ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

Comparison between Trinity Unsupervised Data Extraction and Data Extraction Using Artificial Neural Network

Priyanka Thomas, Garima Singh

P.G. Student, Department of Computer Science, WCEM, Dongargaon, Nagpur, India

Associate Professor, Department of Computer Science, WCEM, Dongargaon , Nagpur, India

ABSTRACT: - In this project we present Trinity Tree Algorithm comparison with Back Propagation Algorithm. Among these the trinity tree algorithm is an unsupervised data extraction and Backpropagation algorithm is a supervised data extraction. Data mining is a growing topic of interest in latest Engineering subject as it has help in the research area to extract important information from raw data. Data mining deals with supervised data as well as unsupervised data. Our first method Backpropagation algorithm is being used through artificial neural networks which require a dataset of the desired output for many inputs, making up the training set. It is most useful for feed-forward networks (networks that have no feedback, or simply, that have no connections that loop). Our second method is Trinity is a unsupervised proposal to generate patterns and to separate them by prefix, suffix and separator in order to generate a regular expression. Aim of our proposed technique is to improve the quality of time in performance. The accuracy of result to be much more and with less error. The graph gives the total clustering of two methods.

KEYWORDS : Data mining, Backpropagation algorithm, KDD cup dataset, Artificial neural network, SOM maps

I. INTRODUCTION

Nowadays the usage of web resources and data are growing exponentially as numbers of users are increasing every day. Growing data storage and usage make the network (WWW) complex and which cannot be handled in future. So, few efficient techniques of web usage mining are useful for retrieving knowledge from huge data available on the web. Web mining basically classifies in 3 major categories content mining, structure mining and usage mining. These techniques are used depending upon what to mine from the web. Now, focus on web usage mining that helps to deal with certain web scaling problems such as user trend analysis of surfing, traffic flow analysis, distributed control handling, web traffic management and many more. Session tracking and website reorganization, distributed traffic sharing on distributed servers can be identified and analysis based on web data can be possible using concepts of neural network. 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, adaptive web sites and website reorganization.

The web provides a huge amount of data which are used in various forms like text, numbers, video, audio, graphs etc which are acting as raw data. Now data mining provides better data which can be used in making better decision. Below example show how data mining can be helpful for human beings. Today various technologies are used in daily life which help us in many ways like for various diseases and treatment in medical area require technology like keeping the records of data of patients, disease status, whether requires further medical attention, dates of admittance and discharge etc .Now this records may have data in large amount and suppose a patient of accident case was admitted. If the relatives of the patient came to hospital than they require the relevant information of patient with his name, disease etc. Now it is not necessary that a person have his own unique name i.e many patient with same name might be present in same hospital so for such situation data mining helps in getting the exact required data in case of emergency. The Data Extraction through data mining summarise the data and pattern discovery from raw data.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

1. ARTIFICIAL NEURAL NETWORK

An artificial neural network is model based on human brain with its neural structure which comprises neurons, dendrites, axon .The neuron here act as a node defined with various features in neural network. The dendrites which help in passing the signal from body to brain is done in human body while in network signal is passed by connectivity between nodes in network. Below diagram shows both biological and artificial neural network. Neural network is a natural intelligent system discovers the data which is required or on which dependency is found than it discovers the patterns in much faster way.

