A Comparative study of classification algorithms and a novel Pre-processing hybrid model for weblog mining

[1]*R.SHANTHI, PART-TIME RESEARCH SCOLAR, SATHYABAMA UNIVERSITY*
[2]*Dr.S.P.RAJAGOPALAN, PROFESSOR, G.K.M .COLLEGE OF ENGINEERING*

ABSTRACT: ***Web is the best medium of communication in modern business. Many companies are redefining their business strategies to improve the business output. In recent years, on-line system business breaks the time and space as compared to the physical office. Large companies around the world are realizing that e-commerce is not just buying and selling over internet, but also it improves the efficiency to compete with other companies. Web mining is the process of extracting useful information from web log server. Web mining is a data mining technique that is applied to the www. Large amount of information is available over the internet. This paper is an attempt to classify the customers based on their interests by using various classification algorithms and a hybrid model is proposed for the effective classification of customers which includes correlated fuzzy logic, k-means clustering and Naïve Bayes classification.***

INDEX TERMS: Fuzzy Logic, K-Means Clustering, Naïve Bayes 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.

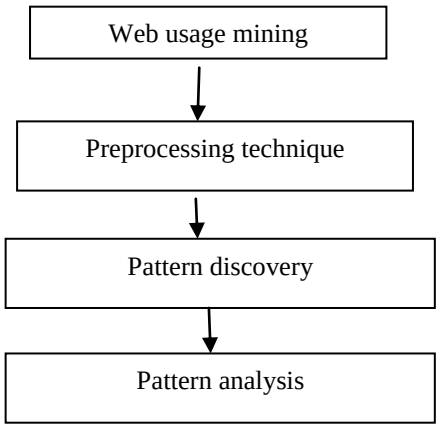
Web mining is broadly categorized into web content mining, web structure mining and web usage mining. All these categories make use of data such as web page content of HTML/XML code for the pages that may be linear or hierarchical. Any kind of web structure that is actual link structure and access information of web pages, such as number of hits, visits etc. Classification and clustering is the e main tasks in web mining for finding the relevant information. Classification is the process of finding the common properties among the set of objects in a database and classifying them into different classes. In the classification problem the class labels the labels of each class is known and discrete. Clustering is also the process of finding the common properties but the class labels are not-known. Classification is treated as supervised-learning algorithms/methods whereas clustering is treated as unsupervised-learning algorithms/methods. In this paper we focus on the web usage mining where, 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 an internet. What type of products they are looking for which includes characteristics of the products such as product price, quality of the product etc. Knowing the customers interests can be used in a variety in improvement of websites, e-commerce, website personalization, and user future request prediction and recommendation systems. In this paper, we compare the performance of the various classification algorithms for web log data.

II. LITERATURE SURVEY

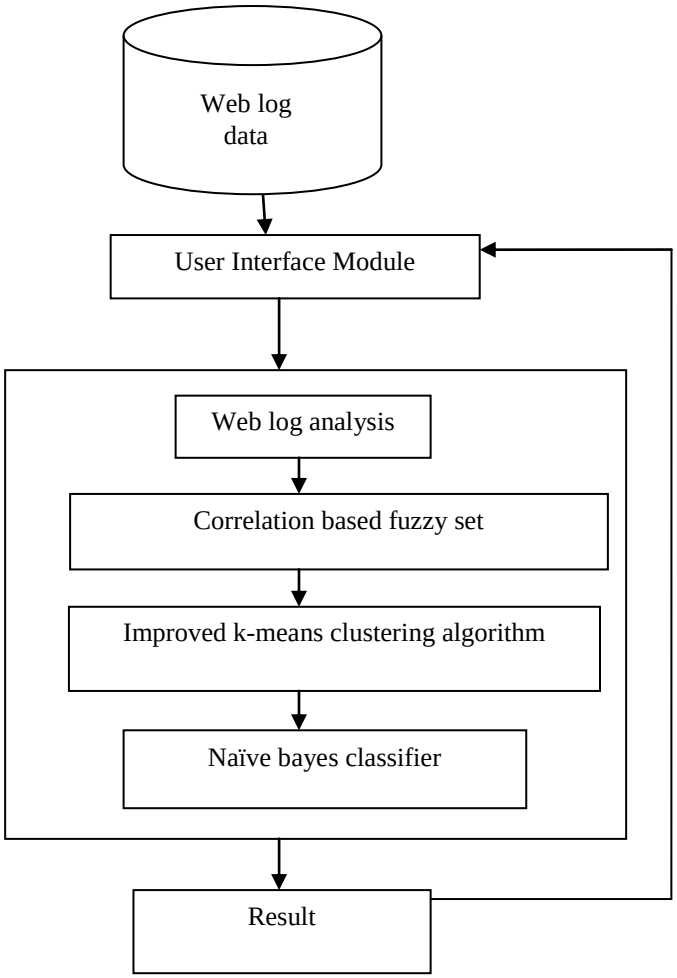
Several web usage mining and classification algorithms have been proposed. In this paper we compare the most significantly used classification algorithms for classifying the customers. Gupta [1] discusses about the web mining in the past, present and future applications. Mishra [2] explains the FP growth algorithm for frequently used pattern. Pre-processing techniques in web log data [3] has to be deal in such a way that irrelevant data and noise should be removed completely to obtain the required results.

