ISSN:0975-9646

Role of soft computing as a tool in data mining

Kanhaiya Lal[#1], N.C.Mahanti[#2]

[#1]*Department of Computer Sc. & Engg., Birla Institute of Technology Patna, Bihar, India*

[#2]*Department of Applied Mathematics, Birla Institute of Technology Mesra, Ranchi, India*

***Abstract*— we live in a world where we can be overwhelmed with information; therefore it has become increasingly important to extract relevant information from the explosive amount of data for. Data Mining is the iterative and interactive process of discovering valid, novel, useful, and understandable patterns or models in massive databases. Data Mining means searching for valuable information in large volumes of data, using exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns and rules. With an enormous amount of data stored in databases and data warehouses develop powerful tools for analysis. Soft computing, are used to identify relationships among a set of items in a database. These relationships are not based on inherent properties of the data themselves (as with functional dependencies), but rather based on co-occurrence of the data items. In this process we discover a set of association rules at multiple levels of abstraction from the relevant set(s) of data in a database. A fundamental challenge is to extend data mining to large data sets.**

In addition to sharing and applying the knowledge in the community, knowledge discovery has become an important issue in the knowledge economic era. Data mining plays an important role of knowledge discovery. Therefore, this study intends to propose a novel framework of data mining which clusters the data first and then followed by association rules mining. Soft computing is being used as the important tool in this area.

The main constitutes of soft computing include fuzzy logic, neural networks, genetic algorithms and rough sets. Each of them contributes a distinct methodology for addressing problems in its domain. This is done in a cooperative, rather than a competitive, manner. The result is a more intelligent and robust system providing a human-interpretable, low-cost, approximate solution, as compared to traditional techniques. This is a review of the role of various soft-computing tools for different data mining tasks.

Keywords— **fuzzy logic, neural networks, genetic algorithms, rough sets, association rule, clustering.**

I. INTRODUCTION

In recent years, there are dramatic changes in the human life, especially the information technology. It has become the essential part of our daily life. Its convenience let us more easily to store any kind of the information regarding science, medicine, finance, population statistics, marketing and so on. However, if there is not a useful method to help us apply these data, then they are only the garbage instead of resources. Due to such demand, there are more and more researchers who pay more attention on how to use the data effectively as well as efficiently. And this is so called data mining.

Data mining includes many areas, in which there are databases techniques, artificial intelligence, machine learning, neural network, statistical techniques, pattern recognition, data visualization etc., is growing up very, estimation, forecasting, clustering, association rule and sequential pattern (Peacock Peter, 1998). Among them, this study intends to propose a framework which integrates both the clustering analysis and association rules mining to discover the useful rules from the database through Soft computing tools[1].

Data mining

Data mining is an increasingly important branch of computer science that examines data in order to find and describe patterns. Because we live in a world where we can be overwhelmed with information, it is imperative that we find ways to classify this input, to find the information we need, to illuminate structures, and to be able to draw conclusions. Data mining is a very practical discipline with many applications in business, science, and government, such as targeted marketing, web analysis, disease diagnosis and outcome prediction, weather forecasting, credit risk and loan approval, customer relationship modeling, fraud detection, and terrorism threat detection. It is based on methods several fields, but mainly machine learning, statistics, databases, and information visualization.

Knowledge discovery in Databases

Knowledge discovery techniques perform data analysis and may uncover important data patterns, contributing greatly to business strategies, knowledge bases, and scientific and medical research. Data mining is an essential step in the process of knowledge discovery in databases. Knowledge discovery as a process consists of an iterative sequence of the following steps:

1. Data cleaning (to remove noise and inconsistent data)
2. Data integration (where multiple data sources may be combined)
3. Data selection (where data relevant to the analysis task are retrieved from the database)
4. Data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance)
5. Data mining (an essential process where intelligent methods are applied in order to extract data patterns)
6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on some interestingness measures)
7. Knowledge representation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user)

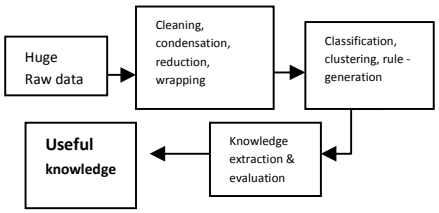


Fig. 1. Block diagram for knowledge discovery

The data mining step may interact with the user or a knowledge base. The interesting patterns are presented to the user, and may be stored as new knowledge in the knowledge base. A pattern is interesting if

1. It is easily understood by humans,
2. valid on new or test data with some degree of certainty,
3. potentially useful, and
4. novel.

