Machine Learning Algorithm to Identify the Fault Data Identification Using Multi-Class Support Vector Machine

Dr. R. Rajesh Kanna
Assistant Professor, Department of Computer Applications,
Dr. N.G.P. Arts and Science College (Autonomous), Coimbatore

Abstract: An experiment was conducted to the raw web log files, in a controlled lab environment, by using KDD technique and M-SVM algorithm. Based on the experiment conducted, the M-SVM algorithm generates 98.68% for true positive rate and 1.32% for false positive rate which indicates the significant efficiency of the new web log file classification and data transformation technique used in this research work. M-SVM model identified fault data identification in more accurate with less time when compared to existing SVM model.

Keywords: *Web log data, SVM, Machine Learning, Web mining*

*****

I. INTRODUCTION

Web usage mining is application of data mining techniques to discover usage patterns from web data, in order to better serve the needs of web based applications.

These days, huge amount of data are generated with different formats and an efficient technique to convert these data is very important. The 5Vs of data in an group continue to expand than its ordinary level. In current situation, the existing Web log files are informative based on website. However, even fault occurrence is high as the data increases enormously. Because of fault occurrence, the browser provides fault pages with the pages to be searched pages. The manual process is not possible because of large amount of complex data. In the existing systems, identifying failure occurrence in web log file is difficult and time-consuming task. The basic reason is the large size and complexity of these systems, and the vast amount of monitoring data they generate. In existing system, fault classification technique does not provide highest accuracy to improve the website.

II. RELATED WORK

A. Vinupriya and S. Gomathi.,2016 [1] proposed a fresh out of the box new plan named as WPP (web page Personalization) for effective net page recommendations. WPP comprise of page hit depend, finish time spent in every hyperlink, number of downloads and connection detachment. Established on these parameters the personalization has been proposed.

A. Yang et al.,2014 [2] have granted an answer that initially distinguishes the clients whose kNN's conceivably tormented by the recently arrived content, after which supplant their kNN's individually. Creators proposed another file constitution named HDR-tree keeping in mind the end goal to support the compelling hunt of influenced clients.

G. Dhivya et al.,2015 [3] dissected individual lead by utilizing mining advanced web section log data. The few net interaction mining approaches for extricating profitable components used to be talked about and utilize every one of these systems to bunch the clients of the area to consider their practices extensively. The commitments of this proposal are a data enhancement that was substance and beginning spot arranged and a treelike perception of bland navigational groupings.

J. Jojo and N. Sugana.,2013 [4] proposed a half breed approach which utilizes the insect established grouping and LCS order techniques to search out and foresee client's route conduct. Subsequently client profile may likewise be followed in powerful pages. Customized inquiry can be utilized to address extend in the web look group, established on the preface that a purchaser's ordinary decision may simply help the mission motor disambiguate the genuine aim of an inquiry.

A. U. R. Khan et al.,2015 [5] have exhibited a cloud transporter to clarify how the status of the broad communications news can be evaluated using clients online use propensities. Creators utilized information from Google and Wikipedia for this correlation challenge. Google data was useful in comprehension the affect of stories on web looks while data from Wikipedia empowered us to comprehend that articles identified with rising data content additionally discover parcel of consideration.

M. A. Potey et al.,2013 [6] inspected and contrasted the with be had ways to deal with display an understanding into the train of question log handling for ability recovery.

M. Nayrolles and A. Hamou-Lhadj.,2016 [7] proposed BUMPER (BUg Meta repository for developers and Researchers), a standard framework for engineers and specialists curious about mining data from numerous (heterogeneous) vaults. Guard used to be an open supply web-established condition that concentrates data from an assortment of BR stores and variation control frameworks. It was once outfitted with a solid web index to help clients rapidly inquiry the vaults using a solitary purpose of get to X.

III. Format of Web Log Files

Raw log files are contains information about website visitor activity. Log files are created through web servers automatically. Each time a visitor requests any file (page, image, text etc.) from the web site information on the request is appended to a current log file. Most log files have text format and each log entry (hit) is saved as a line of text. The web logs do not use graphics, such as graphs and charts. Here is a sample of log entry in Apache combined format:

```
213.135.131.79 - [06/June/2018:19:21:49-0400] "GET
/faculty.htm HTTP/1.1" 200
9955/http://www.drsnsrcas.ac.in/download.htm"
"Mozilla/4.0 (compatible; MSIE 6.0;
Windows NT 5.1; Q312461)"
```

Web Logs

A web log is a listing of page reference data sometimes it is referred to as click stream data. Web plays an important role for extracting useful information. There is a need for data log to track any transaction of the communications. This data can offer valuable information insight into website usage. It characterizes the activity of many users over a potentially long period of time.

