IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 12, December 2015.
www.ijiset.com
ISSN 2348 – 7968

Soft-Computing Based Data Mining: A Review

***Manoj Kumar Jha, **Neelam Sahu, ***Ruchi Trivedi**

*Department of Applied Mathematics, Rungta Engg. College, Raipur, India (manojjha.2010@rediffmail.com)

**Ph.D. Scholar, Computer Science Department, Dr. C.V. Raman University, Bilaspur, India
(neelam.sahu16@rediffmail.com)

***Ph.D. Scholar, Department of Mathematics, Dr. C.V. Raman University, Bilaspur, India
(trivedi.ruchi2201@gmail.com)

Abstract: Soft computing is a consortium of methodologies that works synergistically and provides, in one form or another, flexible information processing capability for handling real-life ambiguous situations. Its aim is to exploit the tolerance for imprecision, uncertainty, approximate reasoning, and partial truth in order to achieve tractability, robustness, and low-cost solutions. The guiding principle is to devise methods of computation that lead to an acceptable solution at low cost by seeking for an approximate solution to an imprecisely/precisely formulated problem. Soft computing methodologies (involving fuzzy sets, neural networks, genetic algorithms, and rough sets) are most widely applied in the data mining step of the overall KDD process. Fuzzy sets provide a natural framework for the process in dealing with uncertainty. Neural networks and rough sets are widely used for classification and rule generation. Genetic algorithms (GAs) are involved in various optimization and search processes, like query optimization and template selection. Other approaches like case based reasoning and decision trees are also widely used to solve data mining problems. The present article provides an over view of the available literature on data mining that is scarce, in the soft computing framework.

Keywords: Fuzzy sets, Fuzzy Logic, Neural Networks, Genetic Algorithms, Data Mining.

I. Introduction

Raw data is rarely of direct benefit. Its true value is predicated on the ability to extract information useful for decision support or exploration, and understanding the phenomenon governing the data source. In most domains, data analysis was traditionally a manual process. One or more analysts would become intimately familiar with the data and, with the help of statistical techniques, provide summaries and generate reports. In effect, the analyst acted as a sophisticated query processor. However, such an approach rapidly breaks down as the size of data grows and the number of dimensions increases. Databases containing number of data in the order 10 and dimension in the order of 10 are becoming increasingly common. When the scale of data manipulation, exploration and inferencing goes beyond human capacities, people look to computing technologies for automating the process. All these have prompted the need for intelligent data analysis methodologies, which could discover useful knowledge from data. The term KDD refers to the overall process of *knowledge discovery in databases*. *Data mining* is a particular step in this process, involving the application of specific algorithms for extracting patterns (models) from data. The additional steps in the KDD process, such as data preparation, data selection, data cleaning, incorporation of appropriate prior knowledge, and proper interpretation of the results of mining, ensures that useful knowledge is derived from the data. The subject of KDD has evolved, and continues to evolve, from the intersection of research from such fields as databases, machine learning, pattern recognition, statistics, artificial intelligence, reasoning with uncertainties, knowledge acquisition for expert systems, data visualization, machine discovery, and high-performance computing. KDD systems incorporate theories, algorithms, and methods from all these fields. Many successful applications have been reported from varied sectors

such as marketing, finance, banking, manufacturing, and telecommunications. Database theories and tools provide the necessary infrastructure to store, access and manipulate data. *Data warehousing*, a recently popularized term, refers to the current business trends in collecting and cleaning transactional data, and making them available for analysis and decision support. A good overview of KDD can be found in literatures. Fields concerned with inferring models from data include statistical pattern recognition, applied statistics, machine learning and neural computing. A natural question that arises is: how is KDD different from those fields? KDD focuses on the overall process of knowledge discovery from large volumes of data, including the storage and accessing of such data, scaling of algorithms to massive data sets, interpretation and visualization of results, and the modeling and support of the overall human machine interaction. Data mining is a form of knowledge discovery essential for solving problems in a specific domain. Individual data sets may be gathered and studied collectively for purposes other than those for which they were originally created. New knowledge may be obtained in the process while eliminating one of the largest costs, *viz.*, data collection. Medical data, for example, often exists in vast quantities in an unstructured format. The application of data mining can facilitate systematic analysis in such cases. Medical data, however, requires a large amount of preprocessing in order to be useful. Here numeric and textual information may be interspersed, different symbols can be used with the same meaning, redundancy often exists in data, erroneous/ misspelled medical terms are common, and the data is frequently rather sparse. A robust preprocessing system is required in order to extract any kind of knowledge from even medium-sized medical data sets. The data must not only be cleaned of errors and redundancy, but organized in a fashion that makes sense to the problem.

II. Data Mining Overview

Data mining refers to extracting or mining knowledge from large amounts of data. Nevertheless, mining is a vivid term characterizing the process that finds a small set of precious nuggets from a great deal of raw material. Knowledge discovery as a process is shown in Fig. 3.1. It consists of the following steps:

☐Data cleaning is to remove noise or irrelevant data

☐Data integration is where multiple data sources may be combined

☐Data selection where data relevant to the analysis task are retrieved from the database

☐ Data transformation where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance, data mining is an essential process where intelligent methods are applied in order to extract data patterns

☐Pattern evaluation is to identify the truly interesting patterns representing knowledge based on some interestingness measures

☐Knowledge presentation where visualization and knowledge representation techniques are used to present the mined knowledge to the user. In order to understand how and why data mining works, it’s necessary to understand the methods and tools of data mining which can be categorized as follows:

Characterization:

Data characterization is a traditional summarization of general features of objects in a target class, and produces what is called characteristic rules. The data relevant to a user-specified class are normally retrieved by a database query and run through a summarization module to extract the essence of the data at different levels of abstractions.

