ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 10, December 2012

Implementation of Decision Tree Algorithm to Analysis the Performance

Pooja Sharma[1], Asst. Prof. Rupali Bhartiya[2]

Department Of Computer Science and Engineering,Shree Vaishnav Institute of Technology & Science, Indore, India[1,2]

ABSTRACT: Data Mining is a very interesting area to mine the data for knowledge. Several techniques are available which makes data mining remarkable. Web mining is also a part of that kind of data mining techniques. Web mining includes data preprocessing, pattern discovery and pattern analysis phase to process the log data. Demand of analyzing and extracting knowledge from different domain databases increases. Classification is a technique to predict the best classifier. In model build methods classification algorithm plays an important role. In this paper we are implementing a proposed decision tree algorithm and existing C4.5 algorithm for comparative study and to analysis the performance.

Keywords: Data mining, Web usage mining, Decision tree algorithm, C4.5 and Cross validation.

I. INTRODUCTION

Web mining is the application of data mining techniques for analysis and extracting useful information from the data set. Web mining works for the three kinds of data like web content, structure and usage of web resources. Web usage mining is the process of discovery of patterns from the transaction stored on the web servers. There are many data sources available for web usage mining-automatic generated log data stored in server, user profiles, meta data, page content and e-commerce for shopping,product rating. Web usage mining plays a significant role to discover prediction of the next page, application objects, distinguish users according to predefined classes, groups of the similar properties and interest, common behavior users. In web usage mining, data preprocessing of session, URL, transaction of data occurs and unwanted data is removed by data cleaning process. Through these processes data is prepared and algorithms applied to build a model. Classification uses training data set to classify the data according to attribute value into predefined classes. Decision tree algorithms are based on classification using attribute values for taking decisions. Decision tree classifies data from the root node to a leaf node till decisions not made. Decision tree represents data in a fashion which is very easily interpreted by users.

II. BACKGROUND

Srivastava [8] has done surveys of web usage mining with the growth of web based application. The main interest of analysis of web usage data and uses that characterization. In paper [2] problem of mining access pattern from log data analyzed. Web access pattern tree (WAPTree) is used to store access pattern but in a very compressed manner. Comparative study of classifier accuracy in user profiling is represented in [4]. Four different classification algorithms are Naïve Bayesian (NB), Instance-Based Learner (IB1), Bayesian networks (BN) and Lazy Learning of Bayesian Rules (LBR) compared to analyze the classifier accuracy. NB and IB1 classifiers performed better than the BN and LBR classifiers with respect to classification accuracy. This paper compares the performance of NB, IB1, Classification and Regression Tree (SimpleCART), Naïve Bayesian Tree (NBTree), Iterative Dichotomister Tree (Id3), J48 -a version of C4.5- and Sequential Minimal Optimization (SMO) algorithms with large user profile data. Decision tree classification algorithms are widely used because of its easier representation of results and easy to implement [7]. Implementation of Decision tree can be done in serial or parallel manner based on scalability, memory used and volume of data.

III. LOG DATA FORMAT

In web usage mining we analyze log data which stored at the server side, client side and proxy side. The format of access log data looks like-

127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200 2326

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 10, December 2012

"http://www.example.com/start.html" "Mozilla/4.08 [en] (Win98; I ;Nav)"

To implement decision tree algorithm we are using log data in ARFF format. Attribute relation file format(ARFF) file shows a list of instances sharing a set of attributes. ARFF developed by machine learning project for using Weka machine learning software[15]. ARFF files have two sections- Header information and Data information. The format of ARFF file for internet access is shown

