Research on Data Mining Using Neural Networks

K. Amarendra, K.V. Lakshmi & K.V. Ramani
Department of Computer Science & Engineering, Dadi Institute of Engineering & Technology,
NH – 5, Anakapalle – 531002, Visakhapatnam District, Andhra Pradesh, INDIA
E-mail : hodcse@dietakp.com, kvenkatalakshmi@dietakp.com, kvramani@dietakp.com

Abstract **-** 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. The data is often voluminous, but as it stands of low value as no direct use can be made of it; it is the hidden information in the data that is useful". 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, Knowledge Discovery, Implementation.*

I. INTRODUCTION

The past two decades has seen a dramatic increase in the amount of information or data being stored in electronic format. This accumulation of data has taken place at an explosive rate. It has been estimated that the amount of information in the world doubles every 20 months and the size and number of databases are increasing even faster. The increase in use of electronic data gathering devices such as point-of-sale or remote sensing devices has contributed to this explosion of available data. The problem of effectively utilizing these massive volumes of data is becoming a major problem for all enterprises.

Data storage became easier as the availability of large amounts of computing power at low cost i.e., the cost of processing power and storage is falling, made data cheap. There was also the introduction of new machine learning methods for knowledge representation based on logic programming etc. in addition to traditional statistical analysis of data. The new methods tend to be computationally intensive hence a demand for more processing power. It was recognized that information is at the heart of business operations and that decision-makers could make use of the data stored to gain valuable insight into the business. Database Management systems gave access to the data stored but this was only a small part of what could be gained from the data. Traditional on-line transaction processing systems, OLTPs, are good at putting data into databases quickly, safely and efficiently but are not good at delivering meaningful analysis in return. Analyzing data can provide further knowledge about a business by going beyond the data explicitly stored to derive knowledge about the business. This is where Data Mining has obvious benefits for any enterprise.

An artificial neural network (ANN), usually called neural network (NN), is a mathematical model or computational model that is inspired by the structure and/or functional aspects of biological neural networks. A neural network consists of an interconnected group of artificial neurons, and it 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. Modern neural networks are non-linear statistical data modeling tools. They are usually used to model complex relationships between inputs and outputs or to find patterns in data.

II. DATA MINING TECHNIQUES

Researchers identify two fundamental goals of data mining: prediction and description. Prediction makes use of existing variables in the database in order to predict unknown or future values of interest, and

description focuses on finding patterns describing the data and the subsequent presentation for user interpretation. The relative emphasis of both, prediction and description differ with respect to the underlying application and technique.

There are several data mining techniques fulfilling these objectives. Some of these are associations, classifications, sequential patterns and clustering.

Another approach of the study of data mining techniques is to classify the techniques as: user-guided or verification-driven data mining and, discovery-driven or automatic discovery of rules. Most of the techniques of data mining have elements of both the models.

A. Association Rules:

An association rule is an expression of the form X => Y, where X and Y are the sets of items. The intuitive meaning of such a rule is that the transaction of the database, which contains X tends to contain Y. Given a database, the goal is to discover all the rules that have the support and confidence greater than or equal to the minimum support and confidence, respectively.

Let L = $\{l_1, l_2, \ldots, l_m\}$ n be a set of literals called items. Let D, the database, is a set of transactions, where each transaction T is a set of items. T supports an item x, if x is in T. T is said to support a subset of items X, if T supports each item x in X. X => y holds with *confidence c,* if c% of the transactions in D that support X also support Y. The rule X => Y has *support s,* in the transaction set D if s% of the transactions in D supports X U Y. Support means how often X and Y occur together as a percentage of the total transactions. Confidence measures how much a particular item is dependent on another. Patterns with a combination of intermediate values of confidence and support provide the user with interesting and previously unknown information.

B. Clustering :

Clustering is a method of grouping data into different groups, so that the data in each group share similar trends and patterns. The algorithm tends to automatically partition the data space into a set of regions or clusters, to which the examples in the tables are assigned, either deterministically or probability-wise. The goal of the process is to identify all sets of similar examples in the data, in some optimal fashion. Clustering according to similarity is a concept which appears in many disciplines. If a measure of similarity is available, then there are a number of techniques for forming clusters. Another approach is to build set functions that measure some particular property of groups. This latter approach achieves what is known as optimal partitioning.