ISSN 2348 – 7968

Discrimination:

Data discrimination produces what are called discriminant rules. These are basically the comparison of the general features of objects between two classes referred to as the target class and the contrasting class.

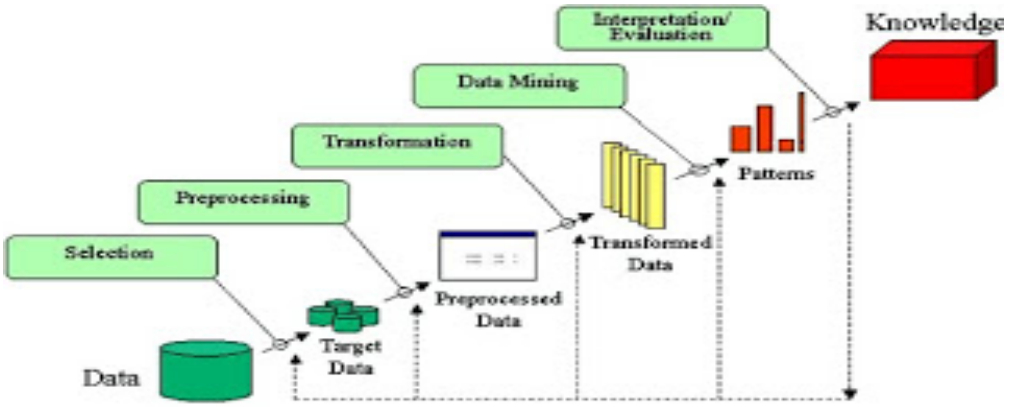


Fig.1 *Knowledge Discovery Process*

Association Rules:

It studies the frequency of items occurring together in transactional databases, and based on a threshold called support, identifies the frequent item sets.

Classification:

A general method for classification is to define a set of features that can be extracted from an item and then derive a model which, given the features of a particular item, can determine the (correct) classification. Also known as supervised classification, the classification uses given class labels to order the objects in the data collection. Classification approaches normally use a training set where all objects are already associated with known class labels. The classification algorithm learns from the training set and builds a model which is used to classify new objects.

Prediction:

Prediction has attracted considerable attention given the potential implications of successful forecasting in a business context. There are two major types of predictions: one can either try to predict some unavailable data values or pending trends, or predict a class label for some data. The latter is tied to classification. Prediction is however more often referred to the forecast of missing numerical values, or increase/ decrease trends in time related data. The major idea is to use a large number of past values to consider probable future values.

Clustering:

It is the process of grouping a set of physical or abstract objects into classes of similar objects. Clustering is also called unsupervised classification, because the classification is not dictated by given class labels. There are many clustering approaches all based on the principle of maximizing the similarity between objects in a same class (intra class similarity) and minimizing the similarity between objects of different classes (inter-class similarity).

Outlier Analysis:

ISSN 2348 – 7968

The analysis is to identify and explain exceptions. Outliers are data elements that cannot be grouped in a given class or cluster. Also known as exceptions or surprises, they are often very important to identify. While outliers can be considered noise and discarded in some applications, they can reveal important knowledge in other domains, and thus can be very significant and their analysis is valuable.

Visualization:

Visualization uses interactive graphs to demonstrate mathematically induced rules and scores, and is far more sophisticated than pie or bar charts. Visualization is used primarily to depict three dimensional geographic locations of mathematical coordinates.

Process of Data Mining:

For all these data mining algorithms and methods mentioned in the above, methodologies for data selection, cleaning, and transformation play a necessary and critical role. For data selection, data needs to extract from different databases and joined, and perhaps sampled. Once selected, the data may need to be cleaned. If the data is not derived from a warehouse but from disparate databases, values may be represented using different notations in the different databases. Also, certain values may require special handling since they may imply missing or unknown information. After the selection and cleaning process, certain transformations may be necessary. These range from conversions from one type of data to another, to deriving new variables using mathematical or logical formulae. Once the mining is done, visualization plays an important role in providing adequate bandwidth between the results of the data mining and the end user Fig.1.

III. An Overview of Soft Computing

Soft computing is a consortium of methodologies that works synergistically and provides, in one form or another, flexible information processing capability. Its aim is to exploit the tolerance for imprecision, uncertainty, approximate reasoning, and partial truth in order to achieve tractability, robustness, and low-cost solutions. Methodologies like fuzzy sets, neural networks, genetic algorithms, and rough sets are most widely applied in the data.

Soft Computing Methods (Fuzzy Logic):

The concept of fuzzy logic was initially conceived by Lotfi Zadeh. Fuzzy logic is an organized method for dealing with imprecise data. This data is considered to be as fuzzy sets. Fig.2. shows how values for the continuous attribute income are mapped into the discrete categories flow, medium, high, as well as how the fuzzy membership or truth values are calculated. Fuzzy logic systems typically provide graphical tools to assist users in this step.

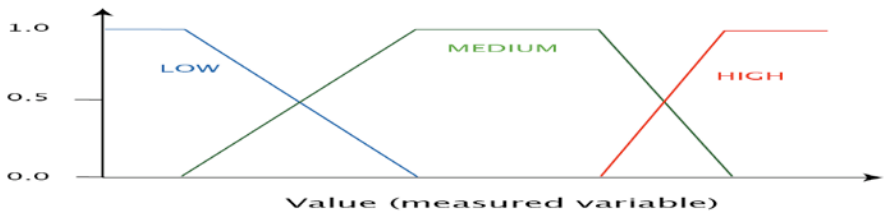


Fig.2 *Fuzzy Logic*

In general, the use of fuzzy logic in rule-based systems involves the following:

ISSN 2348 – 7968

☐Attribute values are converted to fuzzy values. Above figure shows how values for the continuous attribute income are mapped into the discrete categories flow, medium, high, as well as how the fuzzy membership or truth values are calculated. Fuzzy logic systems typically provide graphical tools to assist users in this step.

