net/ashrafmath/naive-bayes-15644818

[3] D.A. Adeniyi, Z. Wei, Y. Yongquan, "Automated web usage data mining and recommendation system using K-Nearest Neighbour (KNN) classification method", Applied Computing and Informatics, ScienceDirect, Vol. 12, pp. 90-108, 2016.

[4] Thomas G. Dietterich, "Approximate Statistical Test For Comparing The Supervised Classification Learning Algorithms", December 1997.

[5] Paul Horton , Kenta Nakai, "Better Prediction of Protein Cellular Localization Sites with the Nearest Neighbours Classifier", American association for Artificial Intelligence, 1997.

[6] Mohammed Hamed Ahmed Elhiber and Ajith Abraham, "Access Patterns in Web Log Data: A Review", Journal of Network and Innovative Computing, Vol. 1, pp. 348-355, 2013.

[7] Chintandeep Kaur , Rinkle Rani Aggarwal, "Review on Classification of Web Log Data using CART Algorithm", International Journal of Computer Applications ,Vol. 80, No. 17 ,October 2013.

[8] D. Jayalatchumy, Dr. P.Thambidurai, "Web Mining Research Issues and Future Directions – A Survey", IOSR Journal of Computer Engineering (IOSR-JCE), Vol. 14, Issue 03, pp. 20-27, Sep. - Oct. 2013.

[9] K. Sudheer Reddy, G. Partha Saradhi Varma, and M. Kantha Reddy, "An Effective Pre-processing Method for Web Usage Mining", International Journal of Computer Theory and Engineering, Vol. 06, No. 05, October 2014.

[10] Sujith Jayaprakash, Balamurugan E., "A Comprehensive Survey on Data Preprocessing Methods in Web Usage Mining", International Journal of Computer Science and Information Technologies, Vol. 06, No 03, pp. 3170-3174, 2015.

[11] Sanjeev Dhawan, Swati Goel, "Web Usage Mining: Finding Usage Patterns from Web Logs", American International Journal of Research in Science, Technology, Engineering & Mathematics.

[12] Nirali H.Panchal, Ompriya Kale, "A Survey on Web Usage Mining", International Journal of Computer Trends and Technology (IJCTT), Vol. 17, No. 04, Nov. 2014.

[13] R. Aruna devi, K. Nirmala, "Analysis of Classification Algorithm in Data Mining", International Journal of Data Mining Techniques and Applications, Vol. 03, pp. 361-364, Jun 2014.

[14] Jaykumar Jagani, Kamlesh Patel, "An Enhanced Approach for Classification in Web Usage Mining using Neural Network Learning Algorithms for Supervised Learning", International Journal of Computer Applications, Vol. 90, March 2014.

[15] Ron Kohavi, "A Study of Cross Validation and Bootstrap for Accuracy Estimation and Model Selection", International Joint Conference on Artificial Intelligence, 1995.

[16] Sofia Visa, Brian Ramsay, Anca Ralescu, Esther van der Knaap, "Confusion Matrix-based Feature Selection", International Joint Conference on Artificial Intelligence, 1995.

[17] M. Seetha, K. V. N. Sunitha, G. Malini Devi, "Performance Assessment of Neural Network and K-Nearest Neighbour Classification with Random Subwindows", International Journal of Machine Learning and Computing, Vol. 2, No. 6, December 2012.

[18] Janmenjoy Nayak, Bighnaraj Naik, H. S. Behera, "A Comprehensive Survey on Support

Vector Machine in Data Mining Tasks: Applications & Challenges", International Journal of Database Theory and Application, Vol. 8, No. 1, pp. 169-186, December 2015.

[19] Anshul Bhargav, Munish Bhargav, "Neuro-Fuzzy Based Hybrid Model for Web Usage Mining", in Eleventh International Multi-Conference on Information Processing, Science Direct, 2015, pp 327 – 334.

[20] A. K. Santra, 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, No 2, January 2012.

[21] Zidrina Pabarskaite, "Decision trees for web log mining", journal of intelligent data analysis, Vol. 7, Issue 2, April 2003.

[22] A. K. Santra, S. Jayasudha, "Web Page Classification using an Ensemble of Support Vector Machine Classifiers", Journal Of Networks, Vol. 6, No. 11, November 2011.