EFFICIENT DATA PRE-PROCESSING FOR DATA MINING USING NEURAL NETWORKS

JothiKumar.R[1], Sivabalan.R.V[2]
[1]Research scholar, Noorul Islam University, Nagercoil, India
Assistant Professor, Adhiparasakthi College of Engineering, G.B. Nagar, Kalvai, India
[2]Professor, Noorul Islam University, Nagercoil, India

ABSTRACT

Organizations are maintaining history of data for future analysis. These huge volume of database is analysed to Predict and improve the benefits and profits of the organization and also for the development. By analysing the history of data, strategic decisions can be made to improve the performance of the organizations by the top level peoples. So organizations are interested in analysing the data which will result in valuable insight. The data subjected to mining consists of inconsistent, blank or null and noisy values which have to be cleaned before mining. Usually the Techniques of Mean, Mode, and Median will be used to clean the data which are inefficient methods. Here I am representing the efficient data pre-processing which is to be carried out before actual mining process can be performed. The data from different databases, different locations and different formats are considered for pre-processing. This results in identification of reasonable patterns to improve the performance of organization. Even the Neural Networks has complex structure, consumes more learning time and difficult to understand the representation of results, it have more acceptance ability to clean impure data with more precise and accuracy in pre-processing which results in efficient data pre-processing for Data Mining. The data pre-processing includes four stages. They are cleaning the Data, Selecting the data, Data Enhancement and Data Transformation. Cleaning the data: is to fill the empty value of the data and to ignore the noisy data and to correct the inconsistencies data. Selecting the data: is choosing the appropriate data which suits for learning. Data Enhancement: is done to enhance the data quality which has been selected. Data Expression: is to transform the data after pre-processing into the form which can be accepted by the data mining algorithm based on neural network. The data mining based on neural network can only handle numerical data, so it is need to transform the sign data into numerical data. The simplest method is to establish a table with one-to-one correspondence between the sign data and the numerical data. The other more complex approach is to adopt the appropriate Hash function to generate a unique numerical data according to given string. Although there are many data types in relational database, but they all basically can be simply come down to sign data, discrete numerical data and serial numerical data. Then, the discrete numerical data can be quantified into continuous numerical data and can also be encoded into coding data which can be easily and efficiently handled by data mining algorithms.

KEYWORDS*—Data mining; neural networks, data mining process, Pre-processing**.*

I. INTRODUCTION

Data pre-processing is an important and critical step in the data mining process and it has a huge impact on the success of a data mining project Data pre-processing is a step of the Knowledge discovery in databases (KDD) process that reduces the complexity of the data and offers better conditions to subsequent analysis. Through this the nature of the data is better understood and the data analysis is performed more accurately and efficiently. Data pre-processing is challenging as it involves extensive manual effort and time in developing the data operation scripts. There are a number of different tools and methods used for pre-processing, including: sampling, which selects a representative subset from a large population of data; transformation, which manipulates raw data to produce a single input; denoising, which removes noise from data; normalization, which organizes data for more efficient access; and feature extraction, which pulls out specified data that is significant in some particular context. Pre-processing technique is also useful for association rules algorithms Like- Apriori, Partitioned, Princer-search algorithms and many more algorithms.

Neural network is a parallel processing network which generated with simulating the image intuitive thinking of human, on the basis of the research of biological neural network, according to the features of biological neurons and neural network and by simplifying, summarizing and refining[1]. It uses the idea of non-linear mapping, the method of parallel processing and the structure of the neural network itself to express the associated knowledge of input and output. Initially, the application of the neural network in data mining was not optimistic, and the main reasons are that the neural network has the defects of complex structure, poor interpretability and long training time.

But its advantages such as high affordability to the noise data and low error rate, the continuously advancing and optimization of various network training algorithms, especially the continuously advancing and improvement of various network pruning algorithms and rules extracting algorithm, make the application of the neural network in the data mining increasingly favored by the overwhelming majority of users. In this paper the data mining based on the neural network is researched in detail.

II. DATA MINING USING NEURAL NETWORK METHODS

There are seven common methods and techniques of data mining which are the methods of statistical analysis, rough set, covering positive and rejecting inverse cases, formula found, fuzzy method, as well as visualization technology. Here, we focus on neural network method.