Fuzzy logic [8] is providing the simplest way in arriving at the definite conclusion depending upon ambiguous, vague, noisy, missing input. Fuzzy logic is much faster in decision making. Clustering [9] is also applied in identifying the user profiles. Every user profile includes those users who are exhibiting the common behavior in browsing and are having similar interests.

III. SYSTEM ARCHITECTURE



Hybrid model



IV. PROPOSED WORK

Web usage mining deals with the application of data mining techniques to discover user access patterns from web server log. The aim of web usage mining is to capture the customer behaviors for a particular web site, so that recommendation can be done based upon their preferences. Usage mining tools helps the designer to improve the design of the web site by predicting and discovering the user behavior. On-line Data has become huge and lot of transactions happens in seconds. In web usage mining huge amount is data and analyzing the relevant data and also un-structured data is present in the web log. So, pre-processing techniques have become necessary for the extraction of the required information from web log.

Classification algorithms:

The need of knowing the users interest in a particular website or e-commerce site to analyze the user preferences has become essential due to the competition in e-business. Classification algorithms are used to categorize the customers.

- Decision tree classifier
- Naïve bayes classifier
- Support vector machines
- Neural networks
- Rule based classifiers
- k-nearest neighbor classifier

Decision tree classifier:

It is the simplest and widely used classification technique in a tree structure. Decision node specifies the test on the single attribute and the leaf node specifies the value of the target attribute. Arc/edge is used to split the attribute. Path is used to make the final decision. It applies an idea to solve the classification problem. Decision tree poses a series of questions about the attributes of the test record. Follow-up question is asked until a conclusion about the class label of the record is reached. Decision tree classifier is computationally expensive because at each node level each candidate splitting field must be before its best split can be found.

Naïve bayes classifier

It is simple probabilistic classifier. In this classifier bayes theorem is applied with strong independent assumptions. Naïve bayes probabilistic classifiers are commonly studied in machine learning because it is used to join the probabilities of words and categories to estimate the probabilities of categories given a document. Naïve bayes classifier is far efficient than the exponential complexity of non-naïve bayes approaches because it does not use word combinations as predictors

Support vector machines

SVM [7] are supervised learning models with associated learning algorithms that analyzed data and recognize patterns used for classification and regression analysis. SVM constructs a hyper plane or set of hyper planes in a high- dimensional space, which can be used for classification, regression, or other tasks. If the margin is larger error can be reduced. SVM simultaneously reduces the empirical classification error and maximize the geometric margin, and hence known as maximum margin classifiers.