The pattern is also interesting if it validates a hypothesis that the user sought to confirm. An interesting pattern represents knowledge. Several objective measures of pattern interestingness exist. These are based on the structure of the discovered patterns and the statistics underlying them. An objective measure for association rules of the form X=>Y is rule support, representing the percentage of transactions from a transaction database that the given rule satisfies. Another objective measure for association rules is confidence, which assesses the degree of certainty of the detected association. Although objective measures help identify interesting patterns, they are insufficient unless combined with subjective measures that reflect the needs and interests of a particular user.

The main steps in the process of Knowledge Discovery includes:

1. Business (or Problem) Understanding
2. Data Understanding
3. Data Preparation (including all the data cleaning and preprocessing)
4. Modeling (applying machine learning and data mining algorithms)
5. Evaluation (checking the performance of these algorithms
6. Deployment

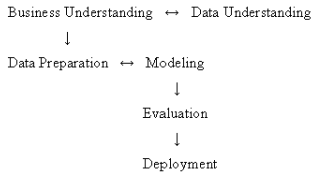


Figure 2: Knowledge Discovery Process Flow

Data mining involves an integration of techniques from multiple disciplines such as database technology, statistics, machine learning, high-performance computing, pattern recognition, neural networks, data visualization, information retrieval, image and signal processing, and spatial data analysis. For large databases, emphasis is placed on efficient and scalable data mining techniques. For an algorithm to be scalable, its running time should grow linearly in proportion to the size of the database, given the available system resources such as main memory and disk space. By performing data mining, interesting knowledge, regularities, or high-level information can be extracted from databases and viewed or browsed from different angles. The discovered knowledge can be applied to decision making, process control, information management, and query processing. Therefore, data mining is considered one of the most important frontiers in database systems and one of the most promising interdisciplinary developments in the information industry.

Association rules

Association rule mining, one of the most important and well researched techniques of data mining. It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories. Association rules are widely used in various areas such as telecommunication networks, market and risk management, inventory control etc. Various association mining techniques and algorithms will be briefly introduced and compared later. Association rule mining is to find out association rules that satisfy the predefined minimum support and confidence from a given database. The problem is usually decomposed into two sub problems. One is to find those item sets whose occurrences exceed a predefined threshold in the database; those item sets are called frequent or large item sets. The second problem is to generate association rules from those large item sets with the constraints of minimal confidence. Suppose one of the large item sets is L_k, $L_k = \{I_1, I_2, \dots, I_k\}$, association rules with this item sets are generated in the following way: the first rule is $\{I_1, I_2, \dots, I_{k-1}\} \in \{I_k\}$, by checking the confidence this rule can be determined as interesting or not. Then other rule are generated by deleting the last items in the antecedent and inserting it to the consequent, further the confidences of the new rules are checked to determine the interestingness of them. Those processes iterated until the antecedent becomes empty. Since the second sub problem is quite straight forward, most of the researches focus on the first sub problem. The first sub-problem can be further divided into two sub-problems: candidate large item sets generation process and frequent item sets generation process. We call those item sets whose support exceed the support threshold as large or frequent item- sets, those item sets that are expected or have the hope to be large or frequent are called candidate item sets. In many cases, the algorithms generate an extremely large number of association rules, often in thousands or even millions. Further, the association rules are sometimes very large. It is nearly impossible for the end users to comprehend or validate such large number of complex association rules, thereby limiting the usefulness of the data mining results. Several strategies have been proposed to reduce the number of association rules, such as generating only "interesting" rules, generating only "non redundant" rules, or generating only those rules satisfying certain other criteria such as coverage, leverage, lift or strength.

