International Journal of Current Trends in Engineering & Technology
ISSN: 2395-3152
Volume: 01, Issue: 06 (September – October 2015)

To Improve Web Performance based on Horizontal Partition of Decision Tree with Layer Architecture

Dharmendra Kumar
Research Scholar, MTech(CSE)
MTRI, Bhopal

Prof. Dhirendra Kumar Jha
HOD (CSE)
MTRI, Bhopal

Abstract:
The World Wide Web has evolved in less than two decades as the major source of data and information for all domains. Web has become today not only an accessible and searchable information source but also one of the most important communication channels, almost a virtual society. The goal of fundamental analysis is to decide the value of a neighborhood preference based on the previously access web pages by users. It receives considerable attention from both researches and practitioners. Here two layers maintain authorization and authentication by users. It also maintains neighborhood profile similarity for next prediction of users. The main achievement of this research is to provide decision tree based of Horizontal Partition with Genetics Algorithm. This analysis is important for the prediction of different close value of website. The existing work for the analysis doesn't provide efficient results and has more error rate. Hence here in this paper we combine two techniques of classification and genetic algorithm to increase the efficiency of the website performance. The classification of items in website is used to provide classification among different values of the user profile and then genetic algorithm provide a close among these values.
Keywords: Fuzzy C-means, Sequential Pattern Mining, Association Rule Mining, SOM clustering.

I. INTRODUCTION

Data mining is primarily used today by many companies with a strong consumer focus just like retail, financial, communication, and marketing organizations areas. It is used to determine the relationships of internal factors such as product positioning, price, or staff skills in the company. It is also determine external factors just like economic indicators, customer interest and the market competition strategy as in the store company. The entire factors are used to make company profits, increase the sales and customer satisfaction etc. It also shows the summary information to view details of transactional data. A retailer store is per day purchase by customer in database for finding some targeted promotions. It is based on an individual's purchase history by mining demographic data. This analysis help to retailer for generation of products and different promotions offer to specific customer segments. The enormous amount of data normally stored in files, databases, and other repositories. It is used to extract of interesting knowledge for analysis and interpretation of data that help in decision making. Data mining and knowledge discovery (or KDD) are frequently treated as synonyms in databases. It is actually part of the knowledge discovery process which having some steps in an iterative knowledge discovery process.

Neural Networks/Pattern Recognition: Neural network is a set of connected input/output units. Each connection has a weight present with it. It predicts the correct class labels of the input tuples during the learning phase of network learns by adjusting weights. It has the remarkable ability to derive meaning from complicated or imprecise data. It can be used to extract patterns and detect trends that are too complex. These are well suited for continuous valued inputs and outputs e.g. handwritten character reorganization, for training a computer to pronounce English text and many real world business problems. It is identifying the patterns or trends in data which well suited for prediction or forecasting needs.

International Journal of Current Trends in Engineering & Technology
ISSN: 2395-3152
Volume: 01, Issue: 06 (September – October 2015)

Clustering Analysis: Clustering analysis is unsupervised classification technique which group the items based on similarity basis. It groups the web users according to web page access pattern in website. It also provides personalized web content to the individual user for market segmentation in e-commerce application. Web page clustering is useful for Internet search engines and Web service providers. There are many clustering approaches which are based on the maximizing the similarity between objects in a same class and minimizing the similarity between objects of different classes. The different types of clustering methods are - a) Partitioning Methods, b) Density based methods, c) Hierarchical Agglomerative (divisive) methods, d) Model-based methods, and e) Grid-based methods etc.

Genetic Algorithms: GA is a technique which performs like bacteria growing in a Petri dish. The data set gives ability to do different things for whether a direction or outcome is favorable. It optimizes the final result which is used mostly for process optimization, such as scheduling, workflow, batching, and process re-engineering.

