

Research of Data Mining Based on Neural Networks

Xianjun Ni

Abstract—The application of neural networks in the data mining has become wider. Although neural networks may have complex structure, long training time, and uneasily understandable representation of results, neural networks have high acceptance ability for noisy data and high accuracy and are preferable in data mining. In this paper the data mining based on neural networks is researched in detail, and the key technology and ways to achieve the data mining based on neural networks are also researched.

Keywords—Data mining; neural networks, data mining process, implementation.

I. INTRODUCTION

WITH the continuous development of database technology and the extensive applications of database management system, the data volume stored in database increases rapidly and in the large amounts of data much important information is hidden. If the information can be extracted from the database they will create a lot of potential profit for the companies, and the technology of mining information from the massive database is known as data mining.

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 company the data mining tools can answer the problems such as “Which customer is most likely to respond to the e-mail marketing activities of our 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.

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. 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. NEURAL NETWORK METHOD IN DATA MINING

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:

(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.

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

Xianjun Ni is with the Department of Computer Science and Technology, Shandong Institute of Education P. R. China, he is an associate Professor now and his research area is: Data Mining (phone: +86-13688601208; e-mail: nixianjun@gmail.com).

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 Fig. 1.

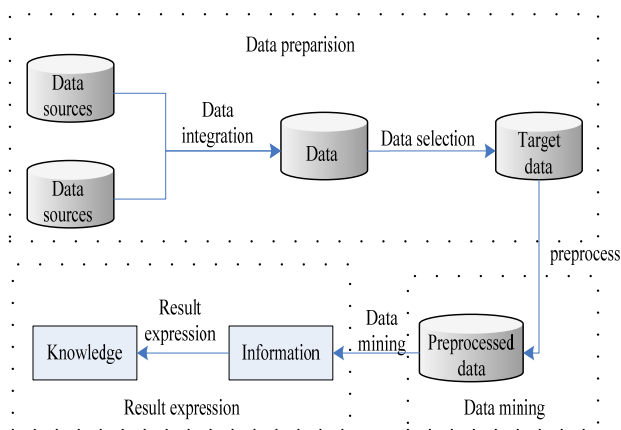


Fig. 1 General data mining process

The data mining based on neural network is composed by data preparation, rules extracting and rules assessment three phases, as shown in Fig. 2.

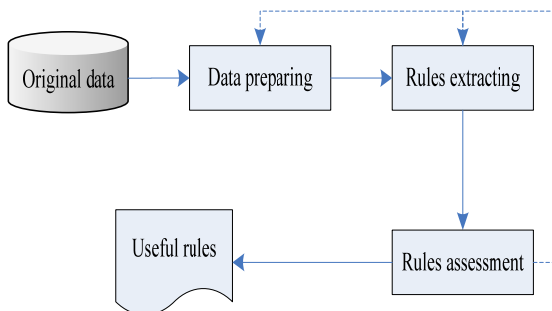


Fig. 2 Data mining process based on neural network

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. 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. Fig. 3 gives the conversion of the three data types. The symbol "Apple" in the figure can be transformed into the corresponding discrete numerical data by using symbol table or Hash function. Then, the discrete numerical data can be quantified into continuous numerical data and can also be encoded into coding data.

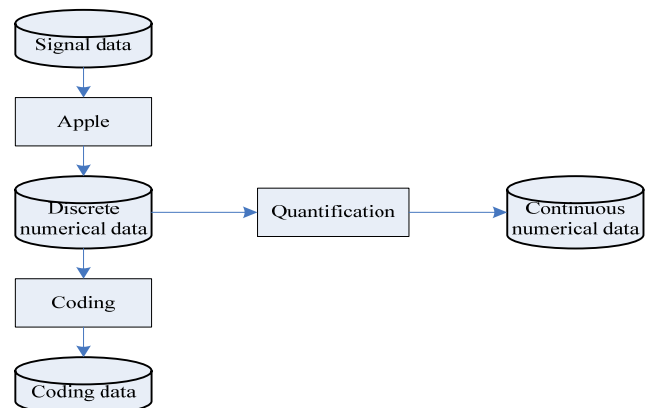


Fig. 3 Data expression and conversion in data mining based on neural network

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 (SOFM).

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. KEY TECHNIQUES AND APPROACHES OF IMPLEMENTATION

A. Effective Combination of Neural Network and Data Mining Technology

The technology almost uses the original ANN software package or transformed from existing ANN development tools, the workflow of data mining should be understood in depth, the data model and application interfaces should be described with standardized form, then the two technologies can be effectively integrated and together complete data mining tasks. Therefore, the approach of organically combining the ANN and data mining technologies should be found to improve and optimize the data mining technology.

B. Effective Combination of Knowledge Processing and Neural Computation

Evaluating whether a data mining implementation algorithm is fine the following indicators and characteristics can be used: (1) whether high-quality modeling under the circumstances of noise and data half-baked; (2) the model must be understood by users and can be used for decision-making; (3) the model can receive area knowledge (rules enter and extraction) to improve the modeling quality. Existing neural network has high precision in the quality of modeling but low in the latter two indicators. Neural network actually can be seen as a black box for users, the application restrictions makes the classification and prediction process can not be understood by users and directly used for decision-making. For data mining, it not enough to depend on the neural network model providing results because that before important decision-making users need to understand the rationale and justification for the decision-making. Therefore, in the ANN data mining knowledge base should be established in order to accede domain knowledge and the knowledge ANN learning to the system in the data mining process. That is to say, in the ANN data mining, it is necessary to use knowledge method to extract knowledge from the data mining process and realize the inoculation of the knowledge processing and neural network. In addition, in the system an effective decision and explanation mechanism should also be considered to be established to improve the validity and practicability of the ANN data mining technology.

C. Input/Output Interface

Considering that the method of using neural network tools or neural network software package to obtain data is laggard, then a good interface with relational database, multi-dimensional database and data warehouse should be established to meet the needs of data mining.

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]. Addison_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] Wooldridge M J. Agent-Based software engineering. IEEE Transactions on Software Engineering [J]. 1999, 144 (1): 26-27.