Basic Concepts & Basic Association Rules Algorithms

Let $I=I_1, I_2, \dots, I_m$ be a set of m distinct attributes, T be transaction that contains a set of items such that $T \subseteq I$, D be a database with different transaction records Ts. An association rule is an implication in the form of $X \Rightarrow Y$, where $X, Y \subset I$ are sets of items called item sets, and $X \cap Y = \phi$. X is called antecedent while Y is called consequent, the rule means X implies Y. There are two important basic measures for association rules, support(s) and confidence(c). Since the database is large and users concern about only those frequently purchased items, usually thresholds of support and confidence are predefined by users to drop those rules that are not so interesting or useful. The two thresholds are called minimal support and minimal confidence respectively. Support(s) of an association rule is defined as the percentage/fraction of records that contain $X \Rightarrow Y$ to the total number of records in the database. Suppose the support of an item is 0.1%, it means only 0.1 percent of the transaction contain purchasing of this item. Confidence of an association rule is defined as the percentage/fraction of the number of transactions that contain $X \Rightarrow Y$ to the

total number of records that contain X. Confidence is a measure of strength of the association rules, suppose the confidence of the association rule X⇒Y is 80%, it means that 80% of the transactions that contain X also contain Y together. In general, a set of items (such as the antecedent or the consequent of a rule) is called an item set. The number of items in an item set is called the length of an item set. Item sets of some length k are referred to as k-item sets. Generally, an association rules mining algorithm contains the following steps:

• The set of candidate k-item sets is generated by 1-extensions of the large (k -1)-item sets generated in the previous iteration.
• Supports for the candidate k-item sets are generated by a pass over the database.
• Item sets that do not have the minimum support are discarded and the remaining
item sets are called large k-item sets. [3]

This process is repeated until no more large item sets are found.

Increasing the Efficiency of Association Rules Algorithms

The computational cost of association rules mining can be reduced in four ways:

• by reducing the number of passes over the database
• by sampling the database
• by adding extra constraints on the structure of patterns
• through parallelization.

The Apriori algorithm

In mining association rules the two important measures are the *support* and the *confidence*. A *large item set* is an item set with support larger than the support threshold. The common algorithm to compute large item set is the Apriori algorithm. The Apriori algorithm [3] has become a data mining classic and most data mining algorithms are based upon it. The algorithm is depicted below. The most important step of this algorithm is step 3 in the prune step in apriori-gen function, which makes sure that all subsets of a candidate item set are frequent. The basic idea is that any subset of a large item set must be large. Therefore, the candidate item sets having k items can be generated by joining large item sets having $k-1$ items, and deleting those that contain any subset that is not large. The algorithm works as follows:

L_1 = {large 1-itemsets}
for ($k = 2$; $L_{k-1} \neq \phi$; k++) do begin
C_k = apriori-gen (L_{k-1}) //New candidates
for all transactions t in database do begin
C_t = subset (C_k, t) //Candidates contained in t
for all candidates $c \in C_t$ do begin
c.count++;
end
$L_k = \{c \in C_k | c.count \geq minsup\}$
end
Answer = $\cup_k L_k$

The apriori-gen function takes as argument L_{k-1}, the set of all large $(k-1)$-item sets. It returns a superset of the set of all large k-item sets. The function works as follows. First, the join step joins L_{k-1} with L_{k-1}:

1. insert into C_k
2. select $p.item_1$, $p.item_2$, $p.item_{k-1}$, $q.item_{k-1}$
3. from $L_{k-1}p$, $L_{k-1}q$
4. where $p.item_1 = q.item_1$, $p.item_{k-2} = q.item_{k-2}$, $p.item_{k-1} < q.item_{k-1}$;

Next, the prune step deletes all item sets $c \in C_k$ such that some $(k-1)$-subset of c is not in L_{k-1}:

1. forall item sets $c \in C_k$ do
2. forall $(k-1)$-subsets s of c do
3. if ($s \notin L_{k-1}$) then
4. delete c from C_k
5. end
6. end
7. end

To illustrate the concepts, we use a small example from the supermarket domain. The set of items is I = {milk,bread,butter,beer} and a small database containing the items (1 codes presence and 0 absence of an item in a transaction) is shown in the table to the right. An example rule for the supermarket could be $\{milk, bread\} \Rightarrow \{butter\}$ meaning that if milk and bread is bought, customers also buy butter.

