Fusion of Selection Parameters for Customer Behavior Prediction based on Web Usage Mining

Ms. Snehal Kulkarni
Department of Computer Engineering
St. Francis Institute of technology
University of Mumbai,
Mumbai, India
snehalkulkarni@sfit.ac.in

Ms. Varsha Nagpurkar
Department of Computer Engineering
St. Francis Institute of technology
University of Mumbai,
Mumbai, India
varshanagpurkari@sfit.ac.in

Ms. Caroline Lopes
Department of Computer Engineering
St Francis Institute of Technology
University of Mumbai,
Mumbai, India
caroline.carry.lopes@gmail.com

Ms. Riya Dodthi
Department of Computer Engineering
St. Francis Institute of technology
University of Mumbai,
Mumbai, India
riyadodthi@gmail.com

Ms. Avril Lopes
Department of Computer Engineering
St. Francis Institute of technology
University of Mumbai,
Mumbai, India
avrilopes2@gmail.com

***Abstract*— System records the behavior and patterns of surfing of customers on websites.These records then mine and this will be used to identify their behavioral pattern through web mining. Ecommerce websites evaluate this data with the purpose of giving better services and also recommend improved options of products and services to customers. This system is customized to note down e-shopping and patterns of buying and track various data analytics features. The system noted down different factors like, stickiness, age of the user, product category, the day for purchasing and other site usage factors. Ecommerce sites have to review and mine for recorded data to monitor their website performance and frequently enhance it as per customer necessities or demands.**

Keywords—SVM, KNN, ANN using back propagation, Web Mining

I. INTRODUCTION

With the tremendous growth of e-businesses or the web-based e commerce system, the online shopping platform has become the greatest vital media for knowing one's consumers and their behavior. Marketing managers seek to gain major understandings of consumers" web navigation behavior" allow them to identify the most frequent visitors and hence derive tailored marketing strategies for loyal customers. Besides this, business consultants create major key performance indicators based on the web information; so that they could offer more personalized recommendations and suggest it to the marketing managers as well. Hence such a practice that analyses the web data and gives insights would be precious.

II. PROBLEM FORMULATION

Tracing of the user's activity on the web can be done through Clickstream i.e. how many moves or clicks a user makes when he or she is visiting the website, may provide us a lot about the way of thinking of the user if it is analyzed in an correct and accurate way with the help of different techniques and algorithms. These movements are sort of the behavior of online customers. We can term this analysis as web mining or web usage mining to find out what are the patterns of customers when they surf on websites. By considering their navigation style and their purchasing patterns we can come with certain innovative pricing strategies or can give some more benefits to the respective class of users

III. REVIEW OF LITERATURE

In paper [1], authors used KNIME. It is Eclipse based open source software for data mining. It exist in "decision tree" and "artificial neural network algorithms" and other open source platforms of data mining e.g. WEKA and R. Thus on KNIME platform we can run different programs and algorithms. For learning, different algorithms like decision trees and artificial neural networks have been used. Rules are formed with the help of decision tree algorithm. In this study bootstrapping is done by the author by using two decision tree algorithms.

The researchers have used are C4.5 and SPRINT i.e. Scalable Parallelizable Induction of Decision Trees. C4.5 algorithm uses entropy function to limit the best attribute or data field to arch from. SPRINT uses Gini function to choose the best data field to create branches. Neural networks are trained through an input layer, a set of hidden layers which depends on how many input parameters we take for the system and an output layer.

The simple algorithm for clustering is K-means clustering algorithm. This proposed in paper [2], for author's work the EPF K-Means algorithm, shows the gain of 12.3 % over conventional systems proving that it increases the clustering efficiency. And authors got the prediction accuracy as well.

In the paper [3], authors work on the theoretical aspect of correlation between the site search query data and daily e-commerce orders given by the customers. Then they have worked on orders and search data empirically ,authors built a model of search index with the help of historical data ,this search index is verifies the cointegration relationship .They come with the results that if customer wants to buy the product they search the information one to two days before they make an order to buy the product, and they found that there is statistically substantial correlation between daily e-commerce orders and respective site search data, and again this search data has good prediction ability for the daily e-commerce orders. The calculation of Pearson correlation

coefficient is done on every search query data with respect to daily e-commerce orders. They have predicted e-commerce orders on a daily basis while the trend was to forecast average e-commerce orders.