☐For a given new sample, more than one fuzzy rule may apply. Each applicable rule contributes a vote for membership in the categories. Typically, the truth values for each predicted category are summed.

☐The sums obtained above are combined into a value that is returned by the system. This process may be done by weighting each category by its truth sum and multiplying by the mean truth value of each category. The calculations involved may be more complex, depending on the complexity of the fuzzy membership graphs.

Neural Networks:

An artificial neural network (ANN), often just called a "neural network" (NN), is a mathematical model or computational model based on biological neural networks, in other words, is an emulation of biological neural system. It consists of an interconnected group of artificial neurons and processes information using a connectionist approach to computation. The human brain is a highly complex structure viewed as a massive, highly interconnected network of simple processing elements called neurons. This behavior of the neuron can be captured by a simple model which can be shown as:

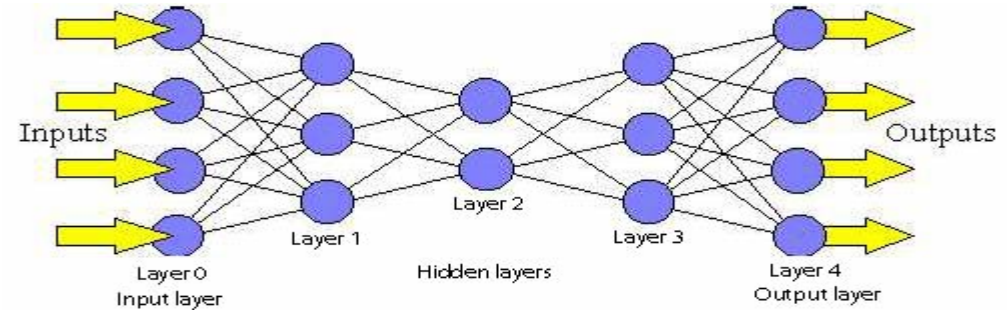


Fig.3 *Neural Network*

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. Every component of the model bears a direct analogy to the actual constituents of a biological neuron and hence is termed as ANN. There are several classes of Neural Networks, classified according to their learning mechanisms. Single layer feed forward network, Multilayer feed forward network, recurrent networks.

Genetic Algorithm:

Genetic algorithm (GA), belonging to a class of randomized heuristic and adaptive search techniques based on the principal of natural selection, is an attractive tool to find near optimal solutions for optimization problems. Genetic algorithms attempt to incorporate ideas of natural evolution. In general, genetic learning starts as follows. An initial population is created consisting of randomly generated rules. Each rule can be represented by a string of bits. As a simple example, suppose that samples in a given training set are described by two Boolean attributes, A1 and A2, and that there are two classes, C1 and C2. The rule \IF A1 and not A2 THEN C2" can be encoded as the bit string \100", where the two leftmost bits represent attributes A1 and A2, respectively, and the rightmost bit represents the class.

ISSN 2348 – 7968

Similarly, the rule \if not A1 and not A2 then C1" can be encoded as \001". If an attribute has k values where k > 2, then k bits may be used to encode the attribute's values. Classes can be encoded in a similar fashion. Based on the notion of survival of the fittest, a new population is formed to consist of the fittest rules in the current population, as well as offspring of these rules. Typically, the fitness of a rule is assessed by its classification accuracy on a set of training samples. Off spring are created by applying genetic operators such as crossover and mutation. In crossover, substrings from pairs of rules are swapped to form new pairs of rules. In mutation, randomly selected bits in a rule's string are inverted. The process of generating new populations based on prior populations of rules continues until a population P evolves where each rule in P satisfies a pre specified fitness threshold. Genetic algorithms are easily parallelizable and have been used for classification as well as other optimization problems. In data mining, they may be used to evaluate the fitness of other algorithms.

IV. Relevance of Soft-computing Methods in Data-mining

Each of the soft computing methods have their own characteristic, based upon which they can be suitably used in data mining process .Encapsulation of each of these methods in the data mining process has brought about a significant difference in the approach of information extraction and processing.

Neural Networks in Datamining:

Neural Network based data mining approach consists of three major phases:

Network Construction and Training:

This phase constructs and trains a three layer neural network based on the number of attributes and number of classes and chosen input coding method.

Network Pruning:

The pruning phase aims at removing redundant links and units without increasing the classification error rate of the network. A small number of units and links left in the network after pruning enable us to extract concise and comprehensible rules.

Rule Extraction:

This phase extracts the classification rules from the pruned network. The rules generated are in the form of "if *(a, Bv,) and* (x, *Bv,) and ... and* ***(x,*** *Bv,)* then *Cy* where *a,s* are the attributes of an input tuple, *v* ,are constants & are relational operators (=, <, *2*, <>), and ***Ci*** is one of the class labels.

The nns exhibit mapping capabilities that is they can map input patterns to their associated output patterns. The nns learn by examples, thus nn architecture can be trained with known examples of a problems before they are tested for their inference capabilities on unknown instances of the problem, they can therefore identify new objects previously untrained. The nns posses the capability to generalize, thus they can predict new outcomes from past trends. The nns are robust systems and are fault tolerant. They can therefore recall full patterns from incomplete partial or noisy patterns. Based upon the above characteristics of nn which are very closely associated with the functionality what is required by the data mining application. They can be efficiently embodied with data mining methods for increasing efficiency of the outcome of different data mining techniques. Data mining, cleaning and validation could be achieved by determining

ISSN 2348 – 7968

which records suspiciously diverge from the patterns of their peers. Hence for this proper approaches for combining the ANN and data mining technologies should be found to improve and optimize data mining technology. Fuzzy neural networks and self-organizing neural networks are gaining fast importance in the field of data mining. Neural networks have found wide application in areas such as pattern recognition, image processing, optimization, fore casting.