Decision Tree/Rule Induction: Decision trees use real data mining algorithms. It helps with classification and split out information that is very descriptive, helping users to understand their data. A decision tree process will generate the rules followed in a process. For example, a lender at a bank goes through a set of rules when approving a loan. Based on the loan data a bank has, the outcomes of the loans (default or paid), and limits of acceptable levels of default, the decision tree can set up the guidelines for the lending institution. These decision trees are very similar to the first decision support (or expert) systems.

Web Mining is the extraction of interesting with potentially useful patterns and implicit information from artifacts or activity related to the World Wide Web. It automatically discovers and extracts information from website. It categorize into three areas (as shown in Figure 1.5) of interest based on which part of the web to mine -

a) Web Content Mining, b) Web Structure Mining, c) Web Usage Mining.

II. RELATED WORK

The web mining research is a converging area from several research communities, such as Databases, Information Retrieval, Machine Learning and Natural Language Processing. Due to the widespread computerization and affordable storage facilities, enormous wealth of information is embedded in huge database belonging to different enterprise or scientific experiment. It provides a tremendous interest in the areas of Knowledge Discovery and Data Mining. These areas have motivated allowed statistician and data miners to develop faster analysis tools that can help to analyze the stockpiles of data, turning up in to valuable and often surprising information.

Omar Zaarour et al. (2013) proposed an improvement the web log mining procedure for the prediction of online navigational pattern. Their contribution contains three different components. First they proposed for session identification, a refined time-out based heuristic. Secondly, suggested the practice for navigational pattern detection by using a specific density based algorithm. Finally, a new method for efficient online prediction is also recommended to improve the applicability and effectiveness of the website.

Rahul Moriwal et al. (2013) presented a method for Finding Frequent Sequential Traversal Patterns from Web Logs which is based on Dynamic Weight Constraint, where various frequent sequential pattern mining algorithms have been proposed that mines the set of frequent subsequences pattern which satisfying a min-support constraint in a particular session database. Though, previously sequential pattern mining algorithms gives equal weight age to sequential traversal patterns whereas the pages in sequential patterns have different importance and also have different weight age. Other problem in most of the frequent sequential pattern mining algorithms is that a large number of sequential patterns is generates, when min-support is lowered and here they do not have any alternative ways for adjusting the number of sequential patterns other than increment in the minimum support. The proposed frequent sequential pattern mining algorithm with weights constraint main purpose is to

append the weight constraints in to the sequential pattern while maintaining the downward closure property. In this a weight range is defined for maintaining the downward closure property. The pages are given dissimilar weights and traversal sequences assign a minimum and maximum weight. For scanning a session database maximum and minimum weight in the session database is utilized to cut infrequent sequential subsequence and by this downward closure property is maintained.

V. Chitraa et al. (2012) proposed method, presented, analyzed, and evaluated is to automatically give the actual value of k and select the right initial points based on the datasets objects. The algorithm enhances the k- means clustering algorithm by finding initial points and optimize for accurate results. This algorithm selecting initial points is more complex than the random methods, but this algorithm is stable, running it in different times, the clustering results obtained are the same, the random algorithms cannot ensure this, and different initial points lead to different running time on random algorithms, compared with proposed algorithm, its running time is uncertain and more long.

Nayana Mariya Varghese et al. (2012) are proposed cluster optimization technique using fuzzy logic. Web page access pattern is collected from web log file as input and then eliminate irrelevant data items. The cleaned web log is used for pattern discovery. The web page personalization is used for clustering of web pages. It is based on similar usage of web access patterns by users. Some clustering algorithms have some drawbacks when the number of web user is increased, because the size of cluster also increases. The proposed algorithm is used for eliminating the redundancies occur in data based on fuzzy logic after clustering optimization methodology.

Nanhay Singh et al. (2013) proposed a new framework to improve the performance of web proxy server through cluster (k-means algorithm) based prefeching schemes (LRU and LFU) and Apriori algorithm is applied to generate rules for web pages. Web caching is used to minimize the network traffic at the proxy server level by caching web pages. There is demand to improve the cache performance by using the prefetching technique. It fetches the objects from database and store in advance that are likely to be accessed in the near future. This will result reduction of the response time of the user request.