Neural network method is used for classification, clustering, feature mining, prediction and pattern recognition[3]. It imitates the neurons structure of animals, bases on the M-P model and Hebb learning rule, so in essence it is a distributed matrix structure. Through training data mining, the neural network method gradually calculates (including repeated iteration or cumulative calculation) the weights the neural network connected. The neural network model can be broadly divided into the following three types:

2.1 Feed-forward networks

It regards the perception back-propagation model and the function network as representatives, and mainly used in the areas such as prediction and pattern recognition;

2.2 Feedback network

It regards Hopfield discrete model and continuous model as representatives, and mainly used for associative memory and optimization calculation;

2.3 Self-organization networks

It regards adaptive resonance theory model and mainly used for cluster analysis.

At present, the neural network most commonly used in data mining is BP network. Of course, artificial neural network is the developing science, and some theories have not really taken shape, such as the problems of convergence, stability, local minimum and parameters adjustment. For the BP network the frequent problems it encountered are that the training is slow, may fall into local minimum and it is difficult to determine training parameters[2]. Aiming at these problems some people adopted the method of combining artificial neural networks and genetic gene algorithms and achieved better results.

Artificial neural network has the characteristics of distributed information storage, parallel processing, information, reasoning, and self-organization learning, and has the capability of rapid fitting the non-linear data, so it can solve many problems which are difficult for other methods to solve.

III. DATA MINING PROCESS BASED ON NEURAL NETWORK

Data mining process can be composed by three main phases: data preparation, data mining, expression and interpretation of the results, data mining process is the reiteration of the three phases. The details are shown in Figure. 1.

i. Data Preparation

Data preparation is to define and process the mining data to make it fit specific data mining method. Data preparation is the first important step in the data mining and plays a decisive role in the entire data mining process. It mainly includes the following four processes.

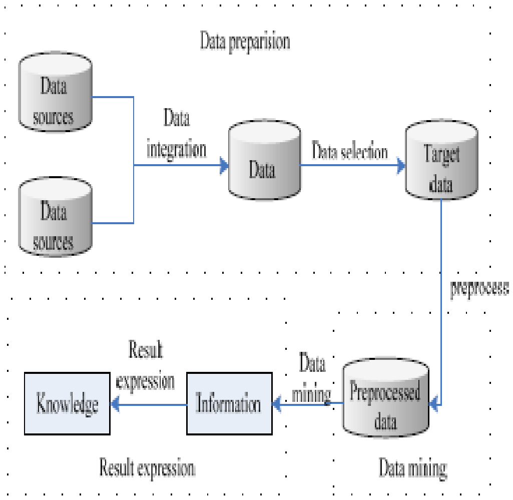
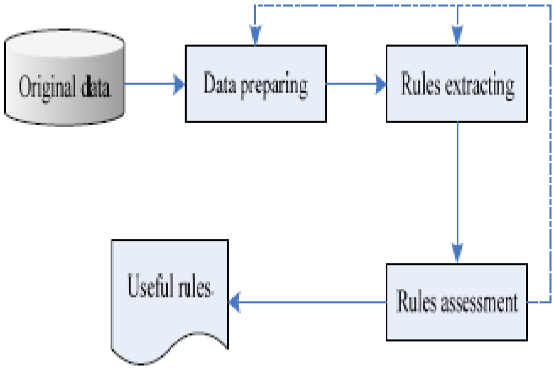


Figure1. General Data mining process

The data mining based on neural network is composed by data preparation, rules extracting and rules assessment three phases, as shown in Figure. 2**.**
(Data mining based on Neural Network)



a) *Data cleaning:* Data cleansing is to fill the vacancy value of the data, eliminate the noise data and correct the inconsistencies data in the data.
b) *Data option:* Data option is to select the data arrange and row used in this mining.
c) *Data preprocessing:* Data preprocessing is to enhanced process the clean data which has been selected.
d) *Data expression:* Data expression is to transform the data after preprocessing into the form which can be accepted by the data mining algorithm based on neural network. The data mining based on neural network can only handle numerical data, so it is need to transform the sign data into numerical data. The simplest method is to establish a table with one-to-one correspondence between the sign data and the numerical data[4]. The other more complex approach is to adopt appropriate Hash function to generate a unique numerical data according to given string. Although there are many data types in relational database, but they all basically can be simply come down to sign data, discrete numerical data and serial numerical data three logical data types.

ii. Rules Extracting

There are many methods to extract rules, in which the most commonly used methods are LRE method, black-box method, the method of extracting fuzzy rules, the method of extracting rules from recursive network, the algorithm of binary input and output rules extracting, partial rules extracting algorithm and full rules extracting algorithm

iii. Rules Assessment

Although the objective of rules assessment depends on each specific application, but, in general terms, the rules can be assessed in accordance with the following objectives

(1) Find the optimal sequence of extracting rules, making it obtains the best results in the given data set;

(2) Test the accuracy of the rules extracted;

(3) Detect how much knowledge in the neural network has not been extracted;

(4) Detect the inconsistency between the extracted rules and the trained neural network.

