ISSN: 2321-7782 (Online)
Volume 3, Issue 7, July 2015
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

To Study Artificial Neural Networks in Data Mining and Its Method

Sujata S.Kharat[1]
M.E 2nd Sem Department of computer Science & Engineering
Anuradha engg.college, Chikhli
Maharashtra, INDIA

Vamshi Krishna[2]
Professor
Department of Computer Science & Engg.
Anuradha engg.college, Chikhli
Maharashtra, INDIA

Abstract: ***Companies have been collecting data for decades, building massive data warehouses in which to store it .Even though this data is available, very few companies have been able to realize the actual value stored in it. The question these companies are asking is how to extract this value. The answer is Data mining. There are many technologies available to data mining practitioners, including Artificial Neural Networks, Regression, and Decision Trees. Many practitioners are wary of Neural Networks due to their black box nature, even though they have proven themselves in many situations. The application of neural networks in the data mining has become wider. Data mining is the search for relationships and global patterns that exist in large databases but are `hidden' among the vast amount of data, such as a relationship between patient data and their medical diagnosis. These relationships represent valuable knowledge about the database and the objects in the database and, if the database is a faithful mirror, of the real world registered by the database. Data mining refers to "using a variety of techniques to identify nuggets of information or decision-making knowledge in bodies of data, and extracting these in such a way that they can be put to use in the areas such as decision support, prediction, forecasting and estimation. This paper is an overview of artificial neural networks and questions their position as a preferred tool by data mining practitioners.***

Keywords: Data Mining, Neural Networks, Data Mining Process, Knowledge Discovery, Implementation.

I. INTRODUCTION

Data mining is the term used to describe the process of extracting value from a database. A data-warehouse is a location where information is stored. The type of data stored depends largely on the type of industry and the company. Many companies store every piece of data they have collected, while others are more ruthless in what they deem to be “important". Consider the following example of a financial institution failing to utilize their data-warehouse. Income is a very important socio-economic indicator. If a bank knows a person’s income, they can offer a higher credit card limit or determine if they are likely to want information on a home loan or managed investments. Even though this financial institution had the ability to determine a customer’s income in two ways, from their credit card application, or through regular direct deposits into their bank account, they did not extract and utilize this information. Another example of where this institution has failed to utilize its data-warehouse is in cross-selling insurance products (e.g. home, life and motor vehicle insurance). By using transaction information they may have the ability to determine if a customer is making payments to another insurance broker. This would enable the institution to select prospects for their insurance products. These are simple examples of what could be achieved using data mining. Four things are required to data-mine effectively: high-quality data, the “right” data, an adequate sample size and the right tool. There are many tools available to a data mining practitioner. These include decision trees, various types of regression and neural networks.

II. Neural Network Method In Data Mining

Artificial Neural Networks:

An Artificial Neural Network (ANN), often just called a "neural network" (NN), is a mathematical model or computational model based on biological neural networks, in other words, is an emulation of biological neural system. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. In most cases an ANN is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase.

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. 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:

Feedforward Neural Network:

One of the simplest feed forward neural networks (FFNN), such as in Figure, consists of three layers: an input layer, hidden layer and output layer. In each layer there are one or more processing elements (PEs). PEs is meant to simulate the neurons in the brain and this is why they are often referred to as neurons or nodes. A PE receives inputs from either the outside world or the previous layer. There are connections between the PEs in each layer that have a weight (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 Back Propagation Algorithm:

Backpropagation, or **propagation of error**, is a common method of teaching artificial neural networks how to perform a given task. The back propagation algorithm is used in layered feedforward ANNs. This means that the artificial neurons are organized in layers, and send their signals “forward”, and then the errors are propagated backwards. The back propagation algorithm uses supervised learning, which means that we provide the algorithm with examples of the inputs and outputs we want the network to compute, and then the error (difference between actual and expected results) is calculated. The idea of the back propagation algorithm is to reduce this error, until the ANN *learns* the training data.