IV. SUPPORT VECTOR MACHINES

It is based on the Structural Risk Minimization principle for which error-bound analysis has been theoretically motivated. The method is defined over a vector space where the problem is to find a decision surface that “best” separates the data points into two classes.

In order to define the “best” separation, a margin has to be introduced between two classes. A decision surface in a linearly separable space is a hyper plane. The SVM problem is to find the decision surface that maximizes the margin between the data points in a training set.

Methodology

Different web servers keep different types of information in the log files. The activity is recorded in web log file when user submit request to a web server. The information in the log files as listed in the following.

User Name: Detects who had visited the web and user identification based on the IP address that is assigned by the Internet Service Provider (ISP).

Visiting Path: The path taken by the user while visiting the website by using search engine, typing Uniform Resource Locator (URL) directly or by clicking on the link.

Path Traversed: Detects the path taken by user using several links within the web site.

Time Stamp: The session or time spent by the user in web site while browsing.

Page Last Visited : The latter page visited by the user before he or she leaves the website.

Success Rate : The rate of success for the website which is determined by the number of copying activity done by the user and the number of downloads.

User Agent : User’s browser information.

URL : The resources accessed by the user.

Request Type : The information transfer method such as GET and POST.

V. MULTI CLASS SUPPORT VECTOR MACHINE

The classical SVM classifiers were originally designed for binary classifications. However, there are more than two classes in some practical classification problems. For instance, one has to divide the potential squeezing effect into several classes according to the magnitude of the normalized convergence so that the severity of the squeezing could be adequately assessed or predicted. This constitutes a typical multiclass classification problem.

Two commonly used strategies for constructing M-SVM are the “one-against-one” and “one-against-all” approaches. In the “one-against-one” approach, we build one SVM for each pair of classes, which means that if there are k classes, then $k(k-1)/2$ binary SVM classifiers are constructed to distinguish the samples of one class from the samples of another class. In the classification, we use a voting strategy, that is, each binary classification is considered to be a web log files, where k by web log data can be cast for all the samples. To predict a new instance, we choose the classifier with the largest decision function value.

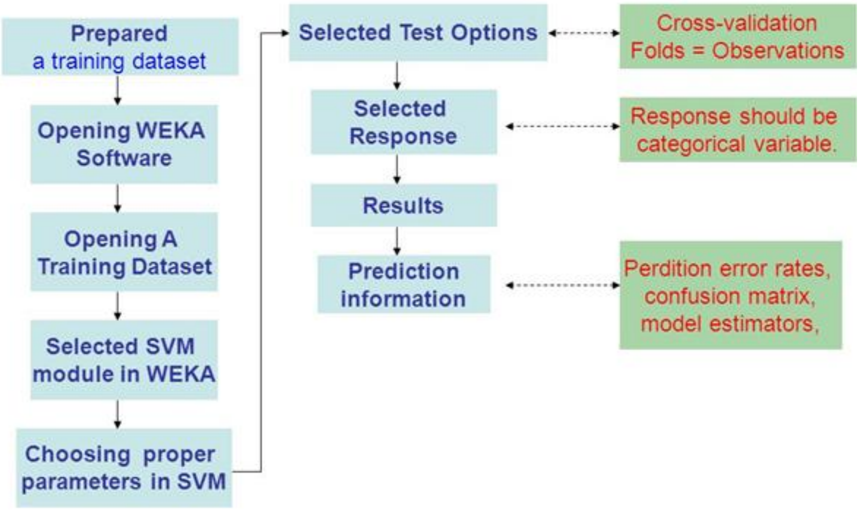

Prepared a training dataset
Opening WEKA Software
Opening A Training Dataset
Selected SVM module in WEKA
Choosing proper parameters in SVM
Selected Test Options
Cross-validation Folds = Observations
Selected Response
Response should be categorical variable.
Results
Prediction information
Perdition error rates, confusion matrix, model estimators,

Figure 2: SVM working nature in Weka

Results and Discussion

The web log data used in M-SVM model experiments are collected from Apache web access log files. The web log files have different levels of faults including error, attack, debugging, warning error, etc. The input dataset size is 50 MB with structured format and unstructured format. In Identifying web fault module is implemented to extract fault related from web log data through the regular expression pattern of websites. Each pattern in the regular expression represents unique information. While applying M-SVM String Matching algorithm ignores the inaccurate and incomplete information data from raw web log data. To simulate the data obtained from heterogeneous sources, web log data using various structured format and stored into training dataset. Classify the identified fault depends upon their category using M-SVM classifier. M-SVM classifier classify the data effectively when compare to existing algorithm.