Transaction ID	Milk	bread	Butter	Beer
1	1	1	0	0
2	0	1	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

Table 1. - Example data base with 4 items and 5 transactions

Note: this example is extremely small. In practical applications, a rule needs a support of several hundred transactions before it can be considered statistically significant, and datasets often contain thousands or millions of transactions.

II. USEFUL CONCEPTS

To select interesting rules from the set of all possible rules, constraints on various measures of significance and interest can be used. The best-known constraints are minimum thresholds on support and confidence.

- The *support* supp(X) of an item set X is defined as the proportion of transactions in the data set which contain the item set. In the example database, the item set {milk,bread,butter} has a support of 1 / 5 = 0.2 since it occurs in 20% of all transactions (1 out of 5 transactions).
- The *confidence* of a rule is defined $\mathrm{conf}(X \Rightarrow Y) = \mathrm{supp}(X \cup Y)/\mathrm{supp}(X)$. For example, the rule $\{\mathrm{milk}, \mathrm{bread}\} \Rightarrow \{\mathrm{butter}\}$ has a confidence of 0.2 / 0.4 = 0.5 in the database, which means that for 50% of the transactions containing milk and bread the rule is correct.
    - Confidence can be interpreted as an estimate of the probability $P(Y \mid X)$, the probability of finding the RHS of the rule in transactions under the condition that these transactions also contain the LHS.
- The *lift* of a rule is defined as $\mathrm{lift}(X \Rightarrow Y) = \frac{\mathrm{supp}(X \cup Y)}{\mathrm{supp}(Y) \times \mathrm{supp}(X)}$ or the ratio of the observed support to that expected if X and Y were independent .The rule $\{\mathrm{milk}, \mathrm{bread}\} \Rightarrow \{\mathrm{butter}\}$ has a lift of $\frac{0.2}{0.4 \times 0.4} = 1.25$.
- The *conviction* of a rule is defined as $\mathrm{conv}(X \Rightarrow Y) = \frac{1 - \mathrm{supp}(Y)}{1 - \mathrm{conf}(X \Rightarrow Y)}$. The rule $\{\mathrm{milk}, \mathrm{bread}\} \Rightarrow \{\mathrm{butter}\}$ has a conviction of $\frac{1 - 0.4}{1 - .5} = 1.2$, and can be interpreted as the ratio of the expected frequency that X occurs without Y (that is to say, the frequency that the rule makes an incorrect prediction) if X and Y were independent divided by the observed frequency of incorrect predictions. In this example, t he conviction value of 1.2 shows that the rule $\{\mathrm{milk}, \mathrm{bread}\} \Rightarrow \{\mathrm{butter}\}$ would be incorrect 20% more often (1.2 times as often) if the association between X and Y was purely random chance.

[4]

III. CLUSTERING

Clustering

Clustering can be considered the most important *unsupervised learning* problem; so, as every other problem of this kind, it deals with finding a *structure* in a collection of unlabeled data. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar in some way". A *cluster* is therefore a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters. We can show this with a simple graphical example:

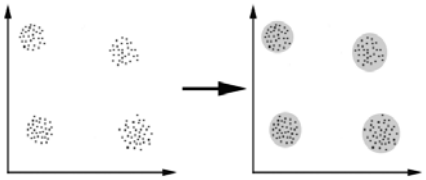


Fig.- 3- clustering

In this case we easily identify the 4 clusters into which the data can be divided; the similarity criterion is *distance*: two or more objects belong to the same cluster if they are "close" according to a given distance (in this case geometrical distance). This is called *distance-based clustering*. Another kind of clustering is *conceptual clustering*: two or more objects belong to the same cluster if this one defines a concept *common* to all that objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures.[5]

The Goals of Clustering

So, the goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But how to decide what constitutes a good clustering? It can be shown that there is no absolute "best" criterion which would be independent of the final aim of the clustering. Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit their needs. For instance, we could be interested in finding representatives for homogeneous groups (*data reduction*), in finding "natural clusters" and describe their unknown properties (*"natural" data types*), in finding useful and suitable groupings (*"useful" data classes*) or in finding unusual data objects (*outlier detection*).