@relation 'KDDTest-21'
@attribute 'duration' real
@attribute 'protocol_type' {'tcp','udp', 'icmp'}
@attribute 'service' {'aol', 'auth', 'bgp', 'courier', 'csnet_ns', 'ctf', 'daytime', 'discard', 'domain', 'domain_u', 'echo', 'eco_i', 'ecr_i', 'efs', 'exec', 'finger', 'ftp', 'ftp_data', 'gopher', 'harvest', 'hostnames', 'http', 'http_2784', 'http_443', 'http_8001', 'imap4', 'IRC', 'iso_tsap', 'klogin', 'kshell', 'ldap', 'link', 'login', 'mtp', 'name', 'netbios_dgm', 'netbios_ns', 'netbios_ssn', 'netstat', 'nnsp', 'nntp', 'ntp_u', 'other', 'pm_dump', 'pop_2', 'pop_3', 'printer', 'private', 'red_i', 'remote_job', 'rje', 'shell', 'smtp', 'sql_net', 'ssh', 'sunrpc', 'supdup', 'systat', 'telnet', 'tftp_u', 'tim_i', 'time', 'urh_i', 'urp_i', 'uucp', 'uucp_path', 'vmnet', 'whois', 'X11', 'Z39_50'}
……

@data
13,tcp,telnet,SF,118,2425,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,1,1,0.00,0.00,0.00,0.00,1.00,0.00,0.00,26,10,0.38,0.12,0.04,0.00,0.00,0.00,0.12,0.30,anomaly
0,udp,private,SF,44,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,4,3,0.00,0.00,0.00,0.00,0.75,0.50,0.00,255,254,1.00,0.01,0.01,0.00,0.00,0.00,0.00,0.00,anomaly
0,tcp,telnet,S3,0,44,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1.00,1.00,0.00,0.00,1.00,0.00,0.00,255,79,0.31,0.61,0.00,0.00,0.21,0.68,0.60,0.00,anomaly
0,udp,private,SF,53,55,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,5
11,511,0.00,0.00,0.00,0.00,1.00,0.00,0.00,255,255,1.00,0.00,0.00,0.87,0.00,0.00,0.00,0.00,0.00,normal
…….

IV. PERFORMANCE ANALYSIS

To analyze the performance of proposed algorithm comparison with decision tree algorithm takes place. After implementing algorithms, a resultant model is built to measure the correctness of results.

A. Comparative study of C4.5 and proposed algorithm in context of accuracy:

Here we are using C4.5 algorithm to compare with proposed algorithm. Here we include the results obtained by the system in five experiments shown in table I.

TABLE I
Accuracy comparison

Data set size	Proposed model	C4.5
102	72.36	77.64
428	73.82	72.48
821	69.24	73.82
1029	76.21	74.93
2193	72.63	70.67

We plot graph for the measured accuracy in which red line shows C4.5 algorithm and blue line shows proposed algorithm. With the help of graph we can see that accuracy of C4.5 is very high when the data size is less. But accuracy of C4.5 algorithm decreases with the increase of data size. Accuracy of proposed model is better than C4.5 for large data size.

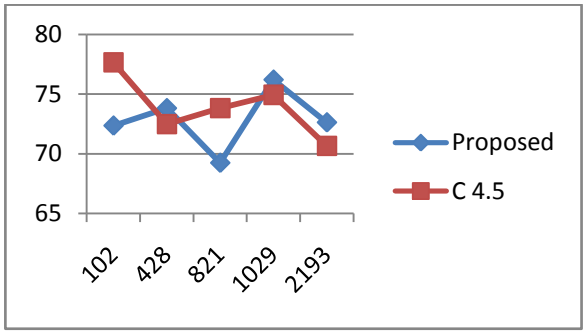

Fig. 1 Graph represents accuracy comparison for C4.5 and proposed algorithm

B. Comparative study of C4.5 and proposed algorithm in context of evaluation time:

Different data size is used to analysis the performance of algorithms. Some algorithms give best results for a small data set but not for large data set.