C. Classification Rules:

Classification involves finding rules that partition the data into disjoint groups. The input for the classification data set is the training data set, whose class labels are already known. Classification analyses the training data set and constructs a model based on the class label, and aims to assign class label to the future unlabelled records. Since the class field is known, this type of classification is known as supervised learning. There are several classification discovery models. They are: the decision tree, neural networks, genetic algorithms and some statistical models.

III. OTHER RELATED AREAS

Data Mining has drawn on a number of other fields, some of which are listed below:

A. Statistics:

Statistics is a theory-rich approach for data analysis, which generates results that can be overwhelming and difficult to interpret. Not with standing this, statistics is one of the foundations on which data mining technology is built. Statistical analysis systems are used by analysts to detect unusual patterns and explain patterns using statistical models. Statistics have an important role to play and data mining will not replace such analyses, but rather statistics can act upon more directed analyses based on the results of data mining.

B. Machine Learning:

Machine learning is the automation of a learning process and learning is tantamount to the construction of rules based on observations. This is a broad field, which includes not only learning from examples, but also reinforcement learning, learning with teacher, etc. A learning algorithm takes the data set and its accompanying information as input and returns a statement e.g. a concept representing the results of learning as output.

C. Inductive Learning:

Induction is the inference of information from data and inductive learning is the model building process where the environment i.e. database is analyzed with a view to finding patterns. Similar objects are grouped in classes and rules formulated whereby it is possible to predict the class of unseen objects. This process of classification identifies classes such that each class has a unique pattern of values, which forms the class description. The nature of the environment is dynamic hence the model must be adaptive i.e. should be able learn. Inductive learning where the system infers knowledge itself from observing its environment has two main strategies: Supervised Learning and Unsupervised Learning.

D. Supervised Learning:

This is learning from examples where a teacher helps the system construct a model by defining classes and supplying examples of each class.

E. Unsupervised Learning:

This is learning from observation and discovery.

F. Mathematical Programming:

Most of the major data mining tasks can be equivalently formulated as problems in mathematical programming for which efficient algorithms are available. It provides a new insight into the problems of data mining.

G. Data Mining Methods:

Various data mining methods are:

- Neural Networks
- Genetic Algorithms
- Rough Sets Techniques
- Support Vector Machines
- Cluster Analysis
- Induction
- OLAP
- Data Visualization

IV. NEURAL NETWORKS

A. Introduction:

Anyone can see that the human brain is superior to a digital computer at many tasks. A good example is the processing of visual information: a one-year-old baby is much better and faster at recognizing objects, faces, and so on than even the most advanced AI system running on the fastest supercomputer. The brain has many other features that would be desirable in artificial systems. This is the real motivation for studying neural computation. It is an alternative paradigm to the usual one (based on a programmed instruction sequence), which was introduced by von Neumann and has been used as the basis of almost all machine computation to date. It is inspired by the knowledge from neuroscience, though it does not try to be biologically realistic in detail.

Neural networks are an approach to computing that involves developing mathematical structures with the ability to learn. The methods are the result of academic investigations to model nervous system learning. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques.

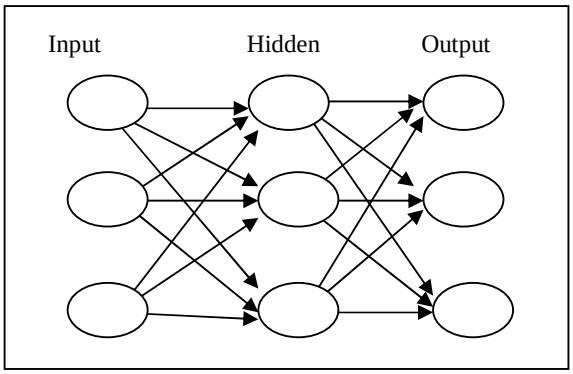

Figure 1. An artificial neural network is an interconnected group of nodes, akin to the vast network of neurons in the human brain.

A trained neural network can be thought of as an "expert" in the category of information it has been given to analyze. This expert can then be used to provide projections given new situations of interest and answer "what if" questions.

Neural networks use a set of processing elements (or nodes) analogous to neurons in the brain. These processing elements are interconnected in a network that can then identify patterns in data once it is exposed to the data, i.e., the network learns from experience just as people do. This distinguishes neural networks from traditional computing programs that simply follow instructions in a fixed sequential order.