Ketki Muzumdar et al. (2013) proposed a method to discover useful knowledge by obtaining secondary data from the access pattern of the web users. The proposed method uses Self Organizing Map (SOM), which is a kind of neural network approach to detect user's patterns. It shows the comparison between the traditional K-Means with SOM algorithm. This process describes the transformations necessaries to modify the data storage in the Web Servers Log files to an input of SOM. Neural Network based method has shown that the trend analysis performance depends on the number of requested cluster.

Srishti Taneja et al. (2014) introduced a novel algorithm proposed which uses some features of algorithm of Univariate tree and if noise remains then that can be removed by implementing some features of Multivariate algorithm with some additional features of new algorithm to be designed.

III. PROPOSED ALGORITHM

The proposed Efficient Prediction of neighborhood users close value using Genetic algorithm based horizontal partition decision tree The proposed methods is implemented using genetic algorithm which includes the concept of decision tree. The scheme is used to find next item prediction in website regarding as a challenging task of the website performance. Genetic algorithms (GAs) are problem solving methods (or heuristics) that mimic the process of natural evolution. Unlike artificial neural networks (ANNs), designed to function like neurons in the brain, these algorithms utilize the concepts of natural selection to determine the best solution for problem.

Algorithm 1 : Genetic Algorithm

for all members of population
sum += fitness of this individual
end for
for all members of population
probability = sum of probabilities + (fitness / sum)
sum of probabilities += probability
end for

International Journal of Current Trends in Engineering & Technology
ISSN: 2395-3152
Volume: 01, Issue: 06 (September – October 2015)

loop until new population is full
do this twice
number = Random between 0 and 1
for all members of population
if number > probability but less than next
probability then you have been selected
end for
end
create offspring
end loop

Algorithm 2: Horizontal Partitioned Based Decision Tree

Define P_1, P_2… P_n Parties.(Horizontally partitioned).
Each Party contains R set of attributes A_1, A_2, …., A_R. C the class attributes contains c class values C_1, C_2… C_c.
For party Pi where i = 1 to n do
    If R is Empty Then
    Return a leaf node with class value
Else If all transaction in T(P_i) have the same class Then
Return a leaf node with the class value
Else
Calculate Expected Information classify the given sample for each party Pi individually. Calculate Entropy for each attribute (A_1, A_2… A_R) of each party P_i. Calculate Information Gain for each attribute (A_1, A_2… A_R) of each party Pi
End If.
End For
Calculate Total Information Gain for each attribute of all parties (TotalInformationGain ( )).
$A_{BestAttribute}$ □ MaxInformationGain ( )
Let V_1, V_2… V_m be the value of attributes. $A_{BestAttribute}$ partitioned P_1, P_2… P_n parties into m parties
P_1 (V_1), P_1 (V_2) ... P_1 (V_m)
P2 (V_1), P_2 (V_2)… P_2 (V_m)
. .
. .
P_n (V_1), P_n (V_2) … P_n (V_m)
Return the Tree whose Root is labeled $A_{BestAttribute}$ and has m edges labeled V_1, V_2… V_m. Such that for every i the edge V_i goes to the Tree
NPPID3(R – $A_{BestAttribute}$, C, (P_1 (V_i), P_2 (V_i)… P_n (V_i)))
End.

Algorithm 3: TotalInformationGain ( ) - To compute the Total Information Gain for every attribute.

For j = 1 to R do {Attribute A_1, A_2… A_R }
Total_Info_Gain (A_j) = 0
For i = 1 to n do {Parties P_1, P_2,…., P_n }
Total_Info_Gain(A_j) = Total_Info_Gain(A_j) + Info_Gain(A_{ij})
End For
End For
End.