The authors of [4] also worked on web navigation behavior and mainly focused on sequence analysis. Methodology used is cluster analysis and "sequence analysis". This methodology makes groups of users based on their web navigation behaviors across different websites. The sequence analysis compares sequences in order to determine similarities and differences in those compared sequences. The algorithm used by this model is Optimal Matching distance (OM) algorithm for comparing sequence patterns. The data sets used in this study is acquired through a self-governing Internet- consulting firm in China, which has noted individual internet usage behavior of customers by using a client based tracking application. In this study, instead of individual site web activities they have considered crossed web sites activities , i.e. the authors have formed the clusters of user's browsing behavior based on their web activities and they have monitored their web activities for some decided duration. Based on the observations and review, ANN using back propagation will be the most suitable for prediction. Back Propagation Neural Network supports high speed classification and multi- classification. ANN using back propagation can be used for linear as well as nonlinear classification.

IV. PROPOSED SOLUTION

We have worked on artificial neural networks model. Neural networks take the input, learn through it and it has a set of hidden layers and an output layer. Setting up of a dummy website is required through which the data sets will be collected. These data sets will then be used for prediction with the help of ANN using back propagation algorithm.

A. System Flow:

The proposed system can be represented as given in Fig No. 1 major blocks of the system start with data preprocessing followed by classification using ANN using back propagation and finally prediction is done.

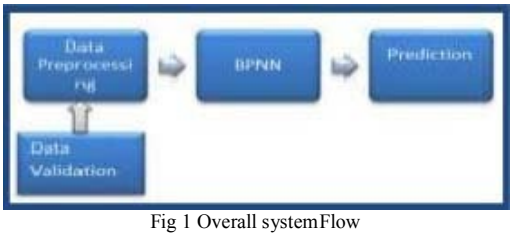

Fig 1 Overall systemFlow

B. Data Validation

1. The integrity and validity of the data can be validated using different mentioned techniques. Data validation predominantly helps in ensuring that the data sent to connected applications is complete, validated, accurate, secure and consistent. There are different ways to do validation like ANOVA, which is a statistical technique for correlation and covariance between the input and respective output data. For more accuracy proper categorization of training and testing data should be opted and will be done by K fold cross validation.
2. ANOVA: It stands for Analysis Of Variance. It is used to compare differences of means among more than 2 groups. Covariance, correlation gives you the relationship between two or more variables. So finding the correlation coefficient and covariance between the input, output factor gives the validation of the selected input and its effect on output. When we will be testing the different groups of each input data to see if there's a difference between them, ANOVA will help us to figure out if we should the null hypothesis or accept the alternate hypothesis.
3. K Fold Cross Validation: This validation procedure includes one parameter called k which gives the number of folds that a given data sample is to be parted into. So this method of selecting the proper folds of training and testing data is known as k-fold cross-validation. It uses sample in order to estimate how the model is expected to perform in general when used to make predictions on data not used during the training of the model.

   For this validation technique we have to follow following steps
   a) Consider the whole dataset.
   b) Split the dataset into k groups.
   c) Consider some folds for training i.e. if in 5 folds we have divided the data so first take first 3 folds for training
   d) Take the remaining groups as a testing data set.
   e) Repeat this training and testing for different fold's combinations and evaluate the training and testing errors
   f) Select the partition for training and testing whose error rate is lowest.

4 .Evaluation on training, testing and checking errors:
Training, testing and checking errors can be evaluated by "mean square error (MSE) "and root mean square error. MSE is the mean of the square of the error based on which the accuracy of the model can be evaluated as difference between actual and predicted values

C. Data Preprocessing

In Data preprocessing, the raw data is transformed into the required format using data normalization.Then this transformed information can be used further. Different selection parameters are selected for the said purpose of prediction. These parameters are decided after analyzing the current e-business scenarios and which are best suitable and impacting customers and the most according to Indian market.

Input selection Parameters (With Linguistic variables used) for prediction-

4. Amount of clicks: the system records the clicks when the customer is surfing through the web site .These amounts of clicks are categorized based on frequency/repetition. e.g. Low, Moderate, High
5. Clicked item: the system will record the items which the customer clicks on and will categorize the item based on the category to which it belongs. e.g.: apparel, home appliances or electronics.