IV. DATA MINING TYPES BASED ON NEURAL NETWORK

The types of data mining based on neural network are hundreds, but there are only two types most used which are the data mining based on the self-organization neural network and on the fuzzy neural network.

i. Data Mining Based on Self Organization Neural Network

Self-organization process is a process of learning without teachers. Through the study, the important characteristics or some inherent knowledge in a group of data, such as the characteristics of the distribution or clustering according to certain feature. Scholars T. Kohonen of Finland considers that the neighboring modules in the neural network are similar to the brain neurons and play different rules, through interaction they can be adaptively developed to be special detector to detect different signal[5]. Because the brain neurons in different brain space parts play different rules, so they are sensitive to different input modes. T_Kohonen also proposed a kind of learning mode which makes the input signal be mapped to the low-dimensional space, and maintain that the input signal with same characteristics can be corresponding to regional region in space, which is the so-called self-organization feature map.

ii. Data Mining Based on Fuzzy Neural Network

Although neural network has strong functions of learning, classification, association and memory, but in the use of the neural network for data mining, the greatest difficulty is that the output results cannot be intuitively illuminated. After the introduction of the fuzzy processing function into the neural network, it can not only increase its output expression capacity but also the system becomes more stable[7]. The fuzzy neural networks frequently used in data mining are fuzzy perception model, fuzzy BP network, fuzzy clustering Kohonen network, fuzzy inference network and fuzzy ART model. In which the fuzzy BP network is developed from the traditional BP network.

In the traditional BP network, if the samples belonged to the first *k* category, then except the output value of the first *k* output node is 1, the output value of other output nodes all is 0, that is, the output value of the traditional BP network only can be 0 or 1, is not ambiguous. However, in fuzzy BP networks, the expected output value of the samples is replaced by the expected membership of the samples corresponding to various types. After training the samples and their expected membership corresponding to various types in learning stage fuzzy BP network will have the ability to reflect the affiliation relation between the input and output in training set, and can give the membership of the recognition pattern in data mining[6]. Fuzzy clustering networks achieved fuzzy not only in output expression, but also introduced the sample membership into the amendment rules of the weight coefficient, which makes the amendment rules of the weight coefficient has also realized the fuzzy.

V. PROPOSED NEURAL NETWORK MODELS

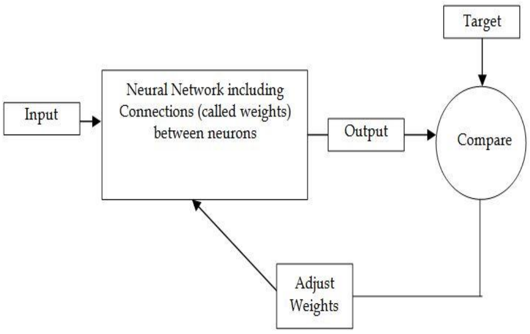
As a starting point, existing knowledge regarding properties of the data such as "data about data" metadata can be used. For example, what is the domain and data type of each attribute? What are the acceptable values for each attribute? What is the range of the length of the values? Do all values fall within the expected range? Are there any known dependencies between attributes? Etc. The data should also be examined regarding unique rules, consecutive rules and null rules. A unique rule says that each value of the given attribute must be different from all other values for that attribute. A Consecutive rule says that there can be no missing values between the lowest and highest values for the attribute, and all values must also unique. A null rule specifies the use of blanks, question marks, special characters or other strings that may indicate the null condition and how such values should be handled.

Simply, the above mentioned way of data pre-processing is inefficient and complex for heart disease prediction systems[8]. So I am suggesting the concept of neural network self-learning algorithm for data pre-processing using heart disease prediction systems.