Neural networks essentially comprise three pieces: the architecture or model; the learning algorithm; and the activation functions. (Fausett (1994)) Neural networks are programmed or “trained to” . . . store, recognize, and associatively retrieve patterns or database entries; to solve combinatorial optimization problems; to filter noise from measurement data; to control ill defined problems; in summary, to estimate sampled functions when we do not know the form of the functions.” (Kosko (1992), p.13) It is precisely these two abilities (pattern recognition and function estimation) which make artificial neural networks (ANN) so prevalent a utility in data mining. As data sets grow to massive sizes, the need for automated processing becomes clear. With their “model-free” estimators and their dual nature, neural networks serve data mining in a myriad of ways.

B. Structure and Function of a Single Neuron:

McCulloch and Pitts (in 1943) proposed a simple model of a neuron as a binary threshold unit. Specifically, the model neuron computes a weighted

sum of its inputs from other units, and outputs a one or a zero according to whether this sum is above or below a certain threshold.

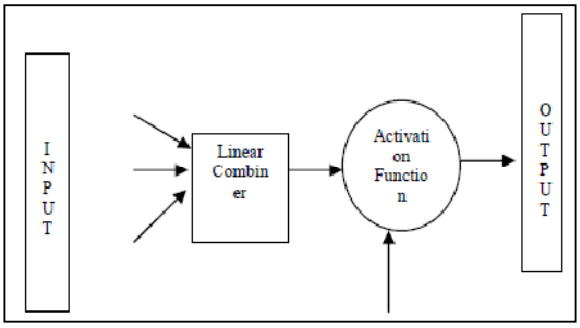

Figure 2. Structure of a Single Neuron: w_{i1}, w_{i2}, …., w_{in} are the Inputs for Linear Combination, Threshold is the Input for Activation Function

The neuron has a set of nodes that connects it to inputs, output, or other neurons, also called synapses (connections / links). A Linear Combiner is a function that takes all inputs and produces a single value. A simple way of doing it is by adding together the weighted inputs. Thus, the linear combiner will produce:

$$(w_{i1} * x_1 + w_{i2} * x_2 + \ldots + w_{in} * x_n) \qquad (1)$$

The Activation Function is a non-linear function, which takes any input from minus infinity to plus infinity and squeezes it into the -1 to 1 or into 0 to 1 interval. This simple model of a neuron makes the following assumptions: The position on the neuron (node) of the incoming synapse (connection) is irrelevant.

Each node has a single output value, distributed to other nodes via outgoing links, irrespective of their positions.

All inputs come in at the same time or remain activated at the same level long enough for computation to occur. (An alternative is to postulate the existence of buffers to store weighted inputs inside nodes).

The threshold is calculated using the Heaviside function as shown below:

$$n_i(t+1) = \Theta\ (\Sigma_j\ w_{ij} n_j(t) - \mu_i) \qquad (2)$$

Here ni is either 1 or 0, and represents the state of neuron i as *firing* or *not firing* respectively. Time t is taken as discrete, with one time unit elapsing per processing step. $\Theta(x)$ is the unit step function, or Heaviside function:

$$\Theta(x) = 1, \text{ if } x >= 0 \text{ and } = 0, \text{ otherwise} \qquad (3)$$

The weight w_{ij} represents the strength of the synapse connecting neuron j to neuron i. It can be positive or negative corresponding to an excitatory or inhibitory synapse respectively. It is zero if there is no synapse between i and j. The cell specific parameter μ_i is the threshold value for unit i; the weighted sum of inputs must reach or exceed the threshold for the neuron to fire.

A simple generalization of the above equation which will consider the activation function is:

$$n_i = g\ (\Sigma_j\ w_{ij} n_j - \mu_i) \qquad (4)$$

The number n_i is now continuous valued and is called state or activation of unit i. The function g(x) is the activation function.

Rather than writing the time t and t+1 explicitly, we now simply give a rule for updating n_i whenever that occurs. Units are often updated asynchronously, in random order, at random times.