6. Click number: the system will record that among a list of items which item was clicked on first, second, etc.
   E.g. firstly preferred, secondly preferred or preferred lastly.
7. Click time: the system records the stickiness of the user to a particular item i.e. amount of time spent looking at an item over a given time period. e.g.: Stickiness High , Medium or Low
8. Discounted item: the system will keep a record of weather the item added by the user to the cart is discounted or not. e.g.: Yes or No
9. Special day: the system will record if the day on which the customer is shopping is a special day or not. e.g.: Yes , No.Day of the week: the system will record if the day on which the customer is shopping is a weekday or the weekend.
10. Period of the day: the system will record the period of the day during which the customer is shopping. e.g.: Anti Meridiem, Meridiem, post Meridiem.
11. Cart: the system will record if the item that the customer is clicking on and viewing is finally added by the customer to the cart or not. e.g.: Yes , No
12. Gender: the system will record if the item is viewed by a male customer or a female customer.
13. Age: the system will record the age of the customer viewing the item and categorize them based on their age groups. e.g.: 14-20 , 21-30 or 31 and above
14. Demography: the system will record the location of the customer. e.g.: Mumbai, Bangalore, etc.

C. ANN using back propagation

Back Propagation Neural Network is a type of NN algorithm that will be used on the data sets for the prediction.

The ANN using back propagation works for non-linear and time series data. It is a multi-layer perceptron algorithm that has two forward and backward directions. Here in the training process three layers are present: input layer which accept the inputs as the parameters which we have decided, hidden layer according to number of parameters, and output layer which will give the final outcome. Due to the presence of the hidden layer, the error rate on ANN using back propagation can be minimized compared to single layer. The hidden layer in ANN using back propagation works to update and adjust the weights so that one will get a new weight value that can be directed to get the desired target. Weight adjustment on the parameters of the ANN using back propagation method is significant because it affects predicted results.

D. Prediction

It is the final result of the algorithm i.e. the output whether customer will buy the product or not.

Advantages of prediction:

1. Analyze customer's behavior patterns.
2. Recognize important customers.
3. Recognize the areas of the website that need improvement.
4. Recognize the demand for items and accordingly make changes in the availability of theseitems.

V. IMPLEMENTATION

A. Algorithm Used

The proposed algorithm for the given prediction is ANN using back propagation which is a supervised algorithm. ANN using back propagation processes learning in two steps

Step 1 Forward pass: In the first step, based on the inputs and associated weights, the outputs will be determined.

Step 2 Backward Propagation of Error: During the second step, error is determined by finding the difference between the predicted output and actual output.

Then this calculated error is back propagated to the preceding layer which is called as concealed or hidden layer. Lastly to get the optimized output i.e. minimum error, forward and backward steps will be repeated. ANN using back propagation can be shown as in fig no.2

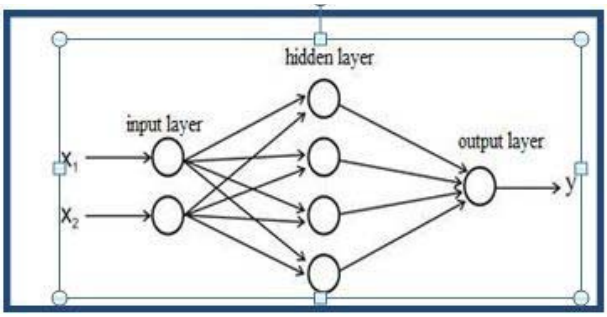

Fig no 2: ANN using back propagation structure

Here the rule based system will be formed using the input parameters for the desired output based on the training dataset which we have used.

Rule-based systems include the rules based on inputs and are quite similar as if-then rules. Some sample rules based on the inputs are given as,

- If the amount of clicks are high and/or click time is high and/or it is a special day and the item is discounted and the customer puts the product in basket then there is a probability that the user will buy the product.
- If the amount of clicks are low and/or click time is medium and/or day it is a week day and the item is not discounted and the customer doesn't put the product in the basket then there is a probability that the user won't buy the product.

For implementation the following steps were carried out during implementation,

1. An ecommerce website was created to collect data.
2. Hosted the website using grow without port forwarding.
3. Created all the tables and codes needed to

collect the datasets.

