Research on a New Effective Data Mining Method Based on Neural Networks

Shanli Wang
Tianjin Polytechnic University, Tianjin, China
wangshlitysx@sina.com; wangshanli@tjpu.edu.cn

Abstract

The application of neural networks in the data mining has become wider and wider. Neural networks have high acceptance ability for noisy data, high accuracy and are preferable in data mining. Focuses on the data mining process based on neural network. In this paper the data mining based on fuzzy neural network is researched in detail, the key technology and ways to achieve the data mining based on neural networks are also researched.

1. Introduction

With the continuous development of computer and internet, it is easy to get the related information . But nowadays, the data volume stored in database increases too rapidly and in the large amounts of data much important information is hidden. It is hard to analyze the mass and wide reference date with the anciently state method. So an intellectualized technology, data mining, emergency as the times require, which integrated apply all kinds of state and analyze, date base and capacity language to analyze mass data. By data mining more and more information can be extracted from the database and will create a lot of potential profit for the companies.

Data mining must by data mining tools. With the development of the data mining, a number of data mining tools were developed. Data mining tools can forecast the future trends and activities to support the decision of people. For example, through analyzing the whole database system of the garment company, the data mining tools can answer the problems such as “Which customer is most likely to respond to the color or style of the company, why”, and other similar problems. Some data mining tools can also resolve some traditional problems which consumed much time, this is because that they can rapidly browse the entire database and find some useful information experts unnoticed.

Scientists of all fields for its developments have interested neural network. Neural network is a complex network system 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. 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 “black-box”, we can not understand the learning and decision-making process in the network, poor interpret ability 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.

2. Neural network method in data mining

According to the kernel techniques in the data mining, there are nine common methods and techniques of data mining which are the methods of similar function, rough set, classification algorithms based on association rules, k-nearest neighbor, decision tree, bayes classification algorithms , fuzzy logic, genetic algorithm and neural network. Here, we focus on neural network method.

Neural network method is used for pattern recognition, expert system, classification, feature mining, clustering, prediction. Neural network comes from the neurons structure theory 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) the weights the neural network connected. The neural network model can be broadly divided into the following four types:

(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) Feedback network: it regards Hopfield discrete model and continuous model as representatives, and mainly used for associative memory and optimization calculation;

(3) Self-organization networks: it regards adaptive resonance theory (ART) model and Kohonen model as representatives, and mainly used for cluster analysis.

(4) Random neural network: it is a special kind of artificial neural network, which is developed recently. As a biological neural mathematical model, it has advantages of associative memory, image processing and combinatorial optimization.

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. At present, artificial neural network has some difficult problems. Aiming at the difficult problems(data convergence, stability, local minimum and the training parameters adjustment) some people adopted the method of combining artificial neural networks and genetic gene algorithms and achieved better results. so it can solve many problems which are difficult for other methods to solve.

3. Data mining process based on neural network

General 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.

The data mining based on neural network process can be composed by four main phases: model option, data preparing, rules option and result assessment, according to the result assessment the process is the reiteration of the four phases. The details are shown in Figure 2.

3.1. Model option

Model option is to choice a data mining model based on neural work. The types of data mining model 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. The theory of the data mining based on self-organization neural network is evolutionism theory that is inheritance-aberrance-choice-transform. It can develop from simpleness to complication by itself. So, self-organization process is a process of learning without teachers.

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. Using fuzzy theory and neural network to structure and train fuzzy neural network, the method can overcome the shortcomings of neural network such as complex structure, long training time and lack of understandable representation of results. Establishment and training of fuzzy neural network which meet the precision requests realize the utilization fuzzy neural network method to withdraw the knowledge from the database. It can not only increase its output expression capacity but also the system becomes more stable.

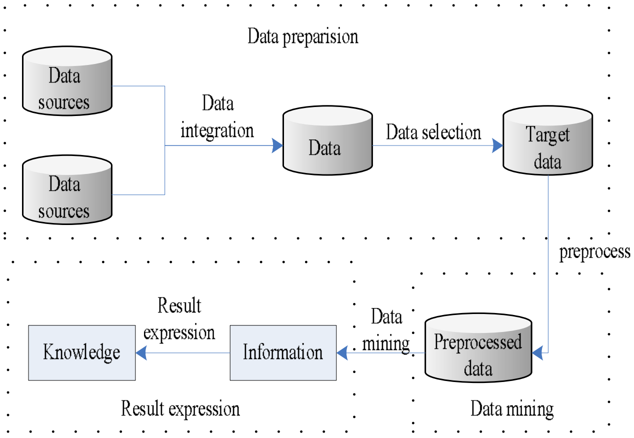

Figure 1. General data mining process

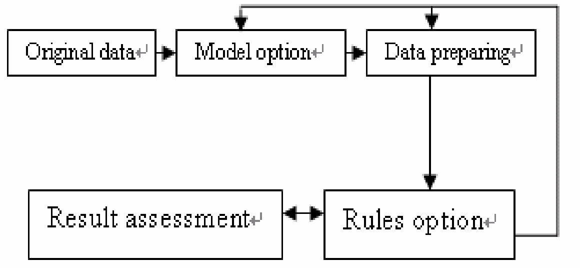

Figure 2 data mining process based on neural network

3.2. Data preparation

Data preparation is to define and process the mining data to make it fit the data mining model which has been selected. Data preparation plays a decisive role in the entire data mining process. It mainly includes the following four processes, and they are data cleaning, data option, data preprocessing, data expression.

Data cleaning: Data cleansing is to eliminate the noise data and find out the disrelated. data, eliminate the vacancy value of the data, correct the inconsistencies data in the data.