Possible Applications

Clustering algorithms can be applied in many fields, for instance:

- *Marketing*: finding groups of customers with similar behavior given a large database of customer data containing their properties and past buying records;
- *Biology*: classification of plants and animals given their features;
- *Libraries*: book ordering;
- *Insurance*: identifying groups of motor insurance policy holders with a high average claim cost; identifying frauds;
- *City-planning*: identifying groups of houses according to their house type, value and geographical location;
- *Earthquake studies*: clustering observed earthquake epicenters to identify dangerous zones;
- *WWW*: document classification; clustering weblog data to discover groups of similar access patterns.
-

Requirements

The main requirements that a clustering algorithm should satisfy are:

- scalability;
- dealing with different types of attributes;
- discovering clusters with arbitrary shape;
- minimal requirements for domain knowledge to determine input parameters;
- ability to deal with noise and outliers;
- insensitivity to order of input records;
- high dimensionality;
- interpretability and usability.

Distance Measure

An important component of a clustering algorithm is the distance measure between data points. If the components of the data instance vectors are all in the same physical units then it is possible that the simple Euclidean distance metric is sufficient to successfully group similar data instances. However, even in this case the Euclidean distance can sometimes be misleading. Figure shown below illustrates this with an example of the width and height measurements of an object. Despite both measurements being taken in the same physical units, an informed decision has to be made as to the relative scaling. As the figure shows, different scaling can lead to different clustering.

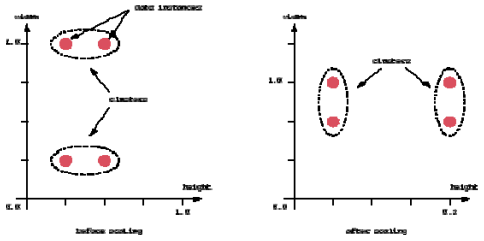


Fig. 4- scaling & clustering

- Notice however that this is not only a graphic issue: the problem arises from the mathematical formula used to combine the distances between the single components of the data feature vectors into a unique distance measure that can be used for clustering purposes: different formulas leads to different clustering. Again, domain knowledge must be used to guide the formulation of a suitable distance measure for each particular application.[6]

IV. FUZZY LOGIC IN DATA MINING

Based on fuzzy set theory, fuzzy logic provides a powerful way to categorize a concept in an abstract way by introducing vagueness. On the other hand, data mining methods are capable of extracting patterns automatically from a large amount of data. The integration of fuzzy logic with data mining methods will help to create more abstract patterns at a higher level than at the data level. Decreasing the dependency on data will be helpful for patterns used in intrusion detection.

Traditionally, a standard set like $S = \{a, b, c, d, e\}$ represents the fact that every member totally belongs to the set S. However, there are many concepts that have to be expressed with some vagueness. For instance, “tall” is fuzzy in the statement of “John’s height is tall” since there is no clear boundary between “tall” and not “tall” (Stefik 1995; Hodges, Bridges, and Yie 1996).

Fuzzy set theory established by Lotfi Zadeh is the basis of fuzzy logic (Stefik 1995). A fuzzy set is a set to which its members belong with a degree between 0 to 1. For example, $S' = \{(a\ 0), (b\ 0.3), (c\ 1), (d\ 0.5), (e\ 0)\}$ is a fuzzy set in which a, b, c, d, and e have membership degrees in the set of S' of 0, 0.3, 1, 0.5, and 0 respectively. So, it is absolutely true that a and

e do not belong to *S'* and *c* does belong to *S'*, but *b* and *e* are only partial members in the fuzzy set *S'*.
A fuzzy variable (also called a linguistic variable) can be used to represent these concepts associated with some vagueness. A fuzzy variable will then take a fuzzy set as a value, which is usually denoted by a fuzzy adjective. For example, "height" is a fuzzy variable and "tall" is one of its fuzzy adjectives, which can be represented by a fuzzy set [7].