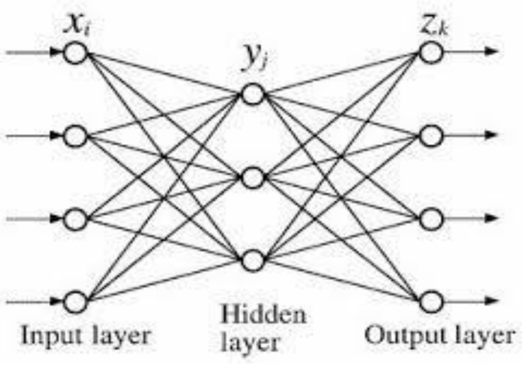
The Methodologies used here are the concept of

(i)Supervised neural networks Models
(ii)Feed forward Neural Networks and
(iii) Back propagation algorithm.

The supervised neural networks such as the multi-layer perception or radial basis functions use training and testing data to build a model. The data involves historical data sets containing input variables, or data fields, which correspond to an output. The training data is what the neural network uses to "learn" how to predict the known output, and the testing data is used for validation. The aim is for the neural networks to predict the output for any record given as the input variables.


Target
Input
Neural Network including Connections (called weights) between neurons
Output
Compare
Adjust Weights

The feedforward neural networks (FFNN), which consists of three layers: an input layer, hidden layer and output layer. In each layer there are one or more processing elements (PEs). Processing elements are meant to simulate the neurons in the brain and this is why they are often referred to as neurons or nodes[9]. A Processing element receives inputs from either the outside world or the previous (parameter) associated with them. This weight is adjusted during training. Information only travels in the forward direction through the network – there are no feedback loops.



The Simplified process for training a FFNN is as follows.

1. Input data is presented to the network and propagated through the network until it reaches output layer. This forward process produces a predicted output.
2. The predicted output is subtracted from the actual output and an error value for the network is calculated.
3. The neural network then uses supervised learning, which in most cases is Back Propagation, to train the network. BackPropogation is a learning algorithm for adjusting the weights. It starts with the weights between the output layer PE's and the last hidden layer PE's and works backwards through network.
4. Once back propagation has finished, the forward process starts again, and this cycle is continued until the error between predicted and actual outputs is minimised.

Neural networks are becoming very popular with data mining practionioners, particularly in medical research, finance and marketing, this is because they have proven their predictive power through comparison with other statistical techniques using real data sets[10]. In most cases neural networks were compared to other statistical techniques, which included various types of discriminant analysis and regression.

VI. CONCLUSION

In recent years, the data mining is a new era in research area and neural network is also very suitable and convenient for solving the problems of data mining and its characteristics of good robustness, self-organizing adaptive, parallel processing, distributed storage and high degree of fault tolerance and accuracy. The combination of data mining method and neural network model can highly improve the efficiency of data mining methods. In future, this combination will attract the researchers enormously.

REFERENCES

[1] Data Pre-processing & Mining Algorithm, Knowledge & Data Mining & Preprocessing, 3rd edition, *Han & Kamber.*

[2] Guan Li, Liang Hongjun. Data warehouse and data mining. Microcomputer Applications. 1999, 15(9): 17-20.

[3] H Lu, R Setiono, H Liu. Effective Data Mining Using Neural Network. IEEE Transactions on Knowledge and Data Engineering, 1996, 8(6): 957-961.

[4] G Towell, J W Shavlik. The extraction of refined rules from knowledge-based neural networks [J]. Machine Learning, 1993(13): 71-101.

[5] Salleb, Ansaf and Christel Vrain, "An Application of Assosiation *Knowledge Discovery and Data Mining (PKDD) 2000*, LNAI 1910, pp. 613-618, Springer Verlag (2000).

[6] Data Pre-processing & Mining Algorithm, Knowledge & Data Mining & Preprocessing, 3rd edition, *Han & Kamber.*

[7] Guan Li, Liang Hongjun. Data warehouse and data mining. Microcomputer Applications. 1999, 15(9): 17-20.

[8] H Lu, R Setiono, H Liu. Effective Data Mining Using Neural Network. IEEE Transactions on Knowledge and Data Engineering, 1996, 8(6): 957-961.

[9] G Towell, J W Shavlik. The extraction of refined rules from knowledge-based neural networks [J]. Machine Learning, 1993(13): 71-101.

[10] Salleb, Ansaf and Christel Vrain, "An Application of Assosiation *Knowledge Discovery and Data Mining (PKDD) 2000*, LNAI 1910, pp. 613-618, Springer Verlag (2000).