2011 International Conference on Computer Science and Information Technology (ICCSIT 2011)
IPCSIT vol. 51 (2012) © (2012) IACSIT Press, Singapore
DOI: 10.7763/IPCSIT.2012.V51.09

The Research of Data Mining Based on Neural Networks

Guoquan Jiang[a,]*, Cuijun Zhao[b]

[a]School of Computer Science and Technology, Henan Polytechnic University, Jiaozuo, 454000, China
[b]School of Resources and Environment Engineering, Henan Polytechnic University, Jiaozuo, 454000, China

Abstract. The traditional data mining algorithms are hard to apply on noisy data, redundant information, incomplete data and sparse data in database, or the application effects are not good. But neural network have many virtues such as robustness, parallelism and anti-noise, so it is very effective on data mining in large and real databases. This paper expounds the process of data mining based neural network in detail, discusses the algorithms of classifying and clustering, indicates the problems at present and makes an expectation for the development.

Keywords: Data Mining (Dm), Neural Network (Nn), Self-Organizing Map

1. Introduction

Along with the unceasing development of information technology and application, data rapid expansion, the surge of data behind much important information, People hope to carry on the top level analysis in order to use these data well. The present database system may effectively realizes many functions such as data input, inquiry, statistics etc, but it is difficult to find the relationships and rules that is in the data, it is not able to forecast future trend of development according to the existing data. So it causes the data to explode, but the knowledge is deficient. Facing this challenge, data mining came into being, and shows great vitality.

The present data mining's methods have many kinds, mainly including statistics, rough sets, fuzzy sets, clustering, association rules, decision trees, etc. However, massive data sets, the data relations are very complex, the misalignment degree is quite high, the noise data universal existence, which causes these methods not to be suitable in many situations. However, the processing of such data is precisely the advantage of neural networks. Moreover, the traditional taxonomic approach is only suitable for solving similar gather, heterogeneous separation problems. But in the objective world, many things (for example, different images, sounds, text, etc.) in sample space's region splitting surface are very complex, the similar sample possibly belongs to the different kind, but the sample being far away is possibly same kind.

Neural network can solve non-linear surface approaching, it has a better classification and recognition capability compared with the traditional sorter [1]. Therefore, the neural network being used for data mining has practical significance and practical value.

2. Data mining and neural network

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different

* Corresponding author. Tel.: +86-391-3987-711 .
E-mail address: jiangguoquan@hpu.edu.cn .

dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases [2].

Neural network is complex network system which connects with the massive simple artificial neuron in order to imitate the human brain neural network.

Take the MP model and the Hebb study rule as a foundation, Network model generally may be divided into three kinds:

- Feed forward network, which takes perceptron, antipropagation network as representative, may use in forecasting, pattern recognition and so on;
- Feedback network, which takes the Hopfield discrete model and continuous model as representative, uses in the associative memory and the optimized computation respectively;
- Self-Organizing Network, which takes SOM model, ART model as representative. It is often used for clustering analysis, etc.

3. Neural network-based data mining process

3.1. Data preparation

Data preparation can improve the quality of the data, thus is helpful to the precision and performance of excavation process.

It plays a pivotal role in the entire data mining process. Data mining based on neural networks, as a result of its method's particularity, the data preparation appears especially important, about 50% to 75% development time expenditure in data processing [3]. Data preparation includes data cleaning and selection, data preprocessing and data expression.

3.1.1. Data cleaning and selection

Generally speaking, the data in data warehouse originates from Heterogeneous database.

These heterogeneous data in the database have some inevitably incomplete, inconsistent or inaccurate, these data are called as the dirty data. When loads the data to the data warehouse, must carry on the clean to the dirty data, because the neural network data mining is typical GIGO (garbage in, garbage out.).

After the data, which is used to train neural network, is cleaned and loaded to the data warehouse, we need to choose the data of this excavation used.

First, select the column or parameter-dimensional; second, select the line or record-dimensional.

3.1.2. Data preprocessing

The data pretreatment is a processing of carrying on enhancement processing to the choice of clean data.

This kind of enhancement processing sometimes contains the new data item according to one or many fields, sometimes means that substituted certain fields with an information content bigger field. For data mining based on neural network, it also needs to be transformed into a data, which mining algorithm can be accepted.

3.1.3. Data expression

The database management system supports many data types, majority may sum up simply as continuous numerical data, discrete numerical data and mark data three logic data type.

- Discrete value expression.

The discrete variable only takes the fixed value. To neural network, the discrete value expression way must be helpful in the neural network differentiates between the difference of these discrete values, and can calculate the difference in size. Therefore there are a variety of data encoding schemes, in which some of the most common encoding methods are 1/N code, binary code, the thermometer code.

- Continuous values expression.

For continuous values, the most common form of data conversion is the transformation of scale. Regarding uniform distribution's variable, the simple linear substitution is enough. But for non uniform data, one method is uses the piecewise linearity transform method to carry on the transformation of scale to the data; another method is to set the threshold value.