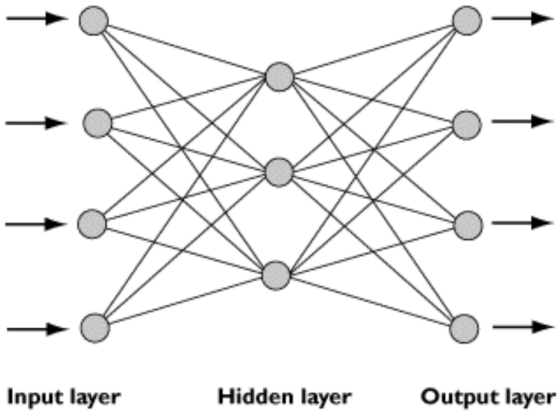

Figure 1.

Self-organization networks:

It regards adaptive resonance theory (ART) model and Kohonen model as representatives, 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. Aiming at these problems some people adopted the method of combining artificial neural networks and genetic gene algorithms and achieved better results.

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 Fig. 2.

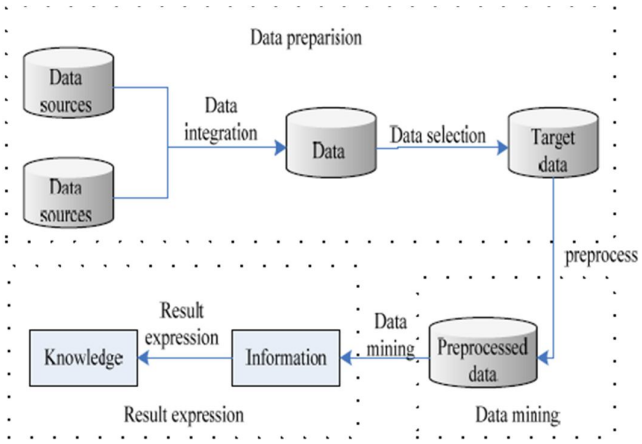

Figure 2.General Data mining process

The data mining based on neural network is composed by data preparation, rules extracting and rules assessment three phases, as shown in Fig. 3

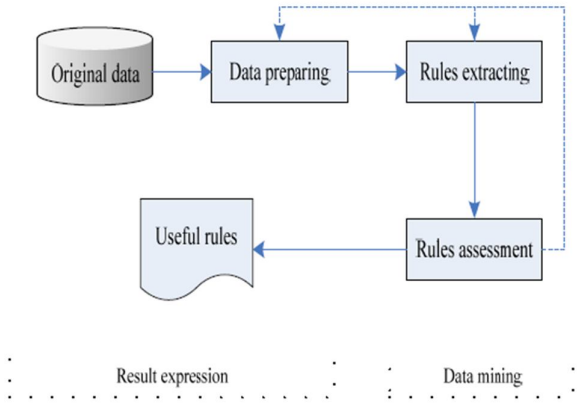

Figure 3.data mining proccess based on neural network

A. 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.

1) 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.

2) Data option: Data option is to select the data arrange and row used in this mining.

3) Data preprocessing: Data preprocessing is to enhanced process the clean data which has been selected.

4) 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.

B. 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 (BIO-RE), partial rules extracting algorithm (Partial-RE) and full rules extracting algorithm (Full-RE).

C. 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.

A. 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. 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 (S0FM).

B. 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 can not 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. 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. Fuzzy clustering Kohonen 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. Developing Neural Network Applications For Data Mining

A. Select Appropriate Paradigm:

Decide on network architecture according to general problem area (e.g., Classification, filtering, pattern recognition, optimization, data compression, prediction), Decide on transfer function, Decide on learning method, Select network size. Eg. How many inputs and output neurons? How many hidden layers and how many neurons per layer? Decide on nature of input/output. Decide on type of training used.

B. Select Input Data and Facts:

Decide the problem domain, the training set should contain a good representation of the entire universe of domain .select input sources and optimal size of training set.

1. Data Set Considerations:

In selecting a data set, the following issues should be considered – Size, Noise, Knowledge domain representation, Training set and test set, insufficient data, Coding the input data.

2. Data Set Size:

Decides the optimal size of the training set, The answer depends on the type of network used. The size should be relatively large. The following is used as a rule of thumb for back propagation networks: Training Set Size = Number of hidden layers + Number of input neuron.

3. Noise:

For back propagation networks, the training is more successful when the data contain noise.