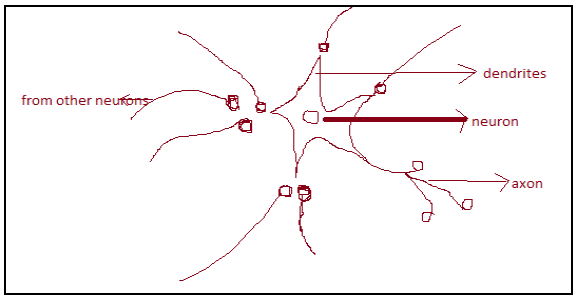

Fig 1(a) Biological neural structure

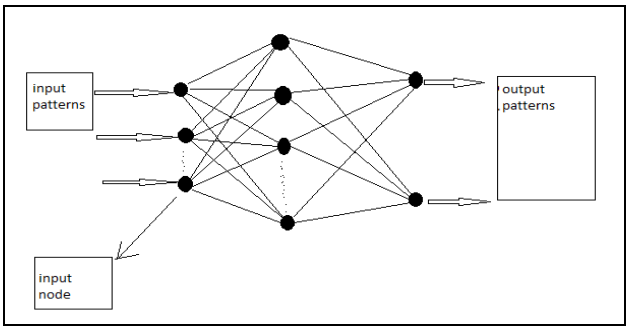

Fig 1(b) Artificial Neural Network

II. RELATED WORK

2. ARTIFICIAL NEURAL NETWORK MINING

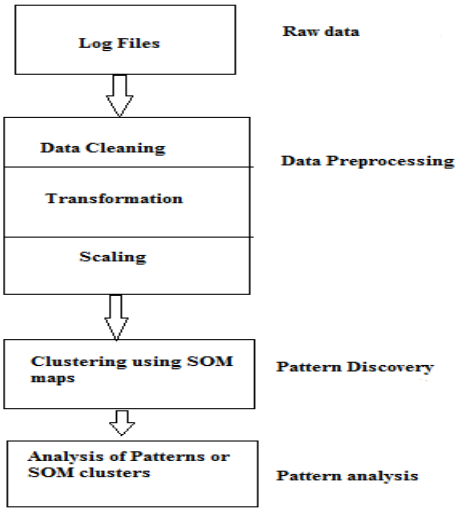

FIG 2.NEURAL NETWORK ANALYSIS

Preprocessing- Data have to be collected from different places and after identifying users to split each user in sessions. The data is sampled on the basis of object identification, feature extraction and Normalization.

Transformation- The input and output transaction formats match so that any number of modules to be combined in any order, as the data analyst sees fit.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

Pattern Recognition-Pattern discovery takes place on the pre-processed data. This pattern discovered by frequent pattern mining.

Clustering –Clustering of web usage log where the clustering is done with the aim of matching groups of users which are having same interest of browsing information clustering uses SOM maps is a visualization tool for visualizing high dimensional , complex data with inherent relationships between the various features of the data.

Pattern Analysis- It is last step which deals with the pattern discovered earlier which are under analysis by using filters which help to overcome the complexity of each pattern.

Backpropagation algorithm:
It helps in performing task in networking and are used in feed forward artificial neural network. This algorithm deals with supervised learning and it send data to nodes which h are associated in network .The data is send in the form of signal which forward pass i.e if pattern matches then it moves in forward direction whereas the error or the mismatched data is returned back i.e backward pass. This helps in reducing the error.
Actual algorithm for a 3-layer network (only one hidden layer):
1. Initialize the weights in the network (often randomly)
2. repeat
For each example e in the training set do
1.O = neural-net-output(network, e) ; .forward pass
2..T = teacher output for e
3..Calculate error (T - O) at the output units
4.Compute delta_wh for all weights from hidden layer to output layer ;
backward pass
5.Compute delta_wi for all weights from input layer to hidden layer ; backward pass continued
6. Update the weights in the network
3. Until all examples classified correctly or stopping criterion satisfied
4.Return the network
It also has two phases of propogation and weight updation till the whole data is fully processed.

III.PROPOSED LOGIC

The objective is to provide an acceptable solution at low cost by seeking for an approximate solution to problems. Soft computing methodologies (involving, neural networks, genetic algorithms and rough sets) hold promise in Web mining. The proposed approach includes Web-log analysis via introducing Back propogation structure for huge distributive information repository of World Wide Web. Back propagation architecture models can self-organize in real time producing stable recognition while getting input patterns beyond those originally stored.

