FRAMEWORK FOR COMPETENT DATA ABSTRACTION USING AI

Anila M
Research Scholar, Department of Computer Science & Engineering
KL University, Vaddeswaram, India

Abstract

In the process of clustering or unsupervised learning the main objective is to identify the instances with similar kind of properties (called clusters) within the data set. In this method, we will not have any information regarding the class label of the data or how many classes are present. Extraction of information from unstructured, ungrammatical data like classified listings is quite difficult because of the facts that traditional structural and grammatical extraction methods do not apply. In supervised classification, the task is to automatically instigate a model based on a set of N instances, called training data. This model then will be used to assign labels for new instances with unknown labels using only the value of their predictor variables. Previous work has made the good use of reference sets to aid such extraction, but it did so using unsupervised machine learning. In this paper, we present a supervised approach that do two things i.e. selecting the relevant reference set(s) automatically and then uses it for supervised extraction. We here are using Artificial Intelligence and emulate it with unsupervised method. In this project, we present Trinity Tree Algorithm comparison with Back Propagation Algorithm.

Index Terms: Data mining, web mining, Classification, unsupervised learning, supervised learning, Artificial neural network, Back propagation algorithm, SOM map

1. Introduction

Web data usage is increasing in day to day life and the readiness of essential data from huge web depository is leading following methods. Web mining basically classifies in three major categories content mining, structure mining and usage mining. Usage of these techniques depends upon what to mine from the web. At present, we emphasize on web usage mining that deals with some web scaling problems like user trend analysis of surfing, distributed control handling, traffic flow analysis, web traffic management. etc., Session tracking and website reorganization, distributed traffic sharing on distributed servers can be branded and analysation of web data is possible using neural network concepts. By this, we can say that data mining techniques can be applied on web logs, server log files result in useful usage path extraction, session tracking, session duration, number of session creations, adaptive web sites and website reorganization. Neural network is different from static networks in which each node is self-intelligent, so the network becomes intelligent and use this network long usage. This concept useful for web usage mining in extracting information for web traffic analysis on live servers and frequent usage path analysis and many more concepts. This network endures classification that make models predict categorical class label. This step is called as learning phase.

In the next step, the classification algorithms figure a classifier, from the training set which is made up of database tuples and their related class labels. Each tuple that institute the training class is referred as class. The aim of unsupervised learning or clustering is to discover groups of similar instances within the data. In this approach, we have no information about the class label of data or how many classes are present.

The graphical representation of the clustering looks like a tree structure which is known as 'Trinity tree'. The major task in supervised learning is to automatically persuade a model based on a set of N instances, called training data. This model is used to assign labels for new instances with unknown labels using the value of their predictor variables. The artificial neural network is based on simulating the structure and behaviour of the biological neuronal networks.

1.1 Artificial Neural Network

Artificial Neural Network is a system that is loosely modelled based on the human brain. This field is called with many names, such as connectionism, natural intelligent systems, parallel distributed processing, neuro-computing, machine learning algorithms. It has ability to account for any functional dependency It hardly doesn't require any prompting as the network learns and models the nature of the dependency. We won't be using the application of data mining because of its long training time, complex structure and poor interoperability. Concept of neural networks proves it as strong technique to solve many real-world problems. Neural networks have the ability to learn from experience in order to improve their performance and to adapt themselves to changes in the environment. They are also able to deal with any incomplete information or noisy data and can be very effective.

1.2 Experimental Detail

Data Mining

Information mining methods can be executed quickly on existing programming and equipment stages to upgrade the benefit of existing data assets, and can be coordinated with new items and frameworks as they are brought on-line. At the point when represented on client/server or parallel processing PCs, data mining instruments can dissect gigantic databases to convey answers to inquiries, for example, "Which customers are likely to respond to my next promotional mailing, and why?" As shown in Fig1, Data mining process consist of three major phases: -

1. Data pre-processing.
2. Applying data mining techniques.
3. Interpretation of Results

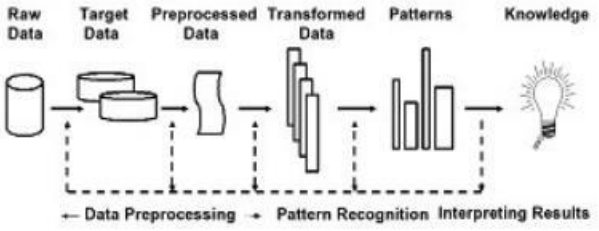
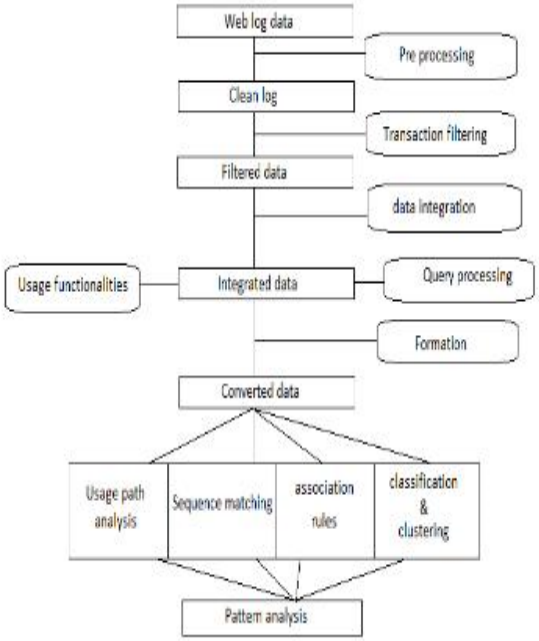


