ISSN:0975-9646
K.Suresh et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (3) , 2012,3952-3957

Knowledge Discovery in Datamining Using Soft Computing

*K.Suresh,*Ch.Jnaneswari,*G.Lakshmi Kranthi, **K.Bindu

*Sri Prakash College of Engineering, Tuni
**HITAM Engineering College, Hyderabad

Abstract:-Many real world domains are inherently spatial temporal in nature. In this work, we introduce significant enhancements to two spatiotemporal relational learning methods, the spatiotemporal relational probability tree and the spatiotemporal relational random forest, that increase their ability to learn using spatiotemporal data. We enabled the models to formulate questions on both objects and the scalar and vector fields within and around objects, allowing the models to differentiate based on the gradient, divergence, and curl and to recognize the shape of point clouds defined by fields. This enables the model to ask questions about the change of a shape over time or about its orientation. These additions are validated on several real-worlds hazardous weather datasets. We demonstrate that these additions enable the models to learn robust classifiers that outperform the versions without these new additions. 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. So 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. The last section exemplifies text mining in the context of a number of successful applications. Text mining offers a solution to this problem by replacing or supplementing the human reader with automatic systems Undeterred by the text explosion. It involves analyzing a large collection of documents to discover previously unknown information. The information might be relationships or patterns that are buried in the document collection and which would otherwise be extremely difficult, if not impossible, to discover. Text mining can be used to analyze natural language documents about any subject, although much of the interest at present is coming from the biological sciences.

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

I INTRODUCTION

The knowledge discovery and data mining (KDD) [3, 8] field draws on findings from statistics, databases, and artificial intelligence to construct tools that let users gain insight from massive data sets [10]. People in business, science, medicine, academia, and government collect such data sets, and several commercial packages now offer general-purpose KDD tools [2, 9] .In today's world data mining is a very important concept that used in Analysis and prediction. Data Mining (DM) is understood as a process of automatically extracting meaningful, useful, previously unknown and ultimately comprehensible information from large databases. Two words are crucial in above definition: DM is an automatic process that - once tailored and started can be run without human intervention (as opposed to OLAP), and databases that DM mines knowledge from are very large, and therefore not subject to human analysis. Data Mining is not a single method or algorithm - it's rather a collection of various tools and approaches sharing the common purpose to "torture the data until they confess". The results of Data Mining analysis can be miscellaneous, ranging from discovering customer behavior, to fraud detection and automatic market segmentation, to full-text document analysis. 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

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)

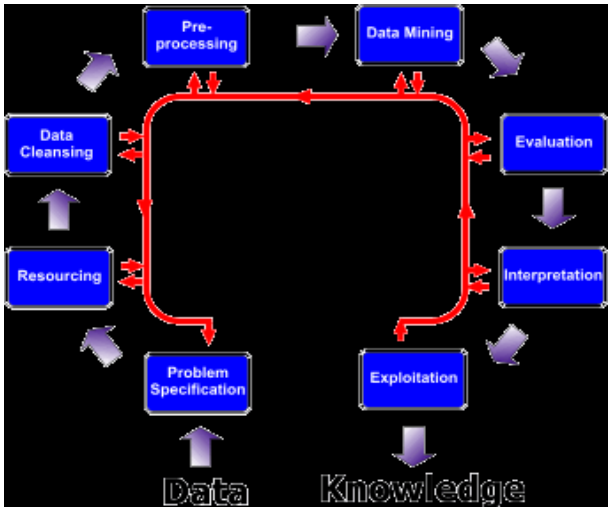

Fig. 1. Block diagram for knowledge discovery

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

Association rules

Association rules finding is perhaps the most spectacular example of Data mining, because it can quickly contribute to sales volume or profit when correctly implemented.

Association models find items that occur together in a given event or record. They try to discover rules of the form: if an event includes object A, then with certain probability7 object B is also part of that event. Consider for example large supermarket network using association rules finding to analyze their databases. These databases contain information about transactions made by customers: articles bought, volume, transaction time etc. During the analysis process such hypothetical rules could be discovered: If a male customer buys beer, then in 80% of cases he also buys potato chips or If a customer is paying at cash desks 1-5, then in 60% of cases he is not buying the daily newspaper. Using potato chips stand could be moved away from the beer stand, to force customers to visit more supermarket space. Special "beer plus chips" bundles could be introduced for customers' convenience. The newspapers stand could be probably installed near cash desks 1-5 and so on.

