Scientific Journal of Impact Factor(SJIF): 3.134

e-ISSN(O): 2348-4470
p-ISSN(P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 1,Issue 11, November -2014

NEURAL NETWORK CLASSIFICATION ALGORITHMS FOR WEB USAGE MINING AND PROPOSED SOLUTIONS FOR HUGE WEB DATA CLASSIFICATION

Disha Patel[1], Shraddha Joshi[2]

[1] *Department of Computer Engineering, Student of PG studies-MEF Group of Institutions, patel.disha62@gmail.com*
[2] *Department of Computer Engineering, Faculty of PG studies-MEF Group of Institutions, shraddha.joshi@marwadieducation.edu.in*

Abstract — *Nowadays, a huge amount of data is present on web, so to extract useful knowledge and to manage those huge files will become mandatory to obtain fruitful business analysis results. To extract useful knowledge from World Wide Web (WWW) is known as web mining. Web usage mining has emerging trends on network traffic control and flow analysis, website management, personalization, etc. Neural network have capability of self organization and is also matched with ant colony behavior and adaptive learning. Such concept is used for information retrieval from huge web data. It is also used for complex classification, optimization and distributed control problems [1].With the help of Neural Network algorithms for classification of web log data; it produces the best result of classification. So, in this paper we have introduced solutions for self-organizing and growing network which helps in information retrieval from huge web data and also discussed various neural network algorithms i.e. GNG (Growing Natural Gas), ART(Adaptive Resonance Theory) model, LVQ(Learning Vector Quantization) and its series. Input for neural algorithms is web log files and expected outcome would be optimal representation of network that is further used for Information extraction in web usage mining. Such trained network is used for classification which gives effective classification of data.*

Keywords- *Web usage mining, Classification, Learning vector Quantization, artificial neural networks, Web log data, GNG, ART*

I. INTRODUCTION

Nowadays the data on the web are growing exponentially as numbers of users are increasing every day, so with the growing of data, data storage and usage make the network (WWW) complex and in future it may be unhandled. So, to retrieve useful knowledge from huge data available on the web, few efficient techniques of web usage mining are useful. Web mining is mainly classified in 3 major categories [2], content mining, usage mining and structure mining and. depending upon what to mine, the above techniques are used. Now, web usage mining helps to deal with various web scaling problems such as user trend analysis, traffic flow analysis, web traffic management and many more[1].Using the concepts of neural network; traffic sharing on distributed servers ,Session tracking, website reorganization can also be identified [2] and analysis based on web data. So, applying data mining techniques on web logs, server log files resulted in useful usage path extraction, session tracking, session duration, number of session creations, and website reorganization[1],[2], [3],[5].

As the Web data is increasing over the time, to find the useful pattern or information efficiently, Ant colony behavior, self organizing [2] concept should be used for the network. Such neural network concept is very useful for adapting manageable usage mining from Web. Neural network is far different from static networks in which, network becomes intelligent because each node is self-intelligent. So, web users can use this network more and more. This concept is widely useful for extracting information in web usage mining for web traffic analysis on live servers and frequent usage path analysis and many more concepts. In this paper, the discussion is based on mining process and parameters and use of neural network and many such algorithms such as LVQ, GNG, ART model, to overcome from various issues and have a great mining result.

II. WEB USAGE MINING ELEMENTS

Web usage mining depends on certain parameters such as, ipaddress, time stamps (creation / release), size of data pages, method [5], authenticated users, distance, Location of web servers & clients. The process of path extraction, session tracking, frequency usage of various services, usage duration has been carried out according to such parameters. We can extract useful knowledge such as traffic analysis, congestion assumptions when traffic is more in particular region which causes low speed reply or session time out scenarios by using this various parameters. So, on the basis of above parameters we can extract hidden and useful information about web usage and it would be very useful for business analysis, traffic control and so on.

2.1 Brief Process of Web Usage Mining:

According to the figure 1,The only difference in data mining and web is that at the initial level in data will come from various data bases and warehouses and in web mining data will come from server log files ,the processing steps and various techniques are similar to the data mining process.