Accuracy

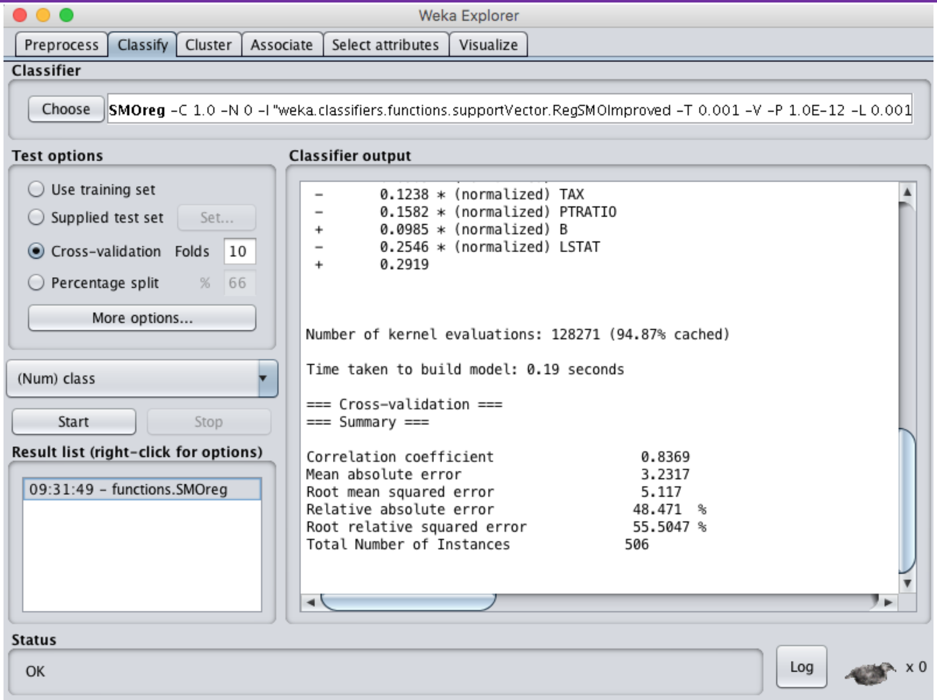
A metric to evaluate the overall performance of fault diagnosis is throughput rate. It can recognize any repetitive structures within a Web log file. Those recognized records may belong to different categories. In additional evaluate the processing time and complexity of fault in the website.

*Accuracy= (Total number of classified fault web log files/ Total number of web log files)*100*

Throughput

Throughput indicates the number of faults occurrences in the web log files can handle, the amount of faults identified over time during a test. Lots of fault types identified from different web log files. To ensure that, load and performance testing is the solution. Also before starting a performance test it is common to have a throughput goal that the application needs to be able to handle a specific number of fault classifications per hour. In below Figure 4, represents the comparative result analysis of proposed(M-SVM) and existing(SVM) system. Throughput performance provides the efficiency of M-SVM. The x axis represents the occurrence of fault from various web log files. The y axis represent the processing time of fault.

Throughput time = Processing time + Identification Fault time + Waiting time + Inspection time



Classifier

SMOreg -C 1.0 -N 0 -I "weka.classifiers.functions.supportVector.RegSMOImproved -T 0.001 -V -P 1.0E-12 -L 0.001

Test options

○ Use training set

○ Supplied test set

● Cross-validation Folds 10

○ Percentage split % 66

(Num) class

Result list (right-click for options)

09:31:49 – functions.SMOreg

Classifier output

```
 -       0.1238 * (normalized) TAX
 -       0.1582 * (normalized) PTRATIO
 +       0.0985 * (normalized) B
 -       0.2546 * (normalized) LSTAT
 +       0.2919



Number of kernel evaluations: 128271 (94.87% cached)

Time taken to build model: 0.19 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.8369
Mean absolute error                      3.2317
Root mean squared error                  5.117
Relative absolute error                 48.471  %
Root relative squared error             55.5047 %
Total Number of Instances              506
```