4. Knowledge Domain Representation:

The most important consideration in selecting a data set for Neural Networks The training set should contain a good representation of the entire universe of the domain may result in an increase in number of training facts, which may cause the networks size to change.

5. Selection of Variables:

It is possible to reduce the size of input data without degrading the performance of the network: Principle Component Analysis, Manual Method.

6. Insufficient Data:

When the data is scarce, the allocation of the data into training and a testing set becomes critical. The following schemes are used when collecting more data is not possible.

7. Rotation Scheme:

Suppose the data set has N facts. Set aside one of the facts, training the system with N-1 facts. Then set aside another fact and retrain the network with the other N-1 facts. Repeat the process N times.

8. Creating Made-up Data: Increase the size of the made up data by including made up-data**,** sometimes the idea of BOOTSTRAPPING is used. The decision should be made as whether the distribution of data should be maintained.

9. Expert-made Data:

Ask an expert to supply additional data. Sometimes a multiple expert scheme is used.

10. Coding the Input Data:

The training data set should be properly normalize.

VI. CONCLUSION

At present, data mining is a new and important area of research, and neural network itself is very suitable for solving the problems of data mining because its characteristics of good robustness, self-organizing adaptive, parallel processing, distributed storage and high degree of fault tolerance. The combination of data mining method and neural network model can greatly improve the efficiency of data mining methods, and it has been widely used. It also will receive more and more attention.

References

1. S Lawrence, C Lee Giles. Accessibility of Information on the Web [J]. Nature, 1999, 400(3): 107-109.
2. Guan Li, Liang Hongjun. Data warehouse and data mining. Microcomputer Applications. 1999, 15(9): 17-20.
3. Adriaans P, Zantinge D. Data mining [M]. Addision_Wesley Longman,1996.
4. Chen Rong, BP arithmetic and its structure optimization tactics. Journal of Autoimmunization. 1997, 23(1), 43-49.
5. G Towell, J W Shavlik. The extraction of refined rules from knowledge-based neural networks [J]. Machine Learning, 1993(13): 71-101.
6. Yang Kun, Liu Dayou. Agents: properties and classifications. Computer Science [J]. 1999, 26(9): 30-34.
7. H Lu, R Setiono, H Liu. Effective Data Mining Using Neural Network. IEEE Transactions on Knowledge and Data Engineering, 1996, 8(6): 957-961.
8. David Hand, Principles of Data Mining [M]. Massachusetts Institute of Technology, 2001.
9. Feng Jiansheng. KDD and its applications, BaoGang techniques. 1999(3): 27-31.
10. Wooldrldge M J. Agent-Based software engineering. IEEE Transactions
11. Neural Networks based Data Mining and Knowledge Discovery in Inventory Applications: Kanti Bansal, Sanjeev Vadhavkar, Amar Gupta
12. Data Mining, An Introduction: Ruth Dilly,Parallel Computer Centre, Queen's University
13. DR. Yashpal Singh, Alok Singh Chauhan, Neural Network In Data Mining, Journal of Theoretical and Applied Information Technology
14. Khajanchi, Amit, Artificial Neural Networks: The next intelligence
15. Zurada J.M., "An introduction to artificial neural networks systems", St. Paul: West Publishing (1992)
16. http://www.sau.ac.in/~vivek/softcomp/neuralnet-datamining.pdf
17. Ainscough, T.L., and J.E. Aronson. (1999). "A Neural Networks Approach for the Analysis of Scanner Data." Journal of Retailing and Consumer Services, Vol. 6.
18. Altman, E.I. (1968). "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy." Journal of Finance, Vol. 23.
19. Bennell, J., D. Crabbe, S. Thomas, and O. Gwilym. (2006, April). "Modelling Sovereign Credit Ratings: Neural Networks versus Ordered Probit," Expert Systems with Applications
20. Bradley, I., Introduction to Neural Networks, Multinet Systems Pty Ltd 1997.
21. Fayyad, Usama, Ramakrishna " Evolving Data mining into solutions for Insights", communications of the ACM 45, no. 8
22. Fausett, Laurene (1994), Fundamentals of Neural Networks: Architectures, Algorithms and Applications, Prentice-Hall, New Jersey, USA.