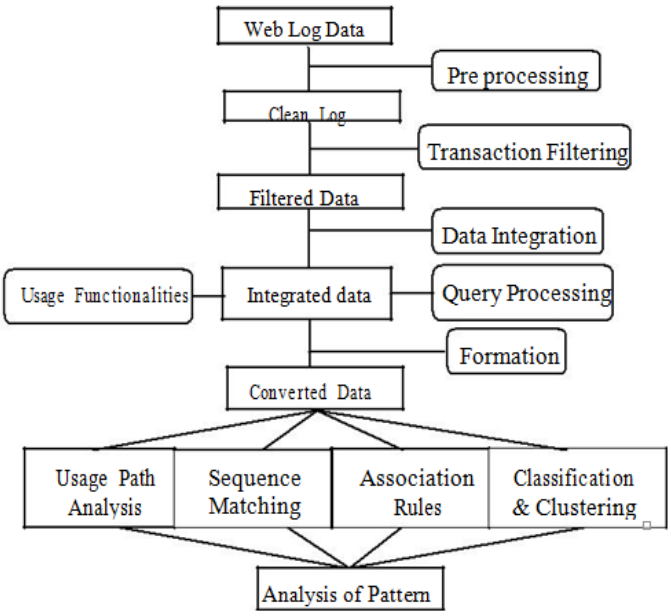

***Fig 1**: General Architecture of web Mining*

2.2 Information Gathering:

Web usage mining applications can gather data mainly from 3 sources [3]. (1)Web servers (2) Proxy servers (3) Web clients [3].The huge mass of data can be available from web servers which is the largest source of web data. In web servers, data is generally presented in Standard common Log Format, extended log format and LogML [3],[6]. For example **ECLF** (Extended common log format) is generally used in web servers.

Important terms in ECLF

Ipaddress- network address of user machine Rfc 931- remote login name of user
Status- as success / page not found like errors Authuser – original user name
User agent- software or browser (web client) Bytes – size of transferred information

2.3 Various Data Usage Issues:

- **Caching**: The data is stored on the server in cache hierarchy. It is possible to mismatch in the local cache data access patterns and web server log records E.g. user has visited page hierarchy as page 1, page2, page1, page3 but due to data in caching server has recorded log as page1, page2, page3 as second time access of page 1 would directly been from cache. So the second entry of page 1 is missed from log [5]. So, we cannot say that log that every time 100% correct data. Thus, caching is a very big issue for accessing web data.
- **CGI data**: CGI is referred as Common Gateway Interface and is used to pass variables and user entered data to respective server. It has a functionality to hide the username and value pairs from URI. So, the data is accessed by whom that cannot be tracked by usage mining methods.
- **Session identification**: Tracking and finding the session creation and usage duration when parallel login with same account through different machines makes identification complex.
- **Dynamicity of Pages**: Dynamic pages may change their content according to user request or fixed time interval. So, even minor change in content makes the log data huge as result.
- **Transaction uniqueness:** Issues in identifying unique users and their unique transactions as same account multiplicity is available.

III. ROLE OF NEURAL NETWORKS

3.1 Introducing ANN:

Artificial neural network (ANN) is motivated from biological nervous system which is a knowledge processing paradigm [1]. The main feature of this technology is its unique structure of system. ANN is the combination of highly connected processing elements, called neurons in medical science (brain), working together to solve a particular problem. It is used to configure to solve a particular problem such as, pattern reorganization or classification [1].Neural networks are highly capable to do the things such as meaning derivation from complex data. It is mostly used in finding usage

trends that are sometimes very common but complex and even not carried out by machines. Neural network has many advantages: Self adaptive learning- How to deal with problems based on initial training or experience from the network. Self organization-Artificial neural network creates its own organization and behavior and representation of knowledge it accepts while learning. Real time operations-Neural network computation may carry out parallel but for that special hardware are required to design. Fault tolerance via multiple information copies –partial destroys or failure of network cannot affect the performance of the network.