Algorithm 4: MaxInformationGain( ) – To compute the highest Information Gain for horizontally partitioned data

MaxInfoGain = -1
For j = 1 to R do {Attribute A_1, A_2,…., A_R }
Gain = TotalInformationGain(A_j)
If MaxInfoGain < Gain then
MaxInfoGain = Gain
ABestAttribute = A_j
End If
Return ($A_{BestAttribute}$)
End For
End.

IV. RESULT ANALYSIS

A decision tree is a tree structure which having node and branches. Here every node having attribute with condition and branches. Every leaf node belongs to a class just like a group of users or condition data. It can handle both categorical and numerical data. It is used in operations research area for decision analysis which provides help to identify a strategy to reach a goal. Online selection model algorithm also used this tree for better response. It can also use as a descriptive means in web mining for calculating conditional probabilities.

International Journal of Current Trends in Engineering & Technology
ISSN: 2395-3152
Volume: 01, Issue: 06 (September – October 2015)

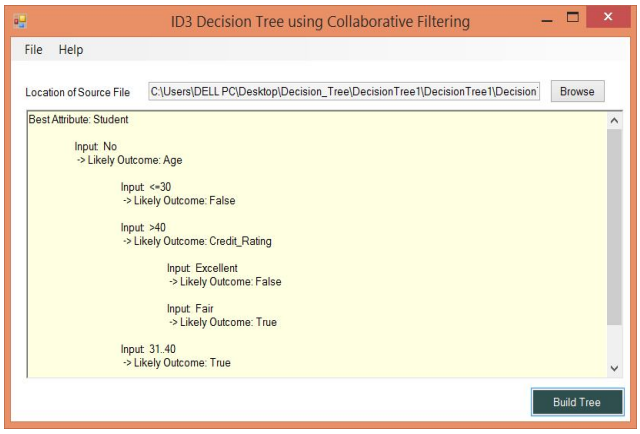

Figure-1: Decision Tree of Web Transaction Database

The Figure-1 shows the decision tree of web transaction database by using highest info-gain.

The Table-1 show the time complexity comparison between existing ID3 based decision tree and Horizontal partition based decision tree and found that the proposed algorithm has less complexity when experimented on different values of dataset.

Table-1: Execution Time of ID3 and HP-Decision Tree

No. of Instances	ID3 Time (ms)	HP Time (ms)
20	78	22
40	92	38
60	112	53
80	123	72
100	137	83
200	158	93

The Figure-2 shows the execution time of ID3 and HP-Decision Tree algorithm. It shows the execution time of HP-Decision is less as compared to HP-Decision tree.

Mean Absolute Error (MAE)

MAE is used to compute the error in time series analysis. The time series is homogeneous space. It must be identical in size. The MAE is given by -

$MAE(X, Y) = (SAE / N) = (\sum_{i=1}^{N} |X-Y|)/N$

Where $\{X_i\}$ is the actual observations time series,

$\{Y_i\}$ is the estimated or forecasted time series,

SAE is sum of the absolute errors,

N is no. of non-missing data points.

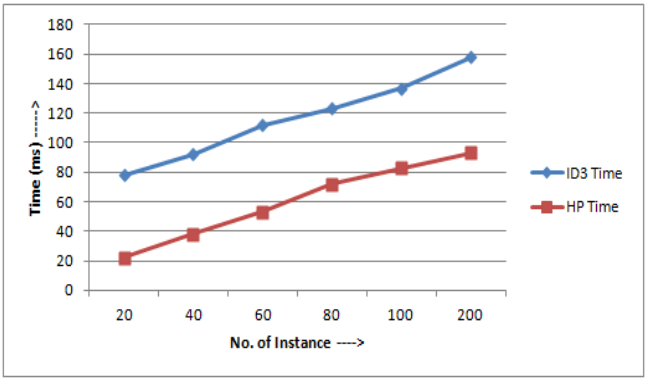

Figure-2: Execution Time of ID3 and HP-Decision Tree

The Table-2 shows the mean absolute error rate between ID3 and HP-Decision tree. It show that HP-Decision Tree have less error rate as compared to the existing ID3 decision tree.