- Mark data expression.

Mark data often be encountered in the neural network application. The most common and easiest to handle is the "yes / no" this Boolean variable. In practical applications, a third state of the "uncertainty" be often needed.

3.2. Rule extraction

There are many rule extraction methods, the most common methods includes: the LRE (Link Rule Extraction) method, the black box method, two values input output rule extraction algorithm (BIO-RE), from the recursion network extracts rule method and so on.

3.3. Rule evaluation

In general, the rules can be assessed according to the following objectives:

- search for the optimal extraction order, and make it obtain the best results in a given data set;
- test the correctness of the rules to be extracted;
- test how much knowledge of neural networks has not been extracted;
- testing inconsistent places between the rules being extracted and neural networks trained.

4. Data mining method based on neural network

4.1. The classification model based on NN

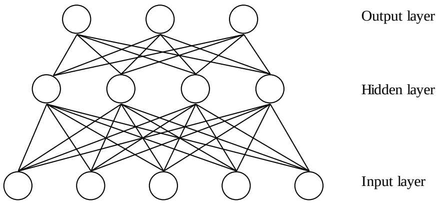

Fig.1. Three layers of BP network topology

Classification is based classification model, the map data in the database, or divided into a number of pre-defined categories. Knowledge discovery in databases also call Supervised Classification. Neural network is a very effective classification tool because of its ability to map any nonlinear function [4].

A continuous transfer function of nonlinear multi-layer feedforward network back propagation network (Back Propagation, BP) is the most used in Supervised Learning of Neural networks, its topology shown in Fig. 1. The Kosmogorov theorem showed that in the reasonable structure and under the condition of suitable weight, three feed-forward network may approach the random continuous function. However, the standard BP algorithm exposes many inherent flaws in the application, such as easy to form a local minimum and the problem of slow convergence and so on, so that some scholars have proposed using the additional momentum and the study rate auto-adapted adjustment to improve BP algorithm, Levenberg-Marquardt optimization method, etc [5].

It is worth mentioning that Shang Lin etc. proposed classification approach based on evolutionary neural networks by the method of improvement the evolution strategy and LM unified, it greatly improved the search efficiency and the classified precision [6].

4.2. Clustering model based on NN

Clustering is to divide a given number of abstract or concrete objects into a finite category. The difference between clustering and classification is that classification is based on training data, and clustering is directly process data. So Clustering is also called Unsupervised Classification. NN-based clustering has two kinds: one is competitive learning; one is Self-Organization feature Mapping.

4.2.1. Competitive learning

Competitive learning is based on a study rule of victor as the king to carry on the competition to the system current processing's object. In the typical competitive learning system, the coupling between the layers is excitatory, that is, the unit of a given level can accept the input of all units from lower level and the layout of active layer represents a high-level input mode.

At a given level, unit in a cluster competes each other, and makes response to the low output mode. A level's relation is suppressed, so that one unit in any cluster is active. The winning unit revises the weight with the other units connected so that it can makes the strong response in the future to the current same or similar object.

If the weight regards as a specimen, the new object will be assigned to the cluster with the most recent samples. Then the number of clusters and unit's number of each cluster is the input parameter. End of the process of clustering, each cluster is considered to be a new "feature", it checks object certain rules. Therefore, the result clusters can be regarded as the mapping from low-level features to high-level features.

4.2.2. Self-organizing feature mapping network

Learning algorithm used by SOM network is called Kohonen algorithm. The algorithm is a non-supervised clustering method, which clusters data by repeatedly learning. SOM is an excellent tool in data mining's investigation stage [7].

Its clustering process also carries on through the certain unit competition current object. Weight vector is closest the unit of current object which becomes active unit or winning unit. In order to close the input object, it is necessary for the winning unit and its nearest neighbor's weight to be adjusted. SOM assumes that the input object has some topological structure or order, the unit will eventually present the structure in space. The organization in unit forms a characteristic mapping.

● Topology

SOM network's basic structure is composed of input layer and competitive layer. Each neuron of input layer collects outside information by weight vector to each neuron of output layer. Input level's form and BP network are the same, the number of nodal point and the sample dimension are equal. Output level is also competition level. Neuron's arrangement has many kinds of forms, such as one-dimensional linear array, two-dimensional array, three- dimensional grid, what is common is the first two types.

SOM network, which output layer is organized according to one-dimensional array, is the simplest self-organizing neural network, the structure shows in Fig. 2 below. May see from the chart, its output level only marks lateral connections between neighboring neurons.

The most typical form of organization of SOM network, which Output layer organizes according to two-dimensional surface, shows in Fig. 3 below. The figure shows its output level's each neuron connects from lateral with other neurons, arranges in chessboard-like plane.