TABLE II
Time comparison

Data set size	Proposed model	C4.5
102	2.398	2.187
428	5.712	4.723
821	8.289	8.179
1029	13.298	12.378
2193	14.278	15.378

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 10, December 2012

According to our analysis we found that model evaluation time simulate similarity there is hardly any difference in evaluation time for small no. of data. But when the data size is large the evaluation time of proposed algorithm is less than the C4.5.

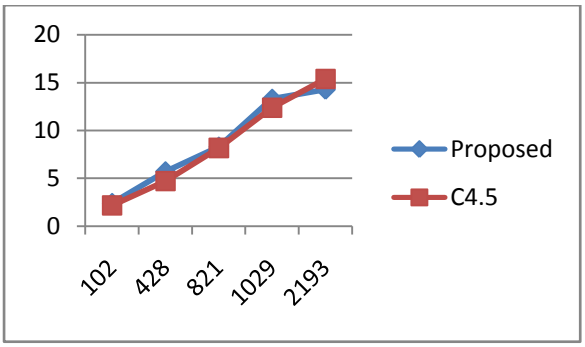

Fig. 2 Graph represents evolution time comparison for C4.5 and proposed algorithm

C. Comparative study of C4.5 and proposed algorithm in context of Memory used:

Comparison table and graph shows that for a small number of data memory used by proposed algorithm is high and for large data is low.

TABLE III
Memory Used

Data set size	Proposed model	C 4.5
102	26391	27822
428	27382	28423
821	28347	27412
1029	28490	28327
2193	29384	28832

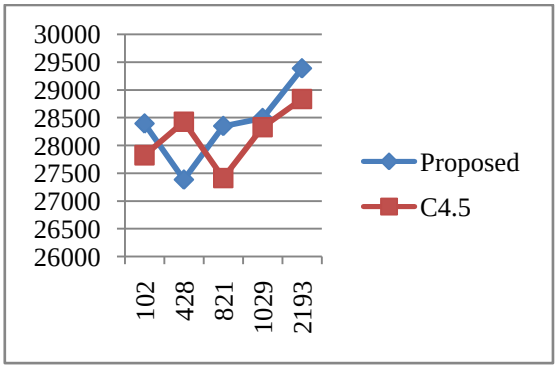

Fig. 3 Graph represents memory used for C4.5 and proposed algorithm

D. Comparative study of C4.5 and proposed algorithm in context of Build time:

TABLE IV
Build Time

Data set size	Proposed model	C 4.5
102	3.273	4.897
428	5.187	6.272
821	9.248	10.437
1029	13.289	13.382
2193	16.389	15.392

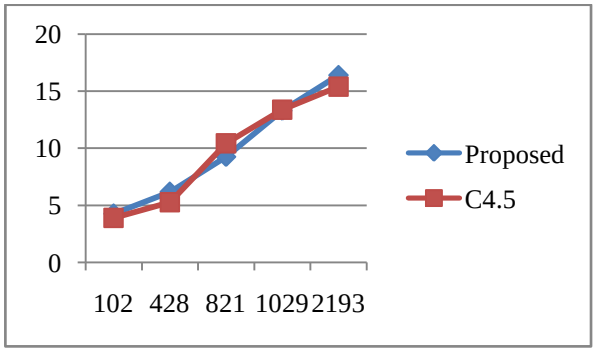

Fig. 4 Graph represents build time for C4.5 and proposed algorithm

Build time of proposed algorithm is also low and is better than C4.5 algorithm

V. CONCLUSION

World wide web is a very lucrative area in the field of computers. Now a days web plays an important part in human life and has become a very good business of earning. Business carried on the web offers the opportunity to find potential customers and their interests. Billions of people regularly/daily accessing the internet for searching different kinds of information. Therefore web server gathers bulk data every day. WUM facilities to analyze the web log files. For pattern discovery in WUM decision tree classification algorithms are used. Implementation of proposed decision tree algorithm results improved performance of WUM.

ACKNOWLEDGEMENT

I would like to thank Mrs. Rupali Bhartiya (Assistant professor, Shri Vaishnav Institute of Technology and Science) for their mentorship.