V. NEURAL NETWORK IN DATA MINING

In this note we provide an overview of the key concepts that have led to the emergence of Artificial Neural Networks as a major paradigm for Data Mining applications. Neural nets have gone through two major development periods -the early 60's and the mid 80's. They were a key development in the field of machine learning. Artificial Neural Networks were inspired by
biological findings relating to the behavior of the brain as a network of units called neurons. The human brain is estimated to have around 10 billion neurons each connected on average to 10,000 other neurons. Each neuron receives signals through synapses that control the effects of the signal on the neuron[10]. These synaptic connections are believed to play a key role in the behavior of the brain. The fundamental building block in an Artificial Neural Network is the mathematical model of a neuron as shown in Figure 5. The three basic components of the (artificial) neuron are:
1. The synapses or connecting links that provide weights, wj , to the input values, xj *for* $j = 1, ...m$;

2. An adder that sums the weighted input values to compute the input to the activation function $v = w0 + m\_j=1\ wjxj$,where $w0$ is called the bias (not to be confused with statistical bias in prediction or estimation) is a numerical value associated with the neuron. It is convenient to think of the bias as the weight for an input whose value is always equal to one, so that $v = m\_j=0\ wjxj$;

3. An activation function g (also called a squashing function) that maps v to $g(v)$ the output value of the neuron[11]. This function is a monotone function.

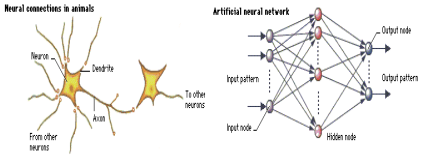


Fig. 5- Neural Network

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 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.

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.6. 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. 7 . and Fig. 8 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. 8 Data expression and conversion in data mining based on neural network

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).

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.

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 cannot 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.

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:

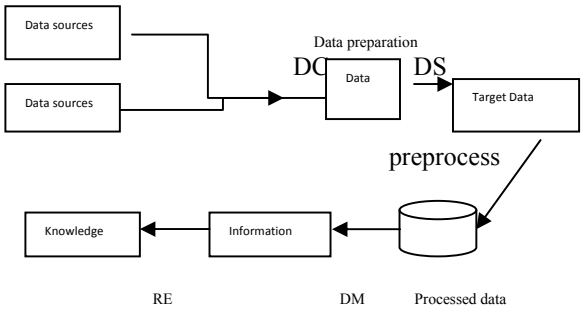
(1) whether high-quality modeling under the circumstances of noise and data half-baked;

(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 cannot 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 inosculation 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[9].