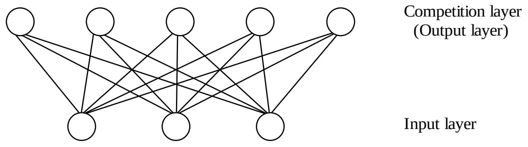

Fig. 2. 1D linear array

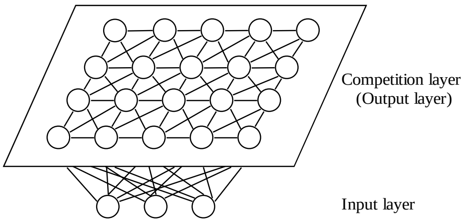

Fig. 3. 2D planar linear array

● Operating Principle

SOM network's movement has two stages: training and work. In the training phase, some node in the output layer will win for producing the maximum response during input sample. Then the winning node and its superior neighborhood all joint connection weight vector makes different degree adjustment to the direction of input vector. After SOM network training, the specific relationship between each node in output layer and each input mode is completely determined, it can be used as a pattern classifier. When inputting a pattern, the pattern particular neuron, which the network output layer represents, will have the greatest response, thus this input automatically classified.

● Improved SOM Clustering

SOM clustering method widely applies in cluster analysis, image processing, voice recognition, combinatorial optimization, data analysis and forecasting [8,9] and so on numerous information processing domain for its unsupervised learning, visualization, and preserving mapping features. However, the traditional

SOM network also has many deficiencies. Its biggest limitation is when the study pattern is small, the network connections weight vector's original state has important influence on the network convergence performance, moreover the network clustering effect depends on the order of the input mode. Alahakoon D[10] presented a growing self-organizing map (GSOM) algorithm, Wang jiacai [11] presented a new dynamic generating algorithm for Self-Organizing Maps controlled by Voronoi region radius(VR2SOM), R.J. May[12] developed a novel approach to stratified sampling, based on Neyman sampling of the self-organizing map (SOM). These research results have the big improvement to the SOM network's cluster performance.

5. Conclusion and Outlook

With a variety of network training algorithms is presented and optimization, in particular a variety of network pruning algorithm and rule extraction algorithms have presented and improved, it makes the data mining based on neural networks more and more to favor for the majority of users. But there are still problems to be solved:

- Algorithm efficiency enhancement. The rule extraction algorithm's computation complexity is a very important limiting factor of neural network in the data mining. Therefore, how to improve algorithm efficiency and reduce computational complexity is also the key point of future work.
- The neural network rule extractions work mainly emphatically the rule which extracts in the enhancement to the network fidelity, namely rule whether can reappear the network really the function. However, in faces the data mining in the application, the regular intelligibility is often more important, in some actual domains, needs to sacrifice certain fidelity to gain a better intelligibility. Therefore, how to strike a medium between the rule intelligibility and the fidelity will be a hot topic in the future.

6. Acknowledgements

This research was supported by the Henan Polytechnic University Doctorial Fund (Grant No. B2010–27).

7. References

[1] Han Liqun. *Artificial Neural Networks Theory, Design and Application*. Beijing: Chemical Industry Publishing House; 2007.

[2] Han J W, Micheline K. *Data Mining: Concepts and Techniques*. San Francisco: Morgan Kaufmann; 2001.

[3] Song Qinbao, Shen Junyi. The Research of Data Preparation for Data Mining with Neural Networks. *Computer Engineering and Applications* 2000; **12**:102–104.

[4] Guo Qiang，P Zh. Neural networks for classification：a survey. *IEEE Tran on system， man，and cybernetics—part C* 2000; **30**: 451－462.

[5] Wang JG， Shang L， Chen SF， Wang YF. Application of fuzzy classification by evolutionary neural network in incipient fault detection of power transformer．*Int'l Joint Conf．on Neural Networks（IJCNN 2004 New York）* 2004; 2279－2283.

[6] Shang Lin, Wang Jin-Gen, Yao Wang-Shu, et al. A Classification Approach Based on Evolutionary Neural Networks. *Journal of Software* 2005; **16**: 1577－1583．

[7] Johan Vesanto，Esa Alhoniemi. Clustering of the Self-Organizing Map. *IEEE Transactions on Neural Networks* 2000; **11**: 586－599．

[8] Lagus K, honkela T, et al. WcbSOM for Textual Data Mining. *Artificial intelligence Review* 1999; **13**: 345–364.

[9] Kohoncn T. Self Organization of a Massive Document Collection. *IEEE Transaction on Neural Networks* 2000; **11**: 574–585.

[10] Alahakoon D， Halgamuge S K. Dynamic Self-Organizing Maps with Controlled Growth for Knowledge Discovery. *lEEE transactions on Neural networks* 2000; **11**: 601–614．

[11] Wang Jiacai; Chen Qi; Yu Ruizhao. A new dynamic generating algorithm for self-organizing maps. *Pattern Recgonition and Artificial Intelligience* 2001; **14**: 360－365.

[12] R.J. May, H.R. Maier, G.C. Dandy. Data splitting for artificial neural networks using SOM-based stratified sampling. *Neural Networks* 2010; **23**: 283－294.