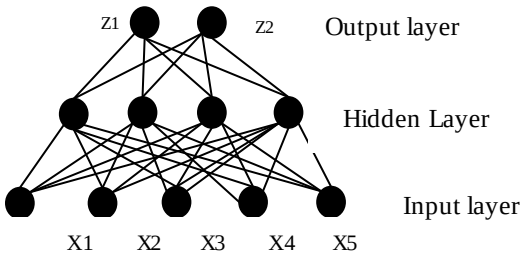

***Fig 2** multi layer recurrent neural network*

Neural networks basically work in three layers as input, hidden and output layer as shown in the figure 2. And multi layer recurrent network is there which is very useful for web usage mining especially for traffic analysis and distributed control.

IV. DIVERSE STAGE WISE ISSUES OF WEB USAGE MINING

From figure 1 we can see many stages of the process of web usage mining that is almost same as KDD process in Data Mining.

- **Log processing:** Logs are present on the web in ECLF format (Table 1). So, to extract information from huge logs for various purposes where each purpose having parameters is difficult.
- **Log Cleaning:** There may be many redundant entries in one server from the same client. e.g. A client may use same server with diff browser at the same time that may create replication in the log data. So, managing such situations is necessary and server logs must be cleared before further processing.
- **Recognition of User:** Once log is cleaned, it is moved forward to identify the user. There are many methods such as using cookie [7] or using Identd [protocol in RFC 1413] [7].
- **Session Tracking:** It is somewhat difficult to detect the users session because when one user is using multiple sessions which session is ending and which is going to start [2]. We can identify using time stamp value. But processing from a large log data and many sessions, it is very complex and difficult.
- **Usage Trend Analysis:** After processing everything there is a need to analyze what is the trend of user, or particular application.

V. SUGGESTIONS TO OVERCOME ISSUES IN WEB USAGE MINING

In general, web usage mining deals with unsupervised data where there are few issues as mentioned in section-4; to overcome these issues some of the work is as below:

4.1 Growing Natural Gas (GNG)

It is a self organizing map and is introduced in 1991 by Thomas maninetz and Klaus schulten and it is one type of artificial intelligent network. It is used for finding optimal representation using feature vectors [2]. The algorithm is named as “natural gas” because during alteration process, they distribute themselves like a gas in the whole network. No need to resolve number of nodes because the parameters of GNG are stable in time and as it is incremental [1]. GNG approach is very much useful in analysis like **user trend analysis.** It is one type of neural network which acquires the users by identifying the pattern of page accesses by users. We require to do some process on web log files to recognize users and their sessions too to get the outcomes. According to the session’s user, we have to train ANN [1]. Now select self organization method [1] because using this method doesn’t require supervising the training. The method of self organizing multilayered recurrent neural network is developed and used to train sample logical neural networks for web usage mining [1]. For heuristic self organization methods, concept of ANN is evolutionary. As a result heuristic self-organization methods depend upon the defined configuration of the Selection-criterion and the freedom of the candidate-structures selection, complexity of the synthesized neural network can’t be optimal [1].

4.2 ART Model Mechanisms:

ART approach does web log analysis via introducing ART structure for huge, widely distributed, highly

heterogeneous, partly structured, interconnected and growing hyper text information warehouse of WWW [3]. This approach has two sub systems, (1) attention subsystem (2) orienting subsystem. In attention sub system stabilization of learning and activation occur. In this method, there is bottom-up input activation and top down expectation [3]. Orienting subsystem used to handle the mismatch occurring in attention sub system.

4.2.1 ***Properties of ART:*** ART model has very crucial four properties. (1) Self scaling computation units (2) self adjusting memory search (3) Previous patterns indirectly access their respective category (4) The System behaves as a teacher and according to environment, it changes its vigilance.

4.2.2 ***Use of ART in Web Usage Mining:*** The ART model was anticipated for unsupervised clustering of binary data [3]. It has one layer neural network in its attention subsystem. In the process of ART, it has fixed no of input neurons to understand the no of dimensions and no of output neurons to map with same amount of utmost clusters. Initially output neurons are not assigned. Once when output neuron is trained from a pattern, it will become assigned. The activation function is calculated from all assigned output neurons. The input is connected by both top-down, and output by bottom-up weights [3]. The main steps of this approach is as follows: (1) Web log data collection (2) data pre-processing (3) clustering unsupervised data (4) Web usage mining after above steps [3].