DC- data concatenation
DS-data selection
DM-data mining
RE- result expression

Fig. 6 - General data mining process

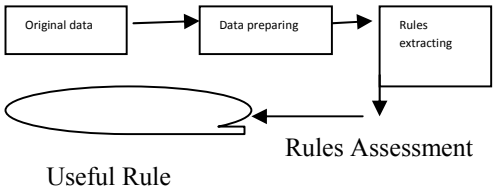


Fig. 7 – data mining process based on neural network

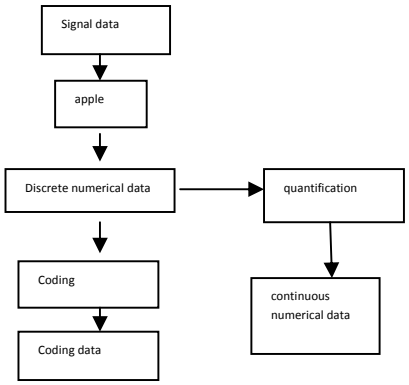


Fig. 8 – Data expression and conversion in data mining based on neural network

VI. GENETIC ALGORITHM IN DATA MINING

Genetic algorithm is the random optimization method based on the principle of natural selection and biological
evolution. Which changes the solution of problems into data individuals of a gene sting structure in genetic space by a certain coding scheme, converts objective function into fitness value, evaluates advantages and disadvantages of
individuals, and as the basis to the genetic operation, It implements through the steps of identification of initial
population, selection, cross, variation, evaluation and screening[12]. Comparing with traditional optimization methods, the use of group search strategy, so information exchange between individuals of group and not dependent on the gradient information when research, processing characteristics of not dependent on problem model, suitable
for parallel processing, with the strong ability of global search function solve problems, strong robustness etc. Now,
It is used in mechanical engineering, electronic engineering, knowledge discovery, combinatorial optimization, machine learning, image processing, knowledge acquisition and data mining, adaptive control and artificial life, and other fields [13]. The major running steps of genetic algorithm are as follows:

1) Establish initial groups randomly with strings.

2) Calculate the fitness value of individuals.

3) According to genetic probability, to create new population by using the following operation.

a) Copy: Add existed excellent individuals copy to a new group, delete poor-quality individuals.

b) Hybrid: Exchange the two selected individual, the new individual of which will be added to the new group.

c) Variability: Random exchange a certain individual characters and then insert into a new group. Repeat the implementation of hybrid and variability, choosing the best individual as the results of genetic algorithm once arrive to the conditions. *Genetic algorithm in the position of data mining* Genetic algorithm plays an important role in data mining technology, which is decided by its own characteristics and advantages[12]. To sum up, mainly in the following aspects:

1) Genetic algorithm processing object not parameters itself, but the encoded individuals of parameters set, which directly operate to set, queue, matrices, charts, and other structure.

2) Possess better global overall search performance; reduce the risk of partial optimal solution. At the same time, genetic algorithm itself is also very easy to parallel.

3) In standard genetic algorithm, basically not use the knowledge of search space or other supporting information, but use fitness function to evaluate individuals, and do genetic Operation on the following basis.

4) Genetic algorithm doesn't adopt deterministic rules, but adopts the rules of probability changing to guide search direction.

VII. ROUGH SETS IN DATA MINING

The theory of rough sets [14] has emerged as a major mathematical tool for managing uncertainty that arises from granularity in the domain of discourse, i.e., from the indiscernibility between objects in a set, and has proved to be useful in a variety of KDD processes. It offers mathematical tools to discover hidden patterns in data and therefore its importance, as far as data mining is concerned, can in no way be overlooked. A fundamental principle of a rough set-based learning system is to discover redundancies and dependencies between the given features of a problem to be classified. It approximates a given concept from below and from above, using *lower* and *upper approximations*.

Fig. 9 provides a schematic diagram of a rough set. A rough set learning algorithm can be used to obtain a set

of rules in IF-THEN form, from a *decision table*. The rough set method provides an effective tool for extracting knowledge from databases. Here one first creates a knowledge base, classifying objects and attributes within the created decision tables. Then a knowledge discovery process is initiated to remove some undesirable attributes. Finally the data dependency is analyzed, in the reduced database, to find the minimal subset of attributes called

reduct. Rough set applications to data mining generally proceed along the following directions.

1) *Decision rule induction from attribute value table* . Most of these methods are based on generation of discernibility matrices and reducts.

2) *Data filtration by template generation* [15]. This mainly involves extracting elementary blocks from data based on equivalence relation. Genetic algorithms are also some- times used in this stage for searching, so that the methodologies can be used for large data sets. Besides these, reduction of memory and computational requirements

for rule generation, and working on dynamic databases [16] are also considered. Some of the rough set-based systems developed for data mining include

1) the KDD-R system based on the variable precision rough set (VPRS) model

2) the rule induction system based on learning from examples based on rough set theory to handle *missing a*ttributes using the closest fit.