The general architecture of web usage mining is described here. The WEBMINER is a system that implements parts of this general architecture. The architecture divides the Web usage mining process into two main parts. The first part includes the domain dependent processes of transforming the Web data into suitable transaction form. This includes preprocessing, transaction identification, and data integration components. The second part includes the largely domain independent application of generic data mining and pattern matching techniques (such as the discovery of association rule and sequential patterns) as part of the system's data mining engine.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

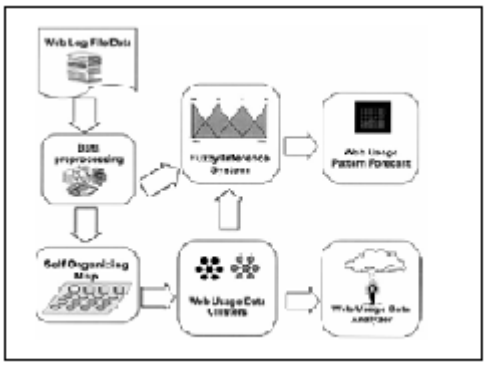

Fig3(a) Neuro-Fuzzy Approach for Web Mining

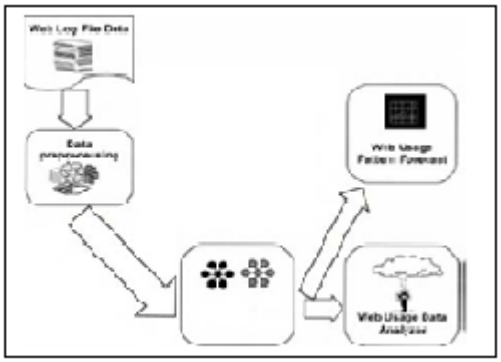

Fig3(b) Reduction of Stages on Neuro-Fuzzy after Back Propagation implementation

IV. EXPERIMENTAL RESULTS

The Back Propagation introduced are deemed as being highly effective as a sophisticated visualization tool for visualizing high dimensional, complex data with inherent relationships between the various features comprising the data. The Back Propagation output emphasizes the salient features of the data and subsequently leads to the automatic formation of clusters of similar data items. This particular characteristic of Back Propagation alone qualifies them as a potential candidate for data mining tasks that involve classification and clustering of data items.

A 'learnt' Back Propagation can be used as an important visualization aid as it gives a complete picture of the data; similar data items are automatically grouped together.

The Back Propagation has proven to be one of the most powerful algorithms in data visualization and exploration against Trinity tree. Application areas include various fields of science and technology, e.g., complex networking, telecommunications systems, databases, and even financial applications. The Back Propagation the high- dimensional input vectors onto a two-dimensional grid of prototype vectors and orders them. For a human interpreter, the ordered prototype vectors are easier to visualize and explore than the original data. The SOM has been widely implemented in various software tools and library. As shown in fig 3(b), Post-processing the SOM extracts qualitative or quantitative information of the data.

Analysing the outcome of the algorithms, we can conclude that with respect Trinity tree we can cover more URL but Back Propagation works better for larger number of cases. With increase in data, learning process of Back Propagation becomes more accurate and we can consider larger number of clusters. Back Propagation is also efficient in time as compared to Trinity tree. Thus we can conclude that Back Propagation has better performance than Trinity tree. For more accurate result we can provide output of Trinity tree as a input to Back Propagation and use the results of Back Propagation for generating recommendation System and analysis of the Web site.

We have analysed our Red Colour Trinity Tree .Blue Colour Back Propagation Algorithm using KDD Dataset for Networking.

Figure shows the results of supervised learning through data extraction using backpropagation algorithm and unsupervised data extraction through trinary tree.
Red Color = Trinary Tree
Blue Color = Back Propagation

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

Fig 4(a) shows the detailed accuracy by clustering method using SOM maps which means with cluster how many attack trinary tree has detected and how many back propagation has detected.