4.3 Learning Vector Quantization:

4.3.1 ***LVQ*:** visualize that a number of 'codebook vectors' cvi (free bound vectors) are situated into the input gap to estimate a variety of domains of the input vector v by their quantized values [8]. Generally a number of codebook vectors are assigned to each class of v values, and v is then determined to fit in to the same class to which the adjacent cvi belongs. Let

$$\boldsymbol{d}=\mathbf{argmi}(\|\boldsymbol{v}-\boldsymbol{cvi}\|) \qquad (1)$$

Eq. describes nearest cvi to v, denoted by dc.

Standards for cvi that just about reduce the misclassification errors in the over nearest-neighbour classification can be found as asymptotic values in the subsequent learning process. Let v(t) be a sample of input and let the cvi(t) stand for sequences of the cvi in the discrete-time domain. Begin with appropriately defined initial values, the following equations define LVQ1.

$$\boldsymbol{dc\ t}+\mathbf{1}=\boldsymbol{dc\ t}+\boldsymbol{alpha\ t}\,[\boldsymbol{v\ t}-(\boldsymbol{t})] \qquad (2)$$

If v and dc fit in to the similar class,

$$\boldsymbol{dc\ t}+\mathbf{1}=\boldsymbol{dc\ t}-\boldsymbol{alpha\ t}\,[\boldsymbol{v\ t}-(\boldsymbol{t})] \qquad (3)$$

If x and mc belong to dissimilar classes, $\boldsymbol{cvi\ t}+\mathbf{1}=(\boldsymbol{t})$ **,** for I is not in c. Here 0 < alpha (t) < 1, and alpha (t) may be steady or it decrease monotonically with time. In the above basic LVQ1 it is suggested that alpha should be lesser than 0.1; in time and is used in this package.

4.3.2 ***LVQ3*:** The LVQ2 algorithm was base on the thought of differentially changing the resolution limitations towards the Bayes restrictions, while no notice was paid to what might has done to the location of the min the extended run if this process were continued. As a result it seems like required having corrections that make sure that the cvi continue resembling the class distributions, at least more or less. By combining those ideas, an improved algorithm is acquired that is named as LVQ3 [9]. As per Equation 2,

$$\boldsymbol{mk\ t}+\mathbf{1}=\boldsymbol{mk\ t}+\boldsymbol{epsilon\ al}(\boldsymbol{t})[\boldsymbol{v\ t}-\boldsymbol{mk}(\boldsymbol{t})] \qquad (4)$$

4.3.3 ***OLVQ*:** The basic LVQ1 algorithm is now adapted in such a way that an individual learning rate alpha (t) is given to each cvi [7]. So we get the distinct time learning process.

Let c be defined by Eq. (1). Then,

$$\boldsymbol{dc\ t}+\mathbf{1}=\boldsymbol{dc\ t}+\boldsymbol{alpha\ t}\,[\boldsymbol{v\ t}-\boldsymbol{d}(\boldsymbol{t})] \qquad (5)$$

If v is classified correctly,

$$\boldsymbol{dc\ t}+\mathbf{1}=\boldsymbol{dc\ t}-\boldsymbol{alpha\ t}\,[\boldsymbol{v\ t}-(\boldsymbol{t})] \qquad (6)$$

If the classification of v is incorrect,

$(t + 1) = cvi(t)$, for i not in c.

We tackle the problem irrespective of the $alphai\ (t)$ can be firmed optimally for best possible union of eq. (6) [7]. If we state (6) in the form,

$$dc\,t + 1 = 1 - st\ alphac\ t\ dc\,t + s\ t\ alphac\ t\ v\ t \qquad (7)$$

Where $(t) = +1$ or -1 if the classification is accurate and incorrect respectively, we primary straight see that (t) is statistically liberated from $v(t)$. It may also be noticeable that the numerical correctness of the scholarly codebook vector values is best if the assets of the corrections made at different times, to the end of the learning period, are of the same weight. Observe that $dc(t + 1)$ contains a "trace" from $v(t)$ through the last term in (7), and "traces" from the earlier $v(t')$; $t' = 1,2,...t-1$ through $dc(t)$. The magnitude of the last "trace" from (t) is scaled down by the factor $alph\ (t)$, and, for illustration, the "outline" from $x(t - 1)$ is scaled down by $[1-(t)alphac(t)]\ alphac(t - 1)$ [7].