C. Characteristics of Neural Networks:

1. Neural Network is composed of a large number of very simple processing elements called neurons.
2. Each neuron is connected to other neurons by means of inter connections or links with an associated weight.
3. Memories are stored or represented in a neural network in the pattern of interconnection strengths among the neurons.
4. Information is processed by changing the strengths of interconnections and/or changing the state of each neurons.
5. A neural network is trained rather than programmed.
6. A neural network acts as an associative memory. It stores information by associating it with other information in the memory. For example, a thesaurus is an associative memory.
7. It can generalize; that is, it can detect similarities between new patterns and previously stored patterns. A neural network can learn the characteristics of a general category of objects on a series of specific examples from that category.
8. It is robust, the performance of a neural network does not degrade appreciably if some of its neurones or interconnections are lost. (distributed memory)

9. Neural networks may be able to recall information based on incomplete or noisy or partially incorrect inputs.
10. A neural network can be self-organizing. Some neural networks can be made to generalize from data patterns used in training without being provided with specific instructions on exactly what to learn.

D. Types of Neural Networks:

A single neuron is insufficient for many practical problems, and network with a large number of nodes are frequently used. The way the nodes are connected determines how computations proceed and constitutes an important early design decision by a neural network developer.

1. Fully Connected Networks:

In this architecture, every node is connected to every node, and these connections may be either excitatory (positive weights), inhibitory (negative weights), or irrelevant (almost zero weights). In a fully connected asymmetric network, the connection from one node to another may carry a different weight than the connection from the second node to the first. In a symmetric network, the weight that connects one node to another is equal to its symmetric reverse. Hidden nodes are the nodes, whose interaction with the external environment is indirect.

2. Layered Networks:

These are networks in which nodes are partitioned into subsets called layers, with no connections that lead from layer j to layer k if j>k. A single input arrives at and is distributed to other nodes by each node of the "input layer" or "layer 0"; no other computation occurs at nodes in layer 0, and there are no intra-layer connections among nodes in this layer. Connections with arbitrary weights, may exist from any node in layer i to any node in layer j for j >= i; intra-layer connections may exist.

3. Acyclic Networks:

This is a subclass of layered networks in which there are no intra-layer connections, as shown in the fig. 2.5. A connection may exist between any node in layer i and any node in layer j for i < j, but a connection is not allowed for i = j. Networks that are not acyclic are referred to as *recurrent* networks.

4. Feed Forward Network:

This is a subclass of acyclic networks in which a connection is allowed from a node in layer i only to nodes in layer i+1. These networks are succinctly described by a sequence of numbers indicating the number of nodes in each layer. These networks, generally with no more than 4 such layers, are among the most common neural nets in use. Conceptually, nodes in successively higher layers abstract successively higher-level features from preceding layers.

5. Feedback Network:

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

6. Modular Neural Networks:

Most problems are solved using neural networks whose architecture consists of several modules, with sparse interconnections between modules. Modularity allows the neural network developer to solve smaller tasks separately using small (neural network) modules and then combine these modules in a logical manner. Modules can be organized in several different ways, some of which are: hierarchical organization, successive refinement and input modularity.

V. DATA MINING PROCESS BASED ON NEURAL NETWORKS

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.

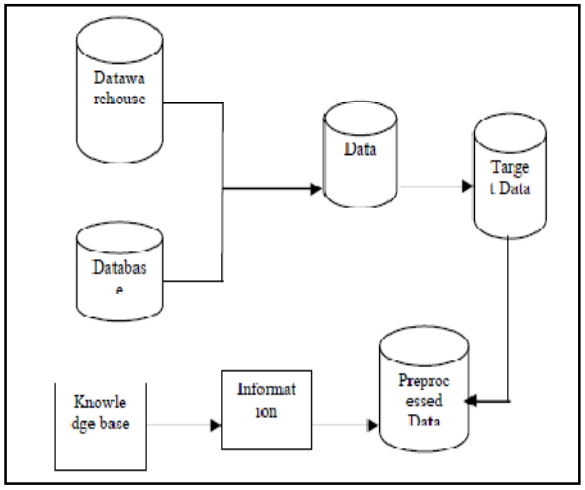

Figure 3. General Data Mining Process

The data mining based on neural network is composed by data preparation, rules extracting and rules assessment three phases, as shown below : 7

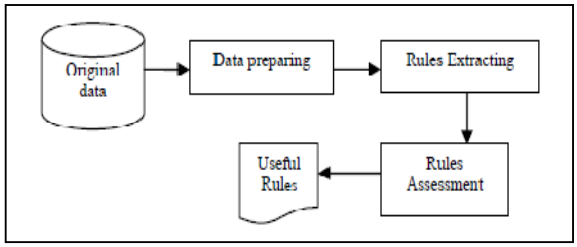

Figure 4. Data mining process based on neural network diagram

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 Clustering:

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 Pre-processing:

Data pre-processing is to enhanced process the clean data which has been selected.