Table-2: Mean Absolute Error of between ID3 and HP-Decision Tree

No. of Instances	ID3_Mean Absolute Error	HP_Mean Absolute Error
20	0.2350	0.2234
40	0.2610	0.2156
60	0.2935	0.2385
80	0.3241	0.3121
100	0.3578	0.3010
200	0.4261	0.4124

V. CONCLUSION

The genetic algorithm can be used as an application to predict the close values in the website data and the performance factor of this algorithm is much better than ANN. Here proposed algorithm is based on the integration of genetic algorithm and the horizontal partition based decision tree and compare the performance factor of the proposed algorithm as compared to the existing decision tree and the proposed algorithm performs better for current trends in the website.

International Journal of Current Trends in Engineering & Technology
ISSN: 2395-3152
Volume: 01, Issue: 06 (September – October 2015)

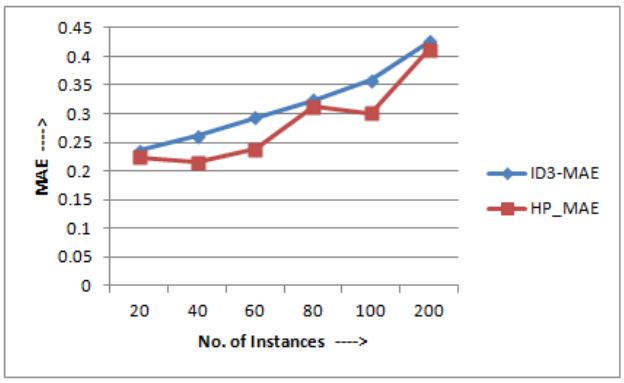

Figure-3: Comparison of MAE between ID3 and HP-Decision Tree

The Figure-3 shows the comparison of mean absolute error between ID3 and HP-Decision tree. It show that HP-Decision Tree have less error rate as compared to the existing ID3 decision tree.

