Soft Computing Techniques and its Impact in Data Mining

Y. K. Mathur¹, Abhaya Nand ²

¹Univ. of Delhi, ²PDMCE, Bahadurgarh, India

Keywords—Computational Intelligence, Data Mining, Information Technology, Maintenance, Portability, Reliability, Reusability, Soft Computing, Web Mining.

Abstract— With the rapid advancement in the field of Information Technology Software organizations are trying to use the strength of various emerging and powerful techniques such as Soft Computing techniques to develop/produce software products at minimum cost with maximum quality in terms of reliability, portability, accessibility, maintainability etc. In fact Soft Computing techniques and its impact as well as its new emerging trends to suit the changing requirements in the area of Data Mining have become the most important topic in recent times. These techniques are used widely for varieties of applications. The need for web intelligence, Internet based applications and the current research on soft web mining in recent times demanded our immediate attention for the development of much more capable and future generation Software systems and services. This includes emerging topics like data mining, Knowledge engineering, natural language processing, Computational intelligence, E-commerce, Bioinformatics and Cognitive computing. Development of such systems helps in addressing problems across different fields and assists in reviewing current progress in the field of Soft Computing and information processing. In the following