Fuzzy Logic in Datamining:

Since fuzzy sets allow partial membership, Fuzzy logic is basically multivalued logic that allows intermediate values to be defined between conventional evaluations such as yes/no, true/false etc. This represents a more human like thinking approach in the programming of computer. Since in data mining with large data bases the most common challenges are the noisy, imprecise, vague data, therefore by suitable extracting the relevant characteristics of fuzzy sets the data mining techniques can be made more efficient. The use of fuzzy techniques has been considered to be one of the key components of data mining systems because of the affinity with human knowledge representation. Wei and Chen have mined generalized association rules with fuzzy taxonomic structures. Fuzzy logic systems have been used in numerous areas for classification, including health care and finance. Fuzzy logic is useful for data mining systems performing classification. It provides the advantage of working at a high level of abstraction. Decision making is very important in data mining which involves social, economic and scientific applications. At this junction fuzzy data mining comes as great help to data miners.

Genetic Algorithm in Datamining:

Genetic algorithm plays an important role in data mining technology, which is decided by its own characteristics and advantages. To sum up, mainly in the following aspects:

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.

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.

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.

Genetic algorithm doesn't adopt deterministic rules, but adopts the rules of probability changing to guide search direction. Genetic algorithm has been efficiently used in multimedia databases. Knowledge discovery systems have been developed using genetic programming concepts.

Knowledge Discovery and Data mining

KDD is defined as *the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data*. *Data* is a set of facts F, and a *pattern* is an expression E in a language L describing the facts in a subset F_E of F.E is called a pattern if it is simpler than the enumeration of all facts in F_E .A measure of certainty, measuring the *validity* of discovered patterns, is a function C mapping expressions in L to a partially or totally ordered measure space M_C. An expression E in L about a subset F_E belongs to F can be assigned a certainty measure c=(E,F). *Novelty* of patterns can be measured by a function N(E,F) with respect to changes in data or knowledge.

ISSN 2348 – 7968

Patterns should potentially lead to some *useful* actions, as measured by some utility function U(E,F) mapping expressions in L to a partially or totally ordered measure space M_U . The goal of KDD is to make patterns *understandable* to humans. This is measured by a function s=S(E,F) mapping expressions E in L to a partially or totally ordered measure space M_S. *Interestingness* of a pattern combines validity, novelty, usefulness, and understandability, and can be expressed as i=I(E,F,C,U,N,S) which maps expressions in L to a measure space M_I. A pattern Eϵ L is called *knowledge* if for some user-specified threshold iϵM_I, I(E,F,C,U,N,S) >I [8]. One can select some thresholds cϵM_C,sϵM_S, and ,uϵM_U and term a pattern E knowledge

IF C(E,F)>c AND S(E,F)>s AND U(E,F)>u

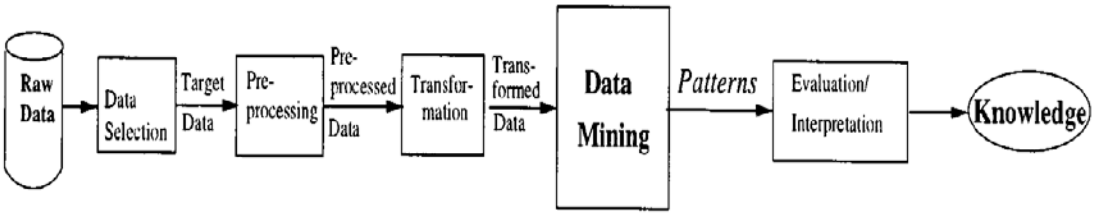
The role of interestingness is to threshold the huge number of discovered patterns and reports only those which may be of some use. There are two approaches to designing a measure of interestingness of a pattern, *viz.*, objective and subjective. The former uses the structure of the pattern and is generally used for computing *rule interestingness*. However often it fails to capture all the complexities of the pattern discovery process. The *subjective* approach, on the other hand, depends additionally on the *user* who examines the pattern. Two major reasons why a pattern is interesting from the subjective (user-oriented) point of view are as follows.

• *Unexpectedness*: when it is 'surprising' to the user.

• *Actionability*: when the user can act on it to her/his advantage.

Though both these concepts are important is has often been observed that actionability and unexpectedness are correlated. In literature, unexpectedness is often defined in terms of the dissimilarity of a discovered pattern from a vocabulary provided by the user.

Data mining is a step in the KDD process consisting of a particular enumeration of patterns E_j over the data, subject to some computational limitations. It uses *historical* data to discover regularities and improve future decisions. The data can consist of (say) a collection of time series descriptions that can be learned to predict later events in the series.



A. KDD Process

The overall KDD process is outlined in Fig.1. It is interactive and iterative involving, more or less, the following steps.

1) Understanding the application domain: includes relevant prior knowledge and goals of the application.

2) Extracting the target data set: includes selecting a data set or focusing on a subset of variables.

ISSN 2348 – 7968

3) Data cleaning and preprocessing: includes basic operations, such as noise removal and handling of missing data. Data from real-world sources are often erroneous, incomplete, and inconsistent, perhaps due to operation error or system implementation flaws. Such low quality data needs to be cleaned prior to data mining.

4) Data integration: includes integrating multiple, heterogeneous data sources.

5) Data reduction and projection: includes finding useful features to represent the data (depending on the goal of the task) and using dimensionality reduction or transformation methods.