REFERENCES

1. Omar Zaarour, Mohamad Nagi, ***"Effective web log mining and online navigational pattern prediction"***, ELSEVIER,2013.
2. Rahul Moriwal and Vijay Prakash, ***"An Efficient Algorithm for Finding Frequent Sequential Traversal Patterns from Web Logs based on Dynamic Weight Constraint"***, 2013**.**
3. Ketki Muzumdar, Ravi Mante, Prashant Chatur, ***"Neural Network Approach for Web Usage Mining"***, International Journal of Recent Technology and Engineering (IJRTE), Vol.-2, Issue-2, May-2013.
4. V. Chitraa, Dr. Antony Selvadoss Thanamani, ***"An Enhanced Clustering Technique for Web Usage Mining"***, International Journal of Engineering Research & Technology (IJERT), Vol.1, Issue 4, June-2012.
5. C. Umapathi, M. Aramuthan and K. Raja, ***"Enhancing Web Services Using Predictive Caching"***, International Journal of Research and Reviews in Information Sciences (IJRRIS), Vol.-1, No.-3, Sept-2011.
6. Song Sun and Joseph Zambreno, ***"Design and Analysis of a Reconfigurable Platform for Frequent Pattern Mining"***, IEEE, Vol.22, No.9, Sept-2011.
7. S. Vijayalakshmi, V. Mohan, S. Suresh Raja, ***"Mining of Users Access behavior for Frequent Sequential Pattern from Web Logs"***, International Journal of Database Management Systems (IJDMS) Vol.2, No.3, August 2010.
8. Mahdi Esmaeili and Fazekas Gabor, ***"Finding Sequential Patterns from Large Sequence Data"***, International Journal of Computer Science (IJCSI), Vol.7, Issue 1, No.1, January 2010.
9. Priyanka Makkar, Payal Gulati, Dr. A.K. Sharma, ***"A Novel Approach for Predicting User Behavior for Improving Web Performance",*** International Journal on Computer Science and Engineering (IJCSE), Vol. 02, No. 04, 2010, 1233-1236.
10. Utpala Niranjan, Dr. R.B.V. Subramanyam, Dr. V.Khanaa, ***"An Efficient System Based On Closed Sequential Patterns for Web Recommendations"***, IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 3, No 4, pages 26-34, May 2010.
11. Jinlin Chen, Member, IEEE, ***"An Up-Down Directed Acyclic Graph Approach for Sequential Pattern Mining",*** IEEE Transactions on Knowledge and Data Engineering, Vol. 22, No. 7, pages 913-928, July 2010.
12. D.Vasumathi and A. Govardhan, **"BC-WASPT : Web Access Sequential Pattern Tree Mining"**, International Journal of Computer Science and Network Security (IJCSNS), Vol.9, No.6, June-2009.
13. Qingqing Gan, Torsten Suel, ***"Improved Techniques for Result Caching in Web Search Engines"***, ACM, pp. 20-24, 2009.
14. Cui Wei, Wu Sen, Zhang Yuan and Chen Lian-Chang, ***"Algorithm of mining sequential patterns for web personalization services"***, ACM SIGMIS Databases, vol. 40, No. 2, pp 57-66, May 2009.
15. Hua Jiang,Dan Zuo, Xin Hu, Yong-Xin Ge, Bin Han, ***"UAP-Minar:A Real-Time Recommendation Algorithm Based on User Access Sequences"***, 7th Int'l Conf. on Machine Learning and Cybernetics, Kunming, 12-15 July 2008.
16. F. Masseglia, P. Poncelet, M. Teissseire, ***"Efficient mining of sequential patterns with time constraint: reducing the combinations"***, Expert systems with applications Elsevier, Vol. 40, N. 3, 29 pp: 2677-2690, 2008.
17. Jatin D Parmar, Sanjay Garg, ***"Modified Web Access Pattern (mWAP) Approach for Seqential Pattern Mining"***, Jan. 2007.
18. Unil Yun and John J. Leggett**,** ***"WSpan: Weighted Sequential Pattern Mining in Large Sequence Databases",*** Proc. Of the Third Int'l Conf. on IEEE Intelligent Systems, pages 512-517, Sep. 2006.
19. Ezeife, C. and Lu, Y. **,** ***"Mining Web Log Sequential Patterns with Position Coded Preorder Linked WAP-Tree"***, International Journal of Data Mining and Knowledge Discovery (DMKD) Kluwer Publishers, pp.5-38, 2005.
20. D. Chiu, Y. Wu and A.L.P. Chen., ***"An Efficient Algorithm for Mining Frequent Sequences by a New Strategy without Support Counting",*** 20th Int'l Conf. on Data Engineering(ICDE), pp. 375-386, Boston, MA, USA, 2004.
21. M. Eirinaki, M. Vazirgiannis, ***"Web Mining for Web Personalization"***, in ACM Transactions on Internet Technology (TOIT), 3(1), pages 1-29, February 2003.
22. B. Mobasher, H. Dai, T. Luo, Y. Sung, J. Zhu, ***"Discovery of Aggregate Usage Profiles for Web Personalization"***, in Proceedings of the Web Mining for ECommerce Workshop (WEBKDD'2000), Boston, August 2000.
23. Dharamveer Sisodia and Beerendra Kumar, ***"Efficient Prediction of Close Value using Genetic Algorithm based Horizontal Partition Decision Tree in Stock Market"***, International Journal of Advance Research in Computer Science and Management Studies, Volume 2, Issue 1, Pages 412-418, January 2014.
24. Abhishek Gupta and Dr. Samidha D. Sharma, ***"Clustering-Classification Based Prediction of Stock Market Future Prediction,"*** International Journal of Computer Science and Information Technologies (IJCSIT), Volume 5(3), Pages 2806-2809, 2014.