Neural network

Neural networks are used for classification and prediction. It combines the input information in a very flexible way that captures complicated relationships among these variables and between them and the response variable. In neural networks the user is not required to specify the correct form but the network tries to learn about such relationships from the data with input, output and hidden layers

Rule based classifier

Rule based classifier makes use of the IF-THEN rules. If condition then conclusion. If part is called antecedent and then part is called consequent. In the antecedent part the condition consists of one or more attribute and the consequent part consists of class predictions which is very easy to interpret and generate.

K-nearest neighbor classifier (KNN)

K-nearest neighbor classifier (KNN) is a non-parametric method for classification and regression. K is a user defined constant assigning that, 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. An object is classified based upon its

neighbors and assigned to the class of the single nearest neighbor.

Novel hybrid model:

In this paper a hybrid model is proposed for classifying the customers. The proposed model is the combination of correlation fuzzy set algorithm. Improved k-means algorithm that is k-means clustering algorithm is used to identify the type of the customers and at last the effective classifier Naïve Bayes classifier is used to classify the customers.

Web log data

Data mining software, Mine set was used in data analysis. The usable customer data of 4263 e-commerce transactions were divided into two groups. Group 1 was about 70% of training data and grou-p2 was about 30% of the total transactions and is of tested data. Five factors are used in the data segmentation which includes age, on-line access time, language, behavior type, gender, address (at home or at work)

Fuzzy rough set:

Fuzzy set theory is used to represent the data uncertainty. Fuzzy set focuses on reasoning using natural languages in which words can have variety of meanings. Fuzzy set properties are used fuzzy rules in which they are used for decision making.

Correlation based fuzzy is used to improve the efficiency in analyzing customer data. The class levels of the customer data can be predicted and inter-relation among classes can also be identified by this combination.

K-means clustering

K-means clustering [4] is the simplest and effective clustering algorithm. The steps are:

1. Randomly select k users as the initial cluster center and consider the grading data of k users the term as the initial clustering center.
2. Integrate the rest users and calculate the similarity of users and k-clustering centers and distribute each user to the cluster of greater similarity.

Improved k-means clustering [5]

1. To the new cluster, calculate the average grading data that all users to the term and produce new clustering center.
2. Repeat the steps until cluster becomes constant.

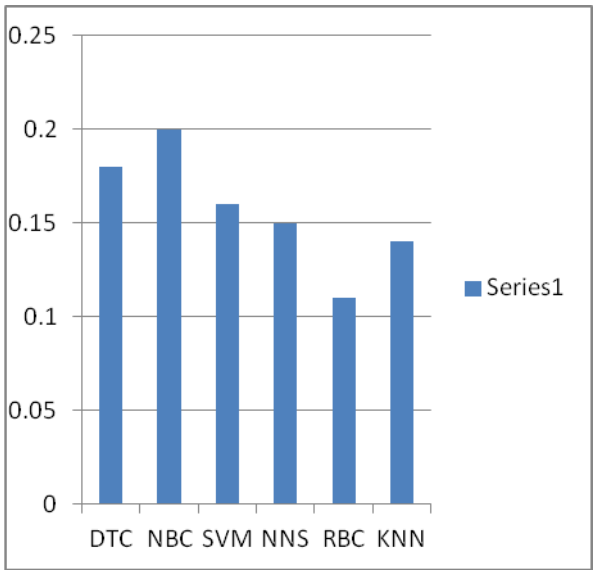
Naïve Bayes classification:

We have proved that Naïve Bayes classifier is an effective and efficient method for classifying the customer data from web log compared to other classification algorithms.

The user interface module retrieves the data from the web log that is from UCI repository. This data is sent to the hybrid module for preprocessing and classification. The web log is analyzed to capture the customer data by using fuzzy logic correlated system and k-means clustering algorithm. This module selects only the required customer data and then naïve bayes classification is used for classifying the customer based on the browsing behavior. Finally, results are produced which can be accessed by the user interface.