6) Choosing the function of data mining: includes deciding the purpose of the model derived by the data mining algorithm (e.g., summarization, classification, regression, clustering, web mining, image retrieval, discovering association rules and functional dependencies, rule extraction, or a combination of these).

7) Choosing the data mining algorithm(s): includes selecting method(s) to be used for searching patterns in data, such as deciding on which model and parameters may be appropriate.

8) Data mining: includes searching for patterns of interest in a particular representational form or a set of such representations.

9) Interpretation: includes interpreting the discovered patterns, as well as the possible visualization of the extracted patterns. One can analyze the patterns automatically or semi-automatically to identify the truly interesting/ useful patterns for the user.

10) Using discovered knowledge: includes incorporating this knowledge into the performance system, taking actions based on knowledge.

B. Data Mining

KDD refers to the overall process of turning low-level data into high-level knowledge. An important step in the KDD process is data mining. Data mining is an interdisciplinary field with a general goal of predicting outcomes and uncovering relationships in data. It uses automated tools employing sophisticated algorithms to discover hidden patterns, associations, anomalies and/or structure from large amounts of data stored in data warehouses or other information repositories. Data mining tasks can be descriptive, i.e., discovering interesting patterns describing the data, and predictive, i.e., predicting the behavior of the model based on available data.

Data mining involves fitting models to or determining patterns from observed data. The fitted models play the role of inferred knowledge. Deciding whether the model reflects useful knowledge or not is a part of the overall KDD process for which subjective human judgment is usually required. Typically, a data mining algorithm constitutes some combination of the following three components.

• ***The model:*** The function of the model (e.g., classification, clustering) and its representational form (e.g., linear discriminants, neural networks). A model contains parameters that are to be determined from the data.

• ***The preference criterion:*** A basis for preference of one model or set of parameters over another, depending on the given data. The criterion is usually some form of goodness- of-fit function of the model to the data, perhaps

ISSN 2348 – 7968

tempered by a smoothing term to avoid over fitting, or generating a model with too many degrees of freedom to be constrained by the given data.

• ***The search algorithm:*** The specification of an algorithm for finding particular models and parameters, given the data, model(s), and a preference criterion. A particular data mining algorithm is usually an instantiation of the model/preference/search components.

Development of new generation algorithms is expected to encompass more diverse sources and types of data that will support mixed-initiative data mining, where human experts collaborate with the computer to form hypotheses and test them. The notion of *interestingness*, which encompasses several features such as validity, novelty, usefulness, and simplicity, can be quantified through fuzzy sets. Fuzzy dissimilarity of a discovered pattern with a user-defined vocabulary has been used as a measure of this interestingness. As an extension to the above methodology *unexpectedness* can also be defined in terms of a *belief system*, where if a belief b is based on previous evidence ξ then d(b/ξ) denotes the degree of belief b

. In soft belief systems, a weight ω_i is attached to each belief b_i

. The degree of a belief may be measured with conditional probability, Dempster-Shafer belief function or frequency of the raw data. Here, the interestingness of a pattern E relative to a belief system B and evidence may be formally defined as

$$I(E, B, \xi) = \sum_{b_i \in B} w_i \, d(b_i \mid E, \xi) - d(b_i \ \ \xi) \, .$$

This definition of interestingness measures the amount by which the degrees of belief change as a result of a new pattern. There is a growing indisputable role of fuzzy set technology in the realm of data mining. Various data browsers have been implemented using fuzzy set theory. Analysis of real-world data in data mining often necessitates simultaneous dealing with different types of variables, *viz.*, categorical/symbolic data and numerical data. Nauck has developed a learning algorithm that creates *mixed* fuzzy rules involving both categorical and numeric attributes. Pedrycz discusses some constructive and fuzzy set-driven computational vehicles of knowledge discovery, and establishes the relationship between data mining and fuzzy modeling. The role of fuzzy sets is categorized below based on the different functions of data mining that are modeled.

1) Clustering:

Data mining aims at sifting through large volumes of data in order to reveal useful information in the form of new relationships, patterns, or clusters, for decision-making by a user. Fuzzy sets support a focused search, specified in linguistic terms, through data. They also help discover dependencies between the data in qualitative/semi-qualitative format. Researchers have developed fuzzy clustering algorithms for this purpose. Russell and Lodwick have explored fuzzy clustering methods for mining telecommunications customer and prospect databases to gain residential and business customer market share. Pedrycz has designed fuzzy clustering algorithms using 1) contextual information and 2) induced linguistic space for better focusing of the search procedure in KDD. Mazlack suggests a converse approach of progressively reducing the data set by partitioning and eliminating the least important attributes to reduce intra item dissonance within the partitions. Wei and Chen have mined generalized association rules with fuzzy taxonomic