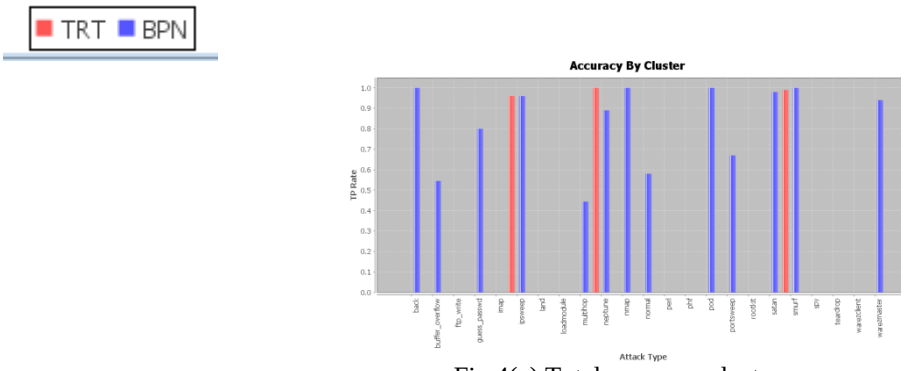

Fig 4(a) Total accuracy cluster

Fig4(b) shows statistical analysis of accuracy of the two methods of trinary tree and backpropagation algorithm means accuracy of attack detection in percentage from trinary tree and back propagation.

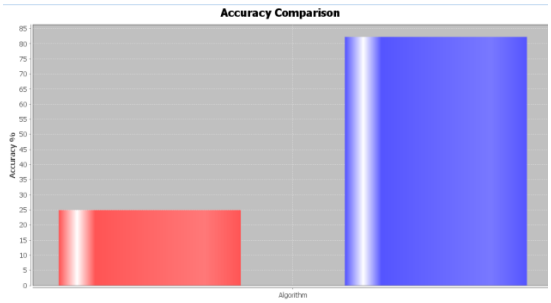

Fig 4(b) Accuracy Comparison

. Fig 4(c) shows the time comparison between the two methods and which used much time and which took less time means how much time is taken by trinary tree to detect attack and how much time is taken by Back Propagation .

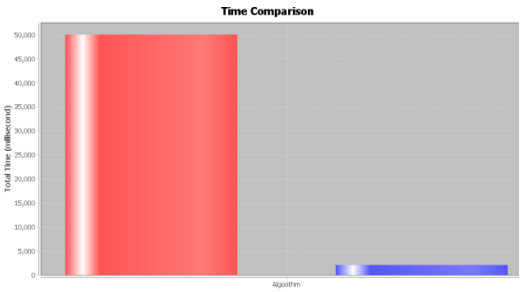

Fig 4(c) Time Comparison

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2015

IV. CONCLUSION

Previously the concept of Trinity which is an unsupervised proposal which extract data on the basis of extraction rule which were developed on same server side template .It works on the basis of hypothesis that shared pattern do not provide relevant data . Now our proposal of neural networks by using backpropagation algorithm is a supervised learning and deals with KDD cup dataset as input and further process of research is done on this basis so that mining could take place building the required pattern. We relax this assumption by exploiting reference sets of KDD dataset to undergoing normalization, transformation and scaling and then creating its required pattern. The main ability of neural network is to learn from its environment and to improve its performance in case of time, accuracy.

REFERENCES

[1] Becker, S & Plumbley, M (1996). Unsupervised neural network learning procedures for feature extraction and classification. International Journal of Applied Intelligence, 6, 185-203
[2] C.-H. Chang, M. Kayed, M. R. Girgis, and K. F. Shaalan, "A survey of web information extraction systems," IEEE Trans. Knowl. Data Eng., vol. 18, no. 10, pp. 1411–1428, Oct. 2006
[3] Artificial neural networks for pattern recognition B YEGNANARAYANA Scidhanci, Vol. 19, P a r t 2, April 1994, pp. 189-238.
[4] Trinity: On Using Trinary Trees for Unsupervised Web Data Extraction Hassan A. Sleiman and Rafael Corchuelo VOL. 26, NO. 6, JUNE 2014.
[5] V. Crescenzi and G. Mecca, "Automatic information extraction from large websites," J. ACM, vol. 51, no. 5, pp. 731–779, Sept. 2004
[6] C.-H. Chang and S.-C. Lui, "IEPAD: Information extraction based on pattern discovery," in Proc. 10th Int. Conf. WWW, Hong Kong, China,2001, pp. 681–688.