V. EXPERIMENTAL RESULTS

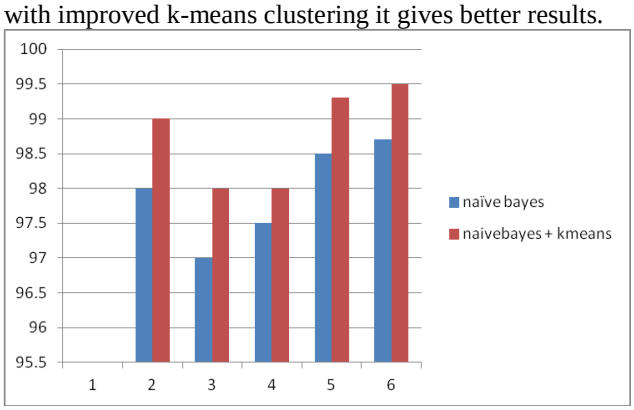
PERFORMANCE STUDY OF CLASSIFICATION ALGORITHMS



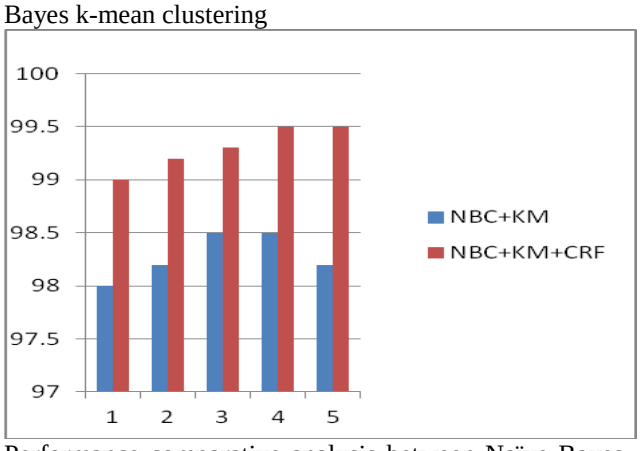
In this paper we present a framework for web usage mining [6] based on classification algorithms. With respect to all factors Naïve Bayes classifier performs well.

Hybrid model

Naïve bayes is the effective classifier but when combined with improved k-means clustering it gives better results.



Comparative analysis between Naïve Bayes and Naïve Bayes k-mean clustering



Performance comparative analysis between Naïve Bayes, k-mean clustering and correlated fuzzy set and Naïve Bayes, k-means

VI. CONCLUSION AND FUTURE ENHANCEMENTS

In the classification algorithms Naïve Bayes classifier yields better results compared to other algorithms. SVM and decision tree also gives better results. These algorithms can be combined to form better results in the future. Correlated fuzzy set plays a major role in customer analysis and recommendation systems but, when combined with improved k-means clustering and Naïve Bayes classification gives far better results. K-means clustering is used to cluster the nearest customers of similar behavior type. After grouping the customer data Naïve Bayes classifier is used for classification of the customers. The proposed hybrid model has improved the efficiency to 99.5%. Fuzzy logic system can be used more effectively for decision making.

REFERENCES

[1]. Aishwarya rastogi, guputa "web mining: a comparative study" international oural of computational engineering research, issn: 2250-3005, issue 2, vol.2, 2012
[2]. Rahul mishra and abha choubey "discovery of frequent patterns from web log by using fp-growth algorithm for web usage mining" international journal of advanced research in computer science and software engineering, issue 9, vol2, 2012
[3]. P.nithya and p.sumathi "novel pre-processing technique for web log mining by removing global noise", international journal of computer applications, vol.53, sep.2012
[4]. Yuantao jiang siqin yu,"mining the e-commerce data to analyze the target customer behavior" workshop on knowledge discovery and data mining, 2008.
[5]. Huang weijian and zhou xuqian "research of cluster-based data mining techniques in e-commerce"ieee, 2009
[6]. Q zhang, r.s.segall"web mining:a survey of current research techniques and software", journal of information technology and decision, 2008
[7].benjamin lenz and bernd barak, " data mining and support vector regression machine learning in semiconductor manufacturing to improve virtual metrology", international confereces on system sciences,,2013
[8]. Doru tanasa and brigitte trousse "advanced data preprocessing for intersites web usage mining"ieee, 2004
[9]. Rong h u, jianxun liu, wanchun dou" a clustering based colloborative filtering approach for big data application," ieee trns, october 2014.