ISSN 2348 – 7968

structures. A crisp taxonomy assumes that a child belongs to its ancestor with degree one. A fuzzy taxonomy is represented as a directed acyclic graph, each of whose edges represents a fuzzy *IS-A* relationship with degree $\mu(0 \leq \mu \leq 1)$. The partial belonging of an item in a taxonomy is taken into account while computing the degrees of support and confidence. Au and Chan utilize an *adjusted difference* between observed and expected frequency counts of attributes for discovering fuzzy association rules in relational databases. Instead of dividing quantitative attributes into fixed intervals, they employ linguistic terms to represent the revealed regularities and exceptions. The algorithm allows one to discover both *positive* and *negative* rules, and can deal with fuzzy class boundaries as well as missing values in databases. Summary discovery is one of the major components of knowledge discovery in databases. This provides the user with comprehensive information for grasping the essence from a large amount of information in a database. Fuzzy set theory is also used for data summarization. Linguistic summaries of large sets of data are derived as linguistically quantified propositions with a degree of validity. This corresponds to the preference criterion involved in the mining task. Chiang *et al.* have used fuzzy linguistic summary for mining time series data. Mining typical user profiles and URL associations from the vast amount of access logs is an important component of Web personalization that deals with tailoring a user's interaction with the Web information space based on information about him/her. Recent increase in the size of *multimedia* information repositories, consisting of mixed media data, has made content-based image retrieval (CBIR) an active research area. Unlike traditional database techniques which retrieve images based on exact matching of keywords, CBIR systems represent the information content of an image by visual features such as color, texture, and shape, and retrieve images based on similarity of features. Frigui has developed an *interactive* and *iterative* image retrieval system that takes into account the *subjectivity* of human perception of visual content. Neural networks were earlier thought to be unsuitable for data mining because of their inherent *black-box* nature. No information was available from them in symbolic form, suitable for verification or interpretation by humans. Recently there has been widespread activity aimed at redressing this situation, by extracting the embedded knowledge in trained networks in the form of symbolic rules. This serves to identify the attributes that, either individually or in a combination, are the most significant determinants of the decision or classification. Unlike fuzzy sets, the main contribution of neural nets toward data mining stems from rule extraction and clustering.

2) Rule Extraction: In general, the primary input to a connectionist rule extraction algorithm is a representation of the trained neural network, in terms of its nodes, links and sometimes the data set. One or more hidden and output units are used to automatically derive the rules, which may later be combined and simplified to arrive at a more comprehensible rule set.

These rules can also provide new insights into the application domain. The use of neural nets helps in 1) incorporating parallelism and 2) tackling optimization problems in the data domain. The models are usually suitable in *data-rich* environments. Typically a network is first trained to achieve the required accuracy rate. Redundant connections of the network are then removed using a pruning algorithm. The link weights and activation values of the hidden units in the network are analyzed, and classification rules are generated.

3) Rule Evaluation: Here we provide some quantitative measures to evaluate the performance of the generated rules. This relates to the *preference criteria/goodness of fit* chosen for the rules. Let N be an l x l matrix whose (i,j)th element n_{ij} indicates the number of patterns actually belonging to class , but classified as class j.

• *Accuracy*: It is the correct classification percentage, provided by the rules on a test set defined as n_{ic}/n_i x 100, where n_i is equal to the number of points in class i and n_{ic} of these points are correctly classified.

ISSN 2348 – 7968

• *User's accuracy*: If n'$_i$ points are found to be classified into class i, then the user's accuracy (U)is defined as $U = n_{ic}/n_i'$.

• *Kappa*: The kappa value for class i(K_i) is defined as

$$K_i = \frac{n \cdot n_{ic} - n_i \cdot n_i'}{n \cdot n_i' - n_i \cdot n_i'}.$$

The numerator and denominator of overall kappa are obtained by summing the respective numerators and denominators of K_i separately over all classes.

• *Fidelity*: It is measured as the percentage of the test set for which network and the rulebase output agree.

• *Confusion*: This measure quantifies the goal that the "*confusion should be restricted within minimum number of classes*". Let be the mean of all for . Then

$$\text{Conf} = \frac{\text{Card}\{n_{ij} : n_{ij} \geq \hat{n}_{ij}, i \neq j\}}{l}$$

for an class l problem.

• *Coverage*: The percentage of examples from a test set for which no rules are fired is used as a measure of the uncovered region. A rule base having a smaller uncovered region is superior.

• *Rule base size*: This is measured in terms of the number of rules. The lower its value, the more compact is the rule base.

• *Computational complexity*: This is measured in terms of the CPU time required.

• *Confidence*: The confidence of the rules is defined by a confidence factor cf . We have

$$cf_j = \inf_{j_i \text{ all nodes in the path}} \frac{(\Sigma_i w_{ji} - \theta_j)}{\Sigma_i w_{ji}}$$

Where ω_{ji} is the th incoming link weight to node j and θ_j is its threshold.

V. Data Mining and Neuro–Fuzzy System

An Interval Type-2 Fuzzy Neural Network (IT2FNN) are used for automatically generate the necessary rules. The phase of data mining using Interval Type-2 Fuzzy Logic Systems (IT2FLS) Castillo et al. (2010); Castro et al. (2010) becomes complicated, as there are enough rules to determine which variables one should take into account. The search method of back-propagation and hybrid learning (BP+RLS) is more efficient in other methods, such as genetic algorithms Rantala and Koivisto (2002); Castro et al. (2008).

ISSN 2348 – 7968

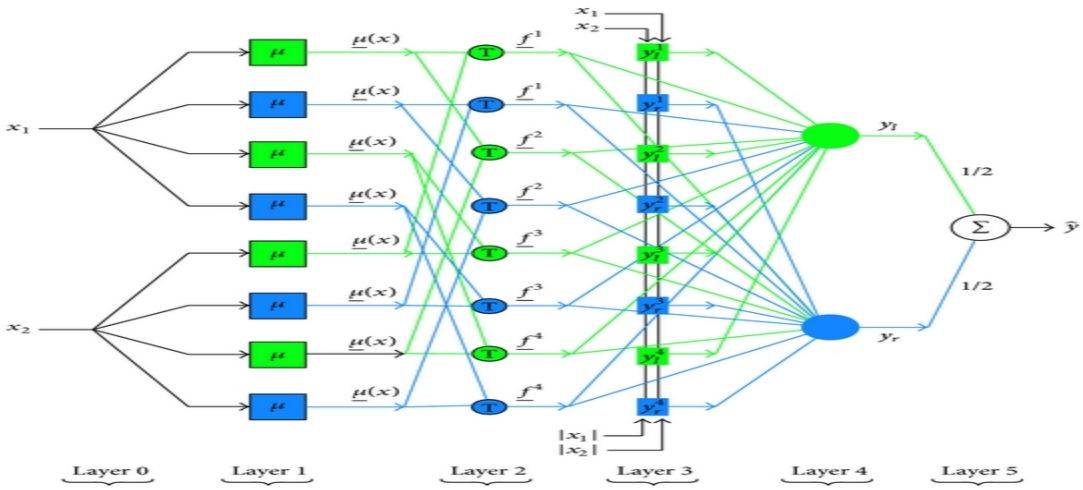


Fig.3 Generation of the necessary rules using an Interval Type–2 Fuzzy Neural Network (IT2FNN)

