Data Mining in the Soft Computing Paradigm

Pradip Kumar Bala
IBAT, Deemed University, India

Shamik Sural
Indian Institute of Technology, Kharagpur, India

Rabindra Nath Banerjee
Indian Institute of Technology, Kharagpur, India

INTRODUCTION

Data mining is a set of tools, techniques and methods that can be used to find new, hidden or unexpected patterns from a large volume of data typically stored in a data warehouse. Results obtained from data mining help an organization in more effective individual and group decision-making. Regardless of the specific technique, data mining methods can be classified by the function they perform or by their class of application.

Association rule is a type of data mining that correlates one set of items or events with another set of items or events. It employs association or linkage analysis, searching transactions from operational systems for interesting patterns with a high probability of repetition. Classification techniques include mining processes intended to discover rules that define whether an item or event belongs to a particular predefined subset or a class of data. This category of techniques is probably the most broadly applicable to different types of business problems. In some cases, it is difficult to define the parameters of a class of data to be analyzed. When parameters are elusive, clustering methods can be used to create partitions so that all members of each set are similar according to a specified set of metrics. Summarization describes a set of data in compact form. Regression techniques are used to predict a continuous value. The regression can be linear or non-linear with one predictor variable or more than one predictor variables, known as multiple regression.

Soft computing, which includes application of fuzzy logic, neural network, rough set and genetic algorithm, is an emerging area in data mining. By studying combinations of variables and how different combinations affect data sets, we develop neural network, a non-linear predictive model that "learns." Machine learning techniques, such as genetic algorithms and fuzzy logic, can derive meaning from complicated and imprecise data. They can extract patterns from and detect trends within the data that are far too complex to be noticed by either humans or more conventional automated analysis techniques. Because of this ability, neural computing and machine learning technologies demonstrate broad applicability in the world of data mining and, thus, to a wide variety of complex business problems. Rough set is the approximation of an imprecise and uncertain set by pair of precise concepts, called the lower and upper approximations.

Each soft computing technique addresses problems in its domain using a distinct methodology. However, they are not substitute of each other. In fact, these soft computing tools work in a cooperative manner, rather than being competitive. This has led to the development of hybridization of soft computing tools for data mining applications (Mitra et al., 2002). It should, however, be kept in mind that soft computing techniques have been traditionally developed to handle small data sets. Extending the soft computing paradigm for processing large volumes of data is itself a challenging task. In the next section, we give a brief background of the various soft computing techniques.

BACKGROUND

Fuzzy rules offer an attractive trade-off between the need for accuracy and compactness on one hand, and scalability on the other, when reasoning systems within a particular knowledge domain become quite complex. Fuzzy rules generalize the concept of categorization because, by definition, the same object can belong to multiple sets with different degrees of membership. In this sense, fuzzy logic eliminates the problems associated with borderline cases: where, for example, a value of degree of membership with 0.9 may cause a rule to fire but a value 0.899 may not. The net result is that fuzzy systems tend to provide greater accuracy than traditional rule-based systems when continuous variables are involved.

Neural network, which draws its inspiration from neuroscience, attempts to mirror the way a human brain works

in recognizing patterns by developing mathematical structures with the ability to learn. An artificial neural network (ANN) learns through training. These are simple computer-based programs whose primary function is to construct models of a problem space based on trial and error. The process of training a neural net to associate certain input patterns with correct output responses involves the use of repetitive examples and feedback, much like the training of a human being.

Rough set theory finds application in studying imprecision, vagueness, and uncertainty in data analysis and is based on the establishment of equivalence classes within a given training data. A rough set gives an approximation of a vague concept by two precise concepts, called the lower and upper approximations. These two approximations are a classification of the domain of interest into disjoint categories. The lower approximation is a description of the domain objects known with certainty to belong to the subset of interest and upper approximation is a description that may possibly belong to the subset.

Genetic algorithms (GAs) are computational models used in efficient and global search methods for optimality in problem solving. These search algorithms are based on the mechanics of natural genetics theory combined with Darwin's theory of "survival of the fittest" and are particularly suitable for solving complex optimization problems as well as applications that require adaptive problem-solving strategies. In data mining, GA finds application in hypothesis testing and refinement.

With this background, we next present how soft computing techniques can be applied to specific data mining problems.

MAIN THRUST

Association Rule Mining: Following are some of the applications of soft computing tools in association rule mining.

- **Fuzzy Logic:** A generalized association rule may involve binary, quantitative or categorical data and hierarchical relation. In quantitative or categorical association rule mining, irrespective of the methodology used, "sharp boundaries" remain a problem which under-estimates or over-emphasizes the elements near the boundaries. This, may, therefore, lead to an inaccurate representation of semantics. To deal with the problem, fuzzy sets and fuzzy items, usually in the form of labels or linguistic terms, are used and defined onto the domains (Chien et al., 2001). In the fuzzy framework, conventional notions of support and confidence could be extended as well. The partial belongingness of an item in a subset is taken into account while computing the degree of support and the degree of confidence. The measures are similar in spirit to the count operator used for fuzzy cardinality. Subsequently, with these extended measures incorporated, several mining algorithms have been developed (Gyenesei, 2000; Gyenesei & Teuhola, 2001; Shu et al., 2001). Instead of dividing quantitative attributes into fixed intervals, linguistic terms can be used to represent the regularities and exceptions the way in which humans perceive the reality. Chen et al. (2002) have developed an algorithm for fuzzy association rules in dealing with partitioning quantitative data domains. Wei & Chen (1999) extended generalized association rules with fuzzy taxonomies, by which partial belongings could be incorporated. Furthermore, a recent effort has been made which incorporates linguistic hedges on existing fuzzy taxonomies (Chen et al., 1999; Chen et al., 2002a). Several fuzzy extensions have been made on interestingness measures. A measure called "Interestingness Degree" has been proposed which can be seen as the increase in probability of an event Y caused by the occurrence of another event X. Attempts have been made to introduce thresholds for filtering databases in dealing with very low membership degrees (Hullermeier, 2001).
- **Genetic Algorithm:** Min et al. (2001) have used a GA-based data mining approach in e-commerce to find association rules of IF-THEN form for adopters and non-adopters of e-purchasing. Association rules of the form IF-THEN form can also be mined, which provides a high degree of accuracy and coverage (Lopes et al., 1999).
- **Clustering:** Following are some of the applications of soft computing tools in clustering.
- **Fuzzy Logic:** A fuzzy clustering algorithm makes an attempt to group the prospects into categories based on their identifying characteristics. For example, for prospective customers of any business, the key attributes can include geographic data, psychographic data and others. Clusters expressed in linguistic terms can be easily handled using fuzzy sets. Using fuzzy sets, we can also find dependencies between data expressed in qualitative format. Use of fuzzy logic can help in avoiding searching for less important, trivial or meaningless patterns in databases. Fuzzy clustering algorithms have been developed for mining telecommunications customer and prospect database to gain customer information for deciding a marketing strategy (Russell et al., 1999).
- **Neural Network:** Self Organizing Map (SOM) is one of the most widely used unsupervised neural network models that employ competitive learning steps.