4.3.4 ***MLVQ****:* It is like supervised version of multi pass SOM where fast rough pass can be made on the model by using OLVQ1 algorithm and then the long very well tuning pass can be made on the model through any of LVQ1, LVQ 2 or LVQ3 [11].

4.3.5 ***HLVQ****:* Here an LVQ model is constructed and each codebook vector is treating as a cluster centroid. All codebook vectors are evaluate and selected as candidates for sub-models. As sub models are constructed for all candidate codebook vectors and those sub models that can do better than their parent codebook vector are kept as a part of the model [10]. During testing a data instance is first mapped onto its BMU, if that BMU has a sub-model, then that sub model is used for classification, else the class value is used for classification. To better model the data, this algorithm has verified most useful for large datasets where a low-complexity model is used as the base model, and dedicated LVQ models are used at each codebook vector [10].

Table 1: *Comparison of various Approaches*

Fields	LVQ	OLVQ	MLVQ	HLVQ
Version	Basic	Optimized	Multiple Passes	Hierarchical
Accuracy	Low	More than LVQ	More than OLVQ	More than MLVQ
Time	High	Less than LVQ	Less than OLVQ	Less than MLVQ
Efficiency	Low	Medium	High	High
Codebooks	More	Moderate	Less	Less
Data	Low	Medium	High	High

VI. EXPECTED OUTCOMES

Using GNG, the possible result makes the network intelligent and stable for other users. Using ART, the best results can be achieved, specially its feature of dynamic vigilance parameter and bottom up input action and top down expectation approach [8] .Network is able to handle large set of web data and optimal representation of the network is possible which is very useful for user trend analysis and usage path frequency analysis and popularity of web clients. Huge amount of web logs can easily been classified using ART model. ART model can classify and cluster any type of complex log data on the basis of specific analysis such pattern identification and session tracking. GNG and ART will be very useful in training of neural network. Once the neural network is trained, it will classify the data at the most accuracy and speed.

V. EXPERIMENTAL RESULT

Medical diagnosis is gaining increasing importance; so there is a certain need to introduce computational intelligence techniques into the traditional bio-medical domain. This makes accurate decision in the least possible time. The experimental result is taken by the use of hierarchical LVQ network for the diagnosis of breast cancer. It uses of two LVQ networks, one after the other, for diagnosis. On the base of the experimental remarks of different neural network, HLVQ gives the maximum accuracy of 98.14% for diagnosis of breast cancer. These techniques may be used for other problems like blood pressure identification, heart attack prediction, and fetal health prediction, blood follow control between pre-natal fetus and maternity abdomen.