Status

OK

Figure 3 : M-SVM Classification for fault identification in Web log data

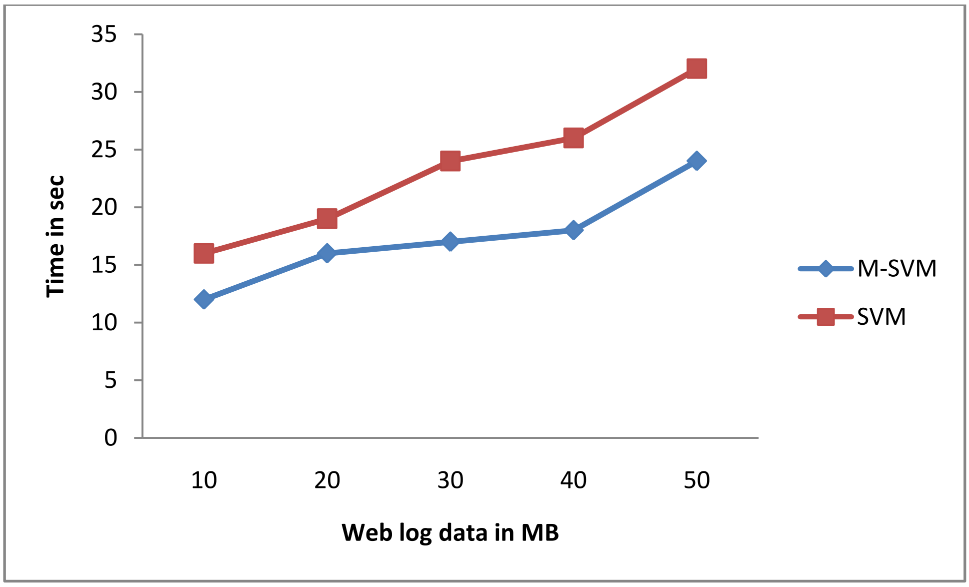


Figure 4 : Throughput vs Fault identification in sec

CONCLUSION

Web is an interface which is used to access remote data, commercial and non-commercial services. Web log file is a file that is automatically created and maintained by a web server. Log files contain the information about the users like user name, visiting path, the path traversed, time stamp, page last visited, success rate, user agent and URL.

In this research paper discussed a detailed review of web log files like web server data, application server data, application level data, web server logs, log file parameter types of log file format, various locations of web log files and types of web log files and identifies the faults in web log files. This work will also improve the loyalty and reliability of the web sites.

REFERENCES

[1] A.Vinupriya and S. Gomathi, "Web Page Personalization and link prediction using generalized inverted index and flame clustering," 2016 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, 2016, pp. 1-8.

[2] A.Yang, X. Yu and Y. Liu, "Continuous KNN Join Processing for Real-Time Recommendation," 2014 IEEE International Conference on Data Mining, Shenzhen, 2014, pp. 640-649.

[3] G. Dhivya, K. Deepika, J. Kavitha and V. N. Kumari, "Enriched content mining for web applications," Innovations in Information, Embedded and Communication Systems (ICIIECS), 2015 International Conference on, Coimbatore, 2015, pp. 1-5.

[4] J. Jojo and N. Sugana, "User profile creation based on navigation pattern for modeling user behaviour with personalised search," Current Trends in Engineering and Technology (ICCTET), 2013 International Conference on, Coimbatore, 2013, pp. 371-374.

[5] A.U. R. Khan, M. B. Khan and K. Mahmood, "Cloud service for assessment of news' Popularity in internet based on Google and Wikipedia indicators," Information Technology: Towards New Smart World (NSITNSW), 2015 5th National Symposium on, Riyadh, 2015, pp. 1-8.

[6] M. A. Potey, D. A. Patel and P. K. Sinha, "A survey of query log processing techniques and evaluation of web query intent identification," Advance Computing Conference (IACC), 2013 IEEE 3rd International, Ghaziabad, 2013, pp. 1330-1335.

[7] M. Nayrolles and A. Hamou-Lhadj, "BUMPER: A Tool for Coping with Natural Language Searches of Millions of Bugs and Fixes," 2016 IEEE 23rd International Conference on Software Analysis, Evolution, and Reengineering (SANER), Suita, 2016, pp. 649-652.

[8] T. Cheng, K. Chakrabarti, S. Chaudhuri, V. Narasayya and M. Syamala, "Data services for E-tailers leveraging web search engine assets," Data Engineering (ICDE), 2013 IEEE 29th International Conference on, Brisbane, QLD, 2013, pp. 1153-1164.

[9] A. Jebaraj Ratnakumar, “An Implementation of Web Personalization Using Web Mining Techniques”, Journal Of Theoretical And Applied Information Technology, 2005,2010 JATIT

[10] Tsuyoshi, M and Saito, K., “Extracting User‟s Interest for Web Log Data”, IEEE 2006, pp. 343-346, ISBN: 0-7695-2747-7