Fig 1.Data Mining Processing



As indicated by the figure2 the handling steps, different methods are just like the procedure of data mining. Fundamental contrast in data mining and web mining is that the underlying level in information originates from different data bases and warehouses in web mining, information originates from server log records.

2. Proposed System

The objective is to impart an adequate solution at low cost by seeking for an approximate solution to problems. Soft computing methodologies (that include neural networks, genetic algorithms and rough sets) hold promise in Web mining.

The proposed approach incorporates Web-log investigation by means of back propagation structure for vast, exceedingly heterogeneous, semi organized, interconnected and hypertext data repository of World Wide Web. Back propagation design models can self-compose continuously delivering stable acknowledgment while getting input designs past those initially put away.

The WEBMINER is a framework that actualizes parts of the general architecture. The design partitions the Web usage mining process into two fundamental parts. The initial segment incorporates the domain dependent processes of changing the Web data into appropriate form. It likewise incorporates pre-processing, transaction identification, and data integration components. The later part incorporates domain independent application of generic data mining and pattern matching techniques as a major aspect of the framework's data mining engine. The general design for the Web mining procedure is portrayed in Figure 2

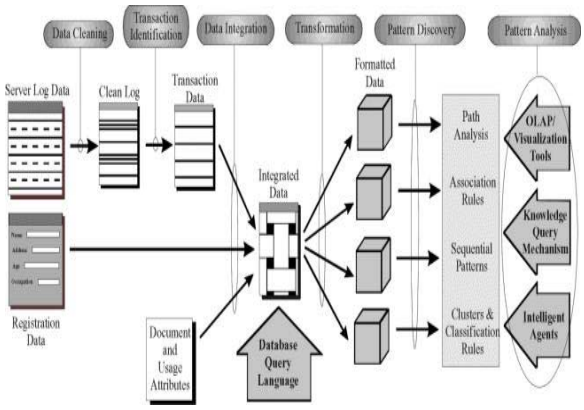


Fig 2. Architecture of web usage mining

1. Data cleaning: It is the initial step performed in the Web usage mining process. Some low-level data integration assignments may likewise be performed at this stage, for example, consolidating various logs, incorporating referrer logs, and so on.

2. Transaction identification: After the data cleaning, the log sections must be divided into logical clusters using one or a progression of transaction identification modules. The objective of this module is to make vital clusters of references for every client. The errand of distinguishing exchanges is one of either isolating a vast exchange into various littler ones or blending little exchanges into less bigger ones.

3. Transformation: The input and output transaction coordinate so that any number of modules can be consolidated in any order, as the data expert sees that it fits in.

4. Pattern discovery: Once the domain-dependent data change stage is finished, the subsequent data must be designed to fit in with the data model of the suitable data mining task. For example, the arrangement of the data for the association rule discovery undertaking might be unique in relation to the organization essential for mining sequential patterns.

5. Pattern analysis: Finally, a query mechanism will permit the client (expert) to give more control over the discovery process by indicating different requirements.

1. Self-scaling computational units. The subsystem depends on focused getting the hang of improving pattern features that however includes stifling commotion.

Pattern discovery is done in the following way:
3. Self-adjusting memory search: The system can look memory in parallel, and change its order of search.

4.Already learned patterns get access to their comparing classification.

4. If the environment disapproves the current recognition of the system, it changes this parameter to be more cautious.

Three major steps of this approach can be integrated as follows: -

a. Web-log data accumulation: The logs we research are of W3C Extended Log File Format under KDD Cup dataset. Web log systems administration or restorative information is gathered from the server of site for the time of one month for trial reason.

b. Data pre-processing: We can utilize database programming Access and Java programming dialect to execute the pre-processing work. Additionally, web-log record pre-processing devices, for example, WEBMINER, AWStat can be utilized for data cleaning, user identification and path completion.

c. Web–usage mining from web-log files: The last stride of web-use mining can be executed utilizing neural system approach by means of. Back Propagation calculation.