Since the IT2FNN method seems to produce more accurate models with fewer rules is widely used as a numerical method to minimize an objective function in a multidimensional space, find the approximate global optimal solution to a problem with N variables, which minimize the function varies smoothly Stefanescu (2007).

With the application of this grouping algorithm we obtain the rules, the agent receives input data from its environment and choose an action in an autonomous and flexible way to fulfill its function Peng et al. (2008).

VI. Neuro-fuzzy inference system

Using the neuro-fuzzy system for the automatic generation of rules, this phase of the data extraction from the data may become complicated, as the process needs to appropriately establish the number of sufficient norms and variables that the study needs to take into account. Using this grouping algorithm, we obtain the appropriate rule-set assigned to each agent representing a inhabitant of it, the agent receives inputs from its geographical environment and in turn much choose an action in an autonomous and flexible fashion Gilbert (2007); Drennan (2005);Wooldridge and Jennings (1995).

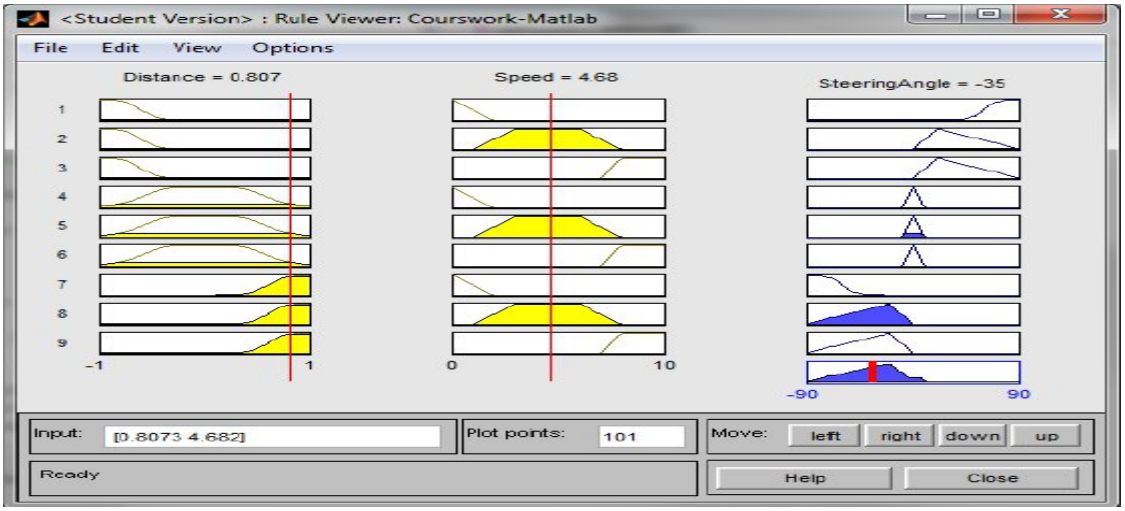


Fig.4 Rules on a Type-2 Fuzzy Inference System.

ISSN 2348 – 7968

The purpose of this structure without central control is to garner agents that are created with the least amount of exogenous rules and to observe the behavior of the global system through the interactions of its existing interactions, such that the system, by itself, generates an intelligent behavior that is not necessarily planned in advance or defined within the agents themselves; in other words, creating a system with truly emergent behavior Botti and Julian (2003); Russell and Norvig (2004). From the 2010 census information, we create a Type-2 Fuzzy Inference System as how we could represent different agencies as a decision-making system into agents. City level Type-2 Fuzzy Inference Systems The figure 5 shows a type-2 fuzzy inference system for Tijuana city. It depicts a set of input-output variables and a rule set. Output variables are catholic and non-catholic as a response of the system. We could use the difference between both values to make decisions into an agent as a preference decision-making system.