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

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

Data expression: Data expression is very important to a mining. Most of the data mining tools can only handle numerical data, but it is impossible that only numerical data in a database. so it is need to transform the sign data into numerical data. According to the character of the sign data, they all basically can be simply come down to sign data, discrete numerical data and serial numerical data three logical data types. Each data type has a appropriate method to transform the sign data. To sign data, the simplest method is to establish a table with one-to-one correspondence between the sign data and the numerical data. To discrete numerical data and serial numerical data, the more complex approach is to adopt appropriate Hash function to generate a unique numerical data according to given string. The detail can be seen in reference 14.

3.3. Rules option

Rule option is the traditional problem of data mining, and is the core problem, too. There are many methods to extract rules, in which the most commonly used methods are black-box method , extract fuzzy rules method , BP network method , the method of extracting rules from recursive network, the algorithm of binary input and output rules extracting, partial rules extracting algorithm (Partial-RE) and full rules extracting algorithm (Full-RE).

3.4. Result assessment

Different application has different rules of assessment , but, in general terms, the result can be assessed in accordance with the following objectives [15].

(1) Whether it can output a right result ,when we input a test example.

(2) Test the accuracy of the best results in the given data set.

(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.

4. Data mining based on fuzzy 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. As an example, we discuss the data mining based on fuzzy neural network.

4.1 The structure of the fuzzy neural network

We can set up a fuzzy neural network model which has five layers[19], seen as in Figure 3.

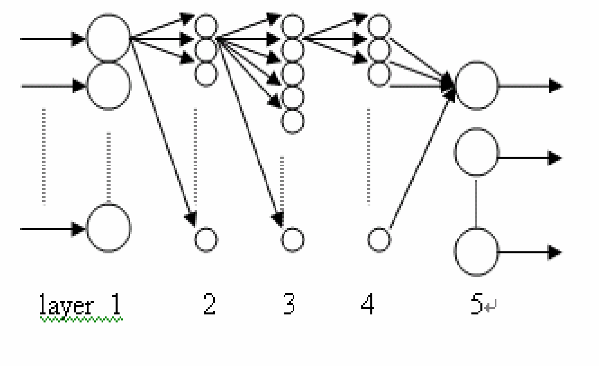

Figure 3. The structure of the fuzzy neural network

In this model, the first layer is input layer; the second compute affiliation relation of the input data; the third is data preprocessing layer; the fourth is rules option, and the fifth is compute the result and output it.

4.2. The standard of the sample data

There is much of the history data in every corporation. We must choice some useful data and extract the attribute of the data, and then prepare all the data according to the data preparation stage. Suppose there are n samples data, they are $x_1,x_2,\ldots,x_k$, and their corresponding history result data. We use all the sample data as training example and get the original affiliation relation , weights .

4.3. Rules option and adjusting

After training the samples and their expected membership corresponding to various types in learning stage fuzzy 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. Because of many cases have not a certain conclude, we must use much of the expert knowledge to enhance quality of the data preprocessing and the rules option. According to the result, adjust the affiliation relations, the weights and the rules again and again, then we can get the good result.

4.4 Effective combination of knowledge processing and neural computation

It may be easy to found a model, but how can we evaluation it? Evaluating whether a data mining implementation algorithm is fine the following indicators and characteristics can be used: (1) whether the data mining implement algorithm can adapt the complex circumstances. Especially, whether high-quality modeling under the circumstances of noise and data half-baked; (2) whether the model is easy to understand by a special user. Especially, the model must be understood by users and can be used for decision-making; (3) whether the model has a good interactive characteristic. In order to improve the modeling quality, the model must receive area knowledge, user's rules enter and weight adjust. As a user, it is not enough to depend on the neural network model providing results, before important decision-making users need to understand the rationale and justification for the decision-making. Therefore, in the ANN data mining knowledge base(common sense, rules , test data) should be established in order to accede domain knowledge and the knowledge ANN learning to the system in the data mining process.

4.5 Input/output interface

Input/output interface is very important to a user. A good interface with relational database, multi-dimensional database and data warehouse should be established to meet the needs of data mining.

5. 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.

6. 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 on Software Engineering [J]. 1999,144 (1): 26-2 7.

[11] Guiqing Wang, Dang Huang . The Summary of The Data Mining Technology . Computer Application Technology [J]2007(69) : 9-14.

[12] Ying Wu, DingFang Chen,etc. .Summarizing of Neural Network Science & Technology Progress and policy [J] 2002(6): 133-134.

[13] Qian Xiaodong, A Review on Classification Algorithms in Data Mining . Library and Information Service[J]. 2007,51(3)::68-71.

[14] Xiuai Zhang, The Research on Data Transform Base on Data Minning**.** Science & Technology Information[J]. 2007(36) :707-708

[15] Zhang Shaobing, Research of Rule Extraction and Classification Algorithm Based on Neural Network. A Dissertation for the Degree of M.ENG ,Harbin Engineering University [M]. 2006.10.13

[16] Duan Lu-ping, Zhou Li-juan ,Wang Yu. Data Mining Based on Neural Networks. Techniques of Automation & Applications [J]. 2007(07) :12,19

[17] Wu Wen-xing, Crucial Technology and Gateway of Redizing Data Mining Based on Nervousl network. Journal of Jiaxing University(Natural Science) [J]. 2006,24(3):70-73

[18] Liu Zhao, Jing Liangxiao, The Research of Data Mining Based on Neural Networks, Computer engineering and application[J], 2004(3)PP:172-173

[19] Li Liangjun, Zhang Bin, Yang Ming, Data Ming Algorithm Based on Fuzzy Neural Network, Computer Engineering[J], 2007(33)PP:63-64