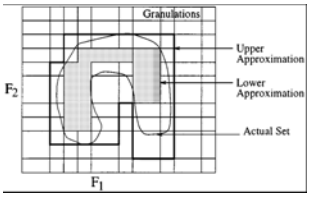


Fig. 9.. Lower and upper approximations in a rough set.

CONCLUSIONS

At present, data mining is a new and important area of research, and soft computing tools itself are 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 soft computing tools like ANN, GA, FL, Rough sets can greatly improve the efficiency of data mining methods, and it has been widely used. It also will receive more and more attention Artificial Neural Networks offer qualitative methods for business and economic systems that traditional quantitative tools in statistics and econometrics cannot quantify due to the complexity in translating the systems into precise mathematical functions. Hence, the use of soft computing tools in data mining is a promising field of research especially given the ready availability of large mass of data sets and the reported ability of neural networks to detect and assimilate relationships between a large numbers of variables. Recently, several commercial data mining tools have been developed based on soft computing methodologies. These include Data Mining Suite, using fuzzy logic; for Thought and IBM Intelligent Miners for Data, using neural networks; and Nuggets, using GAs. Since the databases to be mined are often very large, parallel algorithms are desirable . However, one has to explore a tradeoff between computation, communication, memory usage, synchronization, and the use of problem-specific information to select a suitable parallel algorithm for data mining. One can also also partition the data appropriately and distribute the subsets to multiple processors, learning concept descriptions in parallel, and then combining them. This corresponds to loosely coupled collections of otherwise independent algorithms, and is termed *distributed data mining* [17].

REFERENCES

[1] R. J. Kuo S.Y.Lin and C.W.Shih, Mining Association rules through integration of clustering analysis and ant colony system for health insurance database in Tiwan, Expert system with application, Vol. 33.,Issue 3, 2007.

[2] Han J , Kamber M. “Data Mining: Concepts and Techniques”. 2/e San Francisco: CA. Morgan Kaufmann Publishers, an imprint of Elsevier. 2006. pp-5-38 .

[3] GESTS International Transactions on Computer Science and Engineering, Vol.32 (1), 2006, pp. 71-82.

[4] Kanhaiya Lal etal., International journal of advanced research in computer sc. Vol.(1), 2010, pp . 90-94.

[5]http://www.cs.bris.ac.uk/home/tr1690/documentation/fuzzy_clustering_initial_report/node11.html.

[6] Osmar R. Zaïane: "Principles of Knowledge Discovery in Databases - Chapter 8: Data Clustering".

[7] Jianxiong Luo, "integrating fuzzy logic with data mining methods for intrusion detection, 1999.

[8] Sushmita Mitra, Sankar K. Pal, and Pabitra Mitra," Data Mining in Soft Computing Framework: A Survey, IEEE Transactions on neural networks, Vol. 13, no.. 1, January 2002.

[9] Xianjun Ni," Research of Data Mining Based on Neural Networks " , World Academy of Science, Engineering and Technology 39 2008

[10]Yashpal Singh etal," Neural networks in data mining", **Journal of Theoretical and Applied Information Technology ,2009**

[11] Haykin, S., *Neural Networks*, Prentice Hall International Inc., 1999

[12] Tan Jun-shan, He Wei, Qing Yan, Application of Genetic Algorithm in Data Mining, First International Workshop on Education Technology and Computer Science ,IEEE ,2009

[13] G. Y. Yu, Y. Z. Wang, "Applied Research of improved genetic algorithms," Machinery, vol. 5, 2007, pp. 58-60.

[14] Z. Pawlak, *Rough Sets, Theoretical Aspects of Reasoning about Data*. Dordrecht, : Kluwer, 1991.

[15] L. Polkowski and A. Skowron, *Rough Sets in Knowledge Discovery 1 and 2*. Heidelberg, Germany: Physica-Verlag, 1998.

[16] N. Shan and W. Ziarko, "Data-based acquisition and incremental modification of classification rules," *Comput. Intell.*, vol. 11, pp. 357– 370, 1995.

[17] H. Kargupta and P. Chan, *Advances in Distributed and Paralell Knowledge Discovery*. Cambridge, MA: MIT Press, 2000.