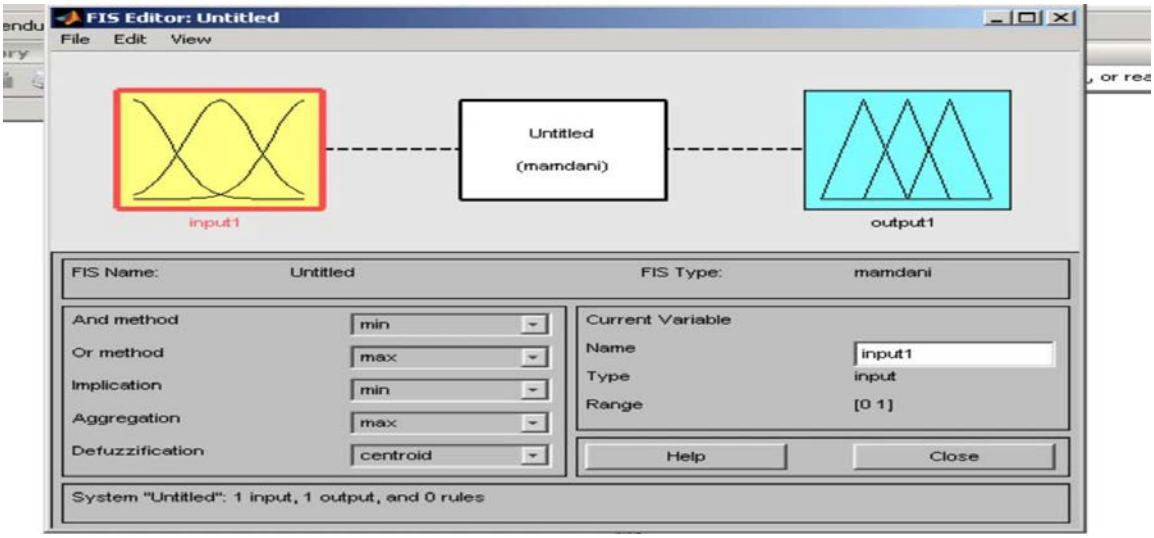


Fig. 5 Fuzzy Inference System for Tijuana City

The Figure 5 depicts the resolution example of the rules by the fuzzy inference system. Different quantitative input values could be introduced and the system resolve creating different responses. Depending of the combination of inputs, we can expect different responses of the system. An agent will use this inference system as a decision-making system to show different behaviours depending of the situation.

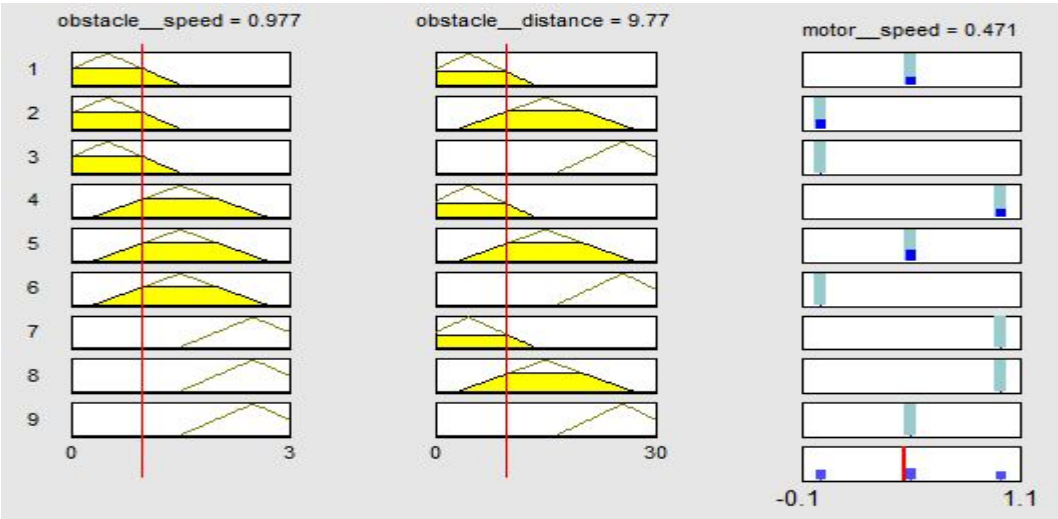


Fig.6 Fuzzy Inference System Rule Set Evaluation for Tijuana City.

ISSN 2348 – 7968

VII. Conclusion and Discussion

Current research in data mining mainly focuses on the discovery algorithm and visualization techniques. There is a growing awareness that, in practice, it is easy to discover a huge number of patterns in a database where most of these patterns are actually obvious, redundant, and useless or uninteresting to the user. To prevent the user from being overwhelmed by a large number of uninteresting patterns, techniques are needed to identify only the useful/interesting patterns and present them to the user. Soft computing methodologies, involving fuzzy sets, neural networks, genetic algorithms, rough sets, and their hybridizations, have recently been used to solve data mining problems. They strive to provide approximate solutions at low cost, thereby speeding up the process. A categorization has been provided based on the different soft computing tools and their hybridizations used, the mining function implemented, and the preference criterion selected by the model. Fuzzy sets, which constitute the oldest component of soft computing, are suitable for handling the issues related to understandability of patterns, incomplete/noisy data, mixed media information and human interaction, and can provide approximate solutions faster. They have been mainly used in clustering, discovering association rules and functional dependencies, summarization, time series analysis, web applications, and image retrieval. Neural networks are suitable in data-rich environments and are typically used for extracting embedded knowledge in the form of rules, quantitative evaluation of these rules, clustering, self-organization, classification and regression. Neuro-fuzzy hybridization exploits the characteristics of both neural networks and fuzzy sets in generating natural/linguistic rules, handling imprecise and mixed mode data, and modeling highly nonlinear decision boundaries. Domain knowledge, in natural form, can be encoded in the network for improved performance.

Recently, several commercial data mining tools have been developed based on soft computing methodologies. These include Data Mining Suite, using fuzzy logic; 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.

Soft computing methods are fundamentally used for information processing by employing methods, which are skilled to deal with imprecision, vagueness and uncertainty, needed in application areas where the problems are not hazy. The outcomes are exact within the error bounds estimated where as in the case of soft computing they are approximate and in some instances they may be interpreted as outcomes from an intelligent behavior. From these basic properties, it may be concluded that, both paradigms have their own merits and by observing these merits synergistically these paradigms can be used in a complementary way for knowledge discovery in databases.

Reference

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. Osmar R. Zaïane: “Principles of Knowledge Discovery in Databases - Chapter 8: Data Clustering”.
6. Jianxiong Luo, “integrating fuzzy logic with data mining methods for intrusion detection, 1999.
7. 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.
8. Xianjun Ni,” Research of Data Mining Based on Neural Networks “ , World Academy of Science, Engineering and Technology 39 2008
9. Yashpal Singh etal,” Neural networks in data mining”, Journal of Theoretical and Applied Information Technology, 2009