Basic Concepts & Basic Association Rules Algorithms

Let I=I1, I2, …, I'm be a set of m distinct attributes, T be transaction that contains a set of items such that T ID be a database with different transaction records Ts. An association rule is an implication in the form of X□Y, where X, Y □ I are sets of items called item sets, and X ∩ Y =□. 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 □ Y to the total number of records in the data- base. 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 □ 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.

II. 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:

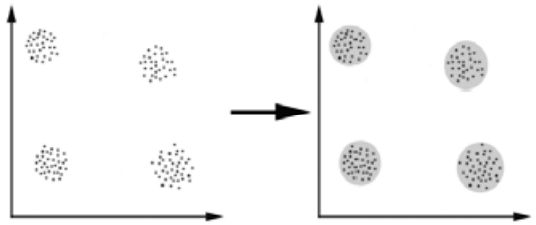

Fig: 3- clustering

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.

III. 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].

IV.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. 5. General data mining process, the data mining based on neural network is composed by data preparation, rules extracting and rules assessment three phases. Data mining process based on neural network

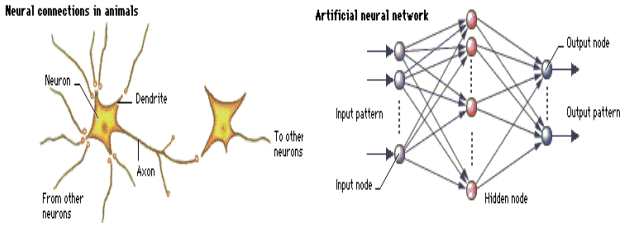

Fig. 5- Neural Network

. 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 consistencies 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.

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.

V. KEY TECHNIQUES AND APPROACHES OF IMPLEMENTATION

A. Effective Combination of Neural Network and Data Mining Technology

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

B. Effective Combination of Knowledge Processing and Neural Computation

Evaluating whether a data mining implementation algorithm is fine the following indicators and characteristics can be used:

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

(2) The model must be understood by users and can be used for decision-making;

(3) The model can receive area knowledge (rules enter and extraction) to improve the modeling quality. Existing neural network has high precision in the quality of modeling but low in the latter two indicators. Neural network actually can be seen as a black box for users, the application restrictions

makes the classification and prediction process 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 needs 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.

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, and 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 functions 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.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.*

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 Taiwan, 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 Mitral, Shankar 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] Arijay Chaudhry & P. S. Deshpande, _Multidimensional Data Analysis and Data Mining Dreamtech Press.

[11] Vikram Pudi & RadhaKrishna, DataMining, Oxford Higher Education.

[12] Chakrabarti, S. _Mining the Web: Discovering knowledge from hypertext data.

[13].M. Jarke, M. Lenzerini, Y. Vassiliou, P. Vassiliadis (ed.), Fundamentals of Data Warehouses_, Springer-Verlag, 1999.

ABOUT AUTHORS

K.suresh working as assistant professor in Sri Prakash College of engineering, Tuni.He has 5 years teaching experience and good knowledge in computer subjects. He completed master degree in computer science and system engineering dept. from A.U College of engineering, in Visakhapatnam. He is pursuing PhD in jntuk.

CH.Jnaneswari working as assistant professor in Sri Prakash College of engineering, Tuni. He has 4 years teaching experience and good knowledge in computer subjects. She completed master degree in computer science and engineering dept. from Godavari institute engineering technology, in Rajahmundry.

Lakshmi kranthi G working as assistant professor in Sri Prakash College of Engineering, Tuni. She has 4.5 years teaching experience and good knowledge in computer subjects. She completed master degree in computer science and Engineering dept. from Sri Vasavi Engineering College in Tadepalligudem.

K.Bindu working as assistant professor in HITAM Engineering College, Hyderabad. She has 5 years teaching experience and good knowledge in computer subjects. She completed master degree in computer science and Engineering dept. from Sri Vasavi Engineering College in Tadepalligudem.