4. Data Expression:

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 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.

B. Data Preparation:

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 Assesment:

1. 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:
2. Find the optimal sequence of extracting rules, making it obtains the best results in the given data set;
3. Test the accuracy of the rules extracted;
4. Detect how much knowledge in the neural network has not been extracted;

Detect the inconsistency between the extracted rules and the trained neural network.

VI. 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 normalized. The training data set should match the design of the network: Zero-mean-unit Variant (Zscore), Min-Max Cut off, Sigmoidale.

C. Data Preparation:

The next step is to think about different ways to represent the information. Data can be non-distributed or distributed. Using a NON Distributed date set, each neuron represents 100% of an item. The data set must be non-overlapping and complete. In this case, the network can represent only a limited number of unique patterns. Using a Distributed data set, the qualities that define a unique pattern are spread out over more than one neuron.

1. Continuous vs. Binary Data:

The developer should decide as whether a piece data is continuous of binary. If a continuous data is represented in a binary form, the network may NOT be able to train properly. The decision as whether a piece data is continuous of binary may not be simple. If the continuous data set is spread evenly within a data range, it may be reasonable to represent it as binary.

2. Actual Values Vs. Change In Values:

An important decision in representing continuous data is whether to use actual amounts or changes in amounts. Whenever possible, it is better to use changes in values. Using the changes in values may make it easier for the network to appreciate the meaning that the data represents.

VII. CONCLUSION

In this paper, we present research on data mining based on neural network. 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] IEEE Transactions on Neural Networks; "Data Mining in a Soft Computing Framework: A Survey", Authors: Sushmita Mitra, Sankar K. Pal and Pabitra Mitra. (January 2002, Vol. 13, No. 1)

[2] Using Neural Networks for Data Mining: Mark W. Craven, Jude W. Shavlik

[3] Data Mining Techniques: Arjun K. Pujari

[4] Introduction to the theory of Neural Computation: John Hertz, Anders Krogh, Richard G. Palmer

[5] Elements of Artificial Neural Networks: Kishan Mehrotra, Chilukuri K. Mohan, Sanjay Ranka.

[6] Artificial Neural Networks: Galgotia Publication

[7] Neural Networks based Data Mining and Knowledge Discovery in Inventory Applications: Kanti Bansal, Sanjeev Vadhavkar, Amar Gupta

[8] Data Mining, An Introduction: Ruth Dilly, Parallel Computer Centre, Queen's University

Belfast: http://www.pcc.qub.ac.uk/tec/courses/ datamining/stu_notes/dm_book_1.html

[9] Introduction to Back propagation Neural Networks: http://cortex.snowseed.com/index.html

[10] Data Mining Techniques: Electronic textbook, Statsoft: http:// www.statsoftinc.com/ textbook/ stdatmin.html#neural

[11] Researchers William J Frawley, Gregory Piatetsky-Shapiro and Christopher J Matheus

[12] S Lawrence, C Lee Giles. Accessibility of Information on the Web [J].Nature, 1999, 400(3): 107-109.

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

[14] Adriaans P, Zantinge D. Data mining [M]. Addision_Wesley Longman,1996.

[15] Chen Rong, BP arithmetic and its structure optimization tactics. Journal of Autoimmunization. 1997, 23(1), 43-49.

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

[17] Yang Kun, Liu Dayou. Agents: properties and classifications. ComputerScience [J]. 1999, 26(9): 30-34.

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

[19] David Hand, Principles of Data Mining [M]. Massachusetts Institute of Technology, 2001.

[20] Feng Jiansheng. KDD and its applications, Bao Gang techniques. 1999(3):27-31.

[21] Wooldrldge M J. Agent-Based software engineering. IEEE Transactionson Software Engineering [J]. 1999,144 (1): 26-27.

❑❑❑