4. Shared the website link to collect clickstream data.
5. The collected data was pre-processed to fit the decided ranges. The following are the selection parameters with linguistic variables used as shown in Table number 1

Table no. 1. Selection Parameters (With Linguistic variables used) for prediction

Parameters	Linguistic Variables		
Amount of clicks	Low (0-2)	Avg.(2-6)	High >6)
Category	Makeup(0),	Clothing(1),	Footwear(2)
Product preference	High(2),	Medium(1),	Low(0)
Stickiness	High(2),	Medium(1),	Low(0)
Discounted item	Yes(1)	No(0)	
Special day	Yes(1)	No(0)	
Day of the week	Week day(1)	Weekend(0)	
More 5, 6 parameters are considered Like period of day ,gender, age,Demography etc.			

For creating a data set we have collected real time data by creating a website. The size of data set is greater than 10000.

VI. RESULT AND DISCUSSIONS

A. *Data validation on the dataset.*

Validation was done using: ANOVA (Analysis of Covariance).After performing ANOVA we found out that:

- The result is most affected by the Day parameter, I.e. whether it is week day, weekend etc…
- The result is least affected by the Cart parameter.
  - The Covariance of each parameter is given as in table no. 2

Table no.2 ANOVA Result

Parameter	Result(Buy)
Day	0.4498
Discount	0.3488
Special Day	0.2257
Time	0.2127
Gender	0.1875
Category	0.1678
Demography	0.1444
Stickiness	0.1148
Period	0.1057
Clicks	0.1047
Preference	0.0894
Age	0.0623
Cart	0.0148

Here from the above result we can conclude that the major parameter which affects the purchase is whether it's a normal day, its weekend or its any festival day, as in based on this factors the different discounts and offers will be there. So if a seller will provide attracting offers, discounts on these days' customers will tend to purchase the products. And on the contrary "cart" is the least decision factor, as it is observed that even if we put the product in the cart, it's not necessary that it will be a buy for the surfer.

B. *K-Fold Cross Validation:*

With the help of cross fold validations we have decided the best partition for training, checking and testing data the result are as shown in fig. No.3.

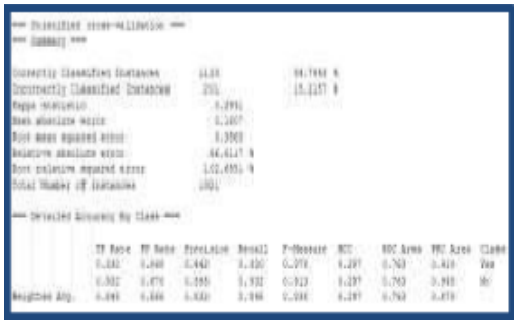

fig. number 3 K-Cross fold validation

C. *Validation was done for the predicted result*

Validation on the predicted result was performed using MAPE which is found to be 0.07%.And testing is done using black box and white box testing methods

VII. CONCLUSION

In this system,the implemented approach is to predict customer's shopping patterns from mouse clicks and website records or logs so we will be able to predict if a customer will finally purchase the items which he/she has added to his /her basket. Back propagation Neural Networks model (ANN using back propagation) is used to predict this online customer's behavior patterns. The main motive behind marketing a product is to fulfil demands and wants of the consumers. So consumer behavior helps to achieve this purpose. With the help of this prediction the seller will be able to convert a web surfer to consumer which can help to improve the sales of the seller.

As a future scope, we can further develop this system into a recommendation system where based on a customer's behavior, certain products will be recommended to him/her and once the system will generate the possibility that the user won't be buying the product, can work on dynamic change in pricing for the customer so that he tend to purchase the product.

REFERENCES

[1] G. Sılahtaroğlu and H. Dönertaşli, "Analysis and prediction e-customers' behaviour by mining clickstream data," 2015 IEEE International Conference on Big Data (Big Data), Santa Clara, CA, 2015, pp. 1466-1472.

[2] S. Cheriyan and K. Chitra, "Web page prediction using Markov model and Bayesian statistics," 2017 Second International Conference on Electrical, Computer and Communication Technologies (ICECCT), Coimbatore, 2017, pp. 1-6.

[3] Q. Jiang, C. Tan, C. W. Phang and K. K. Wei, "Using Sequence Analysis to Classify Web Usage Patterns across Websites," 2012 45th Hawaii International Conference on System Sciences, Maui, HI, 2012, pp. 3600-3609.

[4] L. Na, P. Geng, C. Hang and B. Jiaxing, "A prediction study on E-commerce orders based on site search data," 2013 6th International Conference on Information Management, Innovation Management and Industrial Engineering, Xi'an, 2013, pp. 314-318.