ISSN (Print) : 2319-5940
ISSN (Online) : 2278-1021

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 1, Issue 10, December 2012

REFERENCES

[1] K. R. Suneethe, Dr. R. Krishnamoorti, "Identified User Behavior By Analyzing Web Server Access Log File", Published in Computer Science and Network Security, April 2009 International journal, Vol. 9 No. 4, PP. 327-332.

[2] Jian Pei, Jiawei Han, Behzad Mortaza vi-asl, and Hua Zhu, "Mining Access Patterns Efficiently from Web Logs", School of Computing Science, Simon Fraser University, Canada fpeijian han, mortazav, hzhuag@cs.sfu.ca

[3] Theint Theint Aye,"Web Log Cleaning for Mining of Web Usage Pattern", Publised in 2011 IEEE.

[4] Ayse Cufoglu, Mahi Lohi, Kambiz Madani," A Comparitive Study of Selected Clasifiers with Classification Accuracy in User Profiling", Published in World Congress on Computer Science and Information Engineering(CSIE), 2009 IEEE International Conference,ISBN: 978-0-7695-4/08.

[5] Zhu Xiaoliang, Wang Jian,Yan Hongcan, Wu Shangzhuo,"Research and Application of the improved algorithm C4.5 on Decision Tree", Published in Test and Measurement (ICTM), 2009 IEEE International Conference,ISBN: 978-1-4244-4700-8/09.

[6] Brijendra Singh, Hemant Kumar Singh," Web Data Mining Research: A Survey, 2010 IEEE International Conference,ISBN: 978-1-4244-5967-4/10.

[7] Matthew N. Anyanwu, Sajjan G. Shiva,"Comparitive Analysis of Serial Decision Tree Classification Algorithms", Published in of Computer Science and Security (IJCSS), International Journal, Volume (3): Issue (3).

[8] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan,"Web Usage Mining: Discovery and Application of Usage Patterns from Web Data", Published in ACM SIGKDD, Jan 2000, Volume 1, Issue 2, pp. 12-23.

[9] Sasa Bosnjak, Mirjana Maric, Zita Bosnjak,"The Role of Web Usage Mining in Web Applications Evaluation",Published in Management Information Systems,2010, Vol. 5, No. 1, PP. 031-036.

[10] Song Danwa, Han Ning, Liu dandan, "Construction of forestry resource classification rule decision tree based on ID3 Algorithm", Published in First International Workshop on Education Technology and Computer Science (ETCS), 2009 IEEE.

[11] Jeffrey Xu Yu, Yuming Ou, Chengqi Zhang and Shichao Zhang,"Identifying Interesting Visitors through Web Log Classification", Published by IEEE Computer Society, May/June 2005.

[12] Ting Zhong Wang," The Development of Web Log Mining Based on Improve-K-Means Clustering Analysis", Advances in CSIE, Vol. 2, AISC 169, pp. 613–618.springerlink.com.

[13] Mohamed Koutheaïr Khribi, Mohamed Jemni and Olfa Nasraoui "Automatic Recommendations for E-Learning Personalization Based on Web Usage Mining Techniques and Information Retrieval", Published in Educational Technology & Society, 2009 IEEE, 12 (4), 30–42.

[14] Subhash K.Shinde, Dr.U.V.Kulkarni,"A New Approach For On Line Recommender System in Web Usage Mining", Published in International Conference on Advanced Computer Theory and Engineering (ICACTE), 2008 IEEE

[15] M. Mihut, "Analyzing Log Files using Data-Mining", published in Journal of Applied Computer Science, no.2 (2) /2008, Suceava

BIOGRAPHY

Pooja Sharma is a student of Master of Engineering, Shri Vaishnav Institute of Technology and Science, Indore, Madhya Pradesh, India. She has received B.E. degree in Computer Science and Engineering.