5.1 Analysis of First Stage of HLVQ

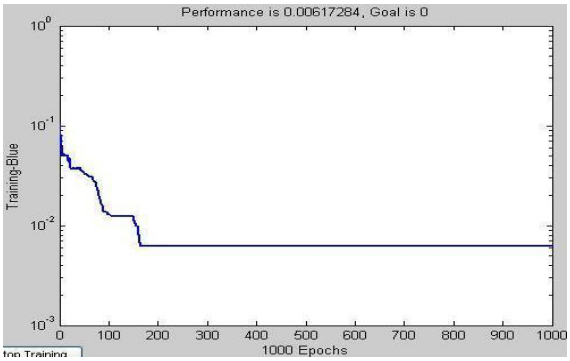

Fig.3. *Result after first stage of HLVQ [11].*

5.2 Analysis of Second Stage of HLVQ

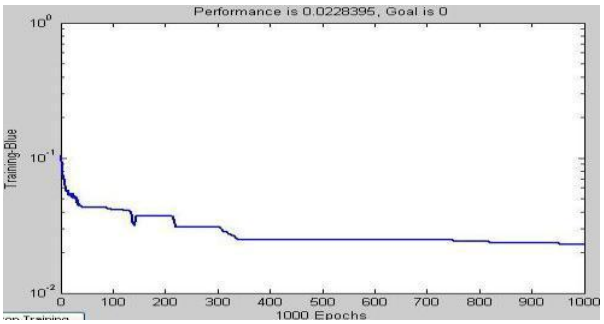

Fig.4. *Result after second stage of HLVQ [11].*

V. CONCLUSION

After the survey on web usage mining, for all the mentioned issues and parameters, because of its adaptive learning nature it is confirmed that the approach of neural network can be very useful. So, as growing the web data, neural network approach will play an important role in today and upcoming days in knowledge extraction from web-logs. Finally, learning and comparing a lot, to handle the large volume of web data, there would be a definite need of neural network concept, because only through neural network learning algorithms, such huge volume of web data can be handled and applied to any application for knowledge extraction. LVQ is a supervised model of SOM and used for giving class label to data. LVQ is capable to summarize large datasets to lesser number of codebook vectors that are useful for visualization and classification. The basic HLVQ algorithm works on the same approach in all hierarchy generation while HHLVQ uses more than one approach to classify the data, so the disadvantages of the same method will excluded and only merits of both approaches are enlighten for accurate classification. Hence, HHLVQ will be proved as a great classification technique and will be useful for huge datasets as well.

REFERENCES

[1] Sonali muddalwar, Shashank Kawan, “Applying artificial neural networks in web usage mining”, international journal of computer science and management research, vol 1 issue 4 [Nov-12]
[2] Anshuman Sharma, “Web usage mining Neural network”, international journal of reviews incomputing, vol 9, [10th april, 2012].
[3] Valishali A. Zilpe, Dr. Mohammad Atique, “Neural network apprach for web usage mining”,ETCSIT, published in IJCA [2011]
[4] Farhad F. Yusifov, “web traffic mining using neural networks”, world acedamy of science,Engineering and technology, 21, [2008]
[5] Jaydeep Srivastava, “Web Mining: Accomplishments and future directions”,http://www.cs.unm.edu/faculty/srivastava.html
[6] John R. Punin, Mukkai S. Krishnamoorthy, Mohammed J. Zaki, “Web usage mining- language and algorithms”, rensseluer polytechnic institute, troy NY 12180.

[7] R Baraglia "Suggest:A web usage mining system,"in proceedings of IEEE International conference on Information technology:coding and computimg,April 2002.
[8] Renata M. C. R. de Souza, Telmo de M. Silva Filho," Optimized Learning Vector Quantization Classifier with an Adaptive Euclidean Distance", 19th International Conference, Limassol, Cyprus, September 14-17[2009], volume 5768.
[9] Diamantini, Claudia , Spalvieri, A. "Certain facts about Kohonen's LVQ1 algorithm", Circuits and Systems, IEEE[1994].
[10] Sang-Woon Kimy and B. J. Oommenz, "Enhancing Prototype Reduction Schemes with LVQ3-Type Algorithms", Natural Sciences and Engineering Research Council of Canada, and Myongji University, Korea, kimsw@mju.ac.kr, oommen@scs.carleton.ca.
[11] R. R. Janghel, Ritu Tiwari, Anupam Shukla, "Breast Cancer Diagnostic System using Hierarchical Learning Vector Quantization", IJCA Proceedings on National Seminar on Application of Artificial Intelligence in Life Sciences[2013].
[12] Mahesh kumar, Uday Kumar," Classification of Parkinson's disease using Lvq, Logistic Model Tree, K-star for Audioset", Hogskolan Darlana University, 2011, roda wagen 3s-781 88.
[13] Suitable MLP Network Activation Functions for Breast Cancer and Thyroid Disease Detection, Second International Conference on Computational Intelligence, Modeling and Simulation, 2010, IEEE, http://eprints.usm.my/20282/1/4262a039.pdf.
[14] Ravita Mishra, "Web Usage Mining Contextual Factor: Human Information Behavior", International Journal of Information Technology and Management Information Systems (IJITMIS), Volume 5, Issue 1, [2014].
[15] M.P.Yadav,"An Efficient Web Mining Algorithm for Web log Analysis:E-Webminer" IEEE[2012].