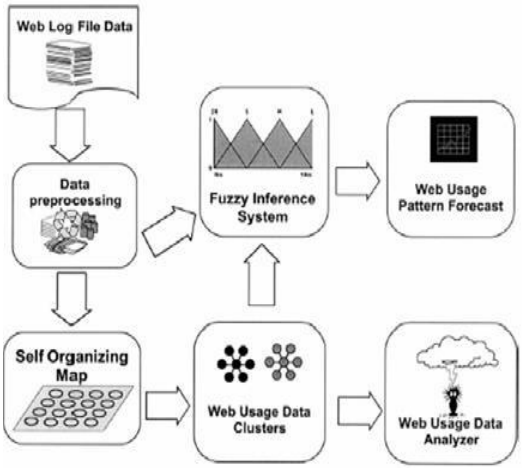


Fig 3: Neuro-Fuzzy Approach for Web Mining

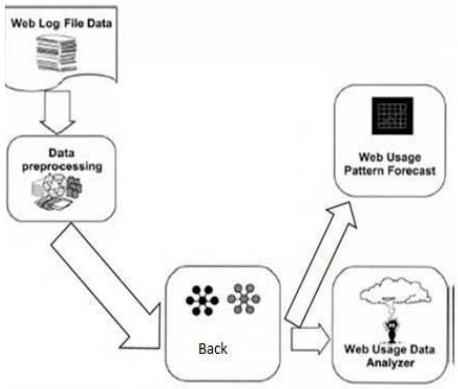


Fig 4: Reduction of phases on Neuro-Fuzzy after Back Propagation implementation.

If any Web-mining researchers apply this Back propagation, they can easily obtain best results than any traditional Web mining techniques with the usage of vigilance parameter, top-down and bottom-up weights.

Also using Back Propagation, it is more beneficiary to minimize the number of steps in Web mining as compare to neuro-fuzzy approach. As neuro-fuzzy approach uses five major steps to produce the Web usage pattern forecast, and Web-usage data analyser; named Web-log data collection, data pre-processing, self-organizing map, Web-usage data cluster, and fuzzy inference system (Figure. 3). But Back propagations use only three steps as Web-log data collection, data pre-processing, and Back propagations itself (Figure 4).

3. Clustering using SOM

The self-arranging maps (SOM) presented are considered as being very powerful as a classy visualization tool for picturing high dimensional, complex information with inalienable connections between the different components involving the data. The SOM‟s yields the remarkable components of the data and accordingly prompts the automatic formation of clusters of comparative data items. This representative of SOMs alone qualifies them as a potential contender for data mining errands that include classification and clustering of data items.

A "learnt‟ SOM can be utilized as a visualization aid as it gives a total picture of the data; comparable data items are naturally gathered together.

The Self-Organizing Map (SOM) has ended up being a standout amongst the most capable calculations in data visualization and exploration. Application areas incorporate different fields of science and innovation, e.g., complex modern procedures, telecommunications systems, document and image databases, and even budgetary applications. The SOM maps the high-dimensional input vectors onto a two-dimensional grid of prototype vectors and requests them. For a human translator, the ordered prototype vectors are simpler to imagine and investigate than the first information. The SOM has been widely implemented in various software tools and libraries as shown in fig3, Post-processing the SOM extracts qualitative or quantitative information of the data.

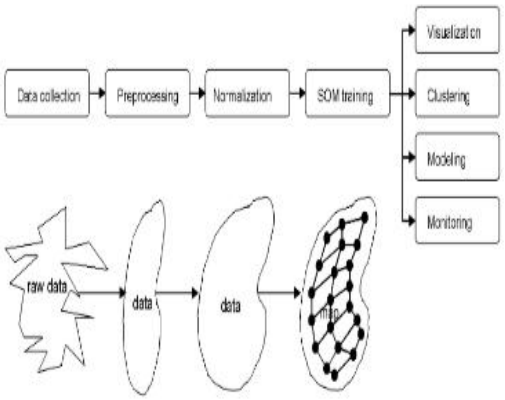


Fig .5 Applying SOM in Data Mining

Table.1. Comparison with respect to SSE with different # clusters and # cases of K-Means and SOM

ISSN: 2321-8134

CLUSTER #	K-MEANS (SSE)	SOM (SSE)
3	7017	321
4	5263	243
5	4210	195
6	3508	162

K-Means cover more URLs but SOM works better for larger number of cases. With increase in data, learning process of SOM becomes more accurate and we can consider larger number of clusters. SOM is also efficient in time as compared to K-Means. Thus, we can conclude that SOM has better performance than K-Mean.

4. CONCLUSION

It can be concluded that supervised learning is more effective than unsupervised. Most of the previous works and discussions conclude that, unsupervised extraction used extraction patterns that make supposition about the harmony of the structure in the data. We relax this assumption by utilising reference sets to assist the extraction. SOM used for clustering is rapid and accurate helping us further in artificial neural network mining that analyse pattern defined in the training set and then can be compared with many unorganised testing set. The comparison will go under the process of pre-processing, classification, clustering and analysing.

REFERENCES

[1] Trinity: On Using Trinary Trees for Unsupervised Web Data Extraction Hassan A. Sleiman and Rafael Corchuelo VOL. 26, NO. 6, JUNE 2014.

[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] Becker, S & Plumbley, M (1996). Unsupervised neural network learning procedures for feature extraction and classification. International Journal of Applied Intelligence, 6, 185-203.

[5] V.Crescenzi and G. Mecca, "Automatic information extraction from large websites," J. ACM, vol. 51, no. 5, pp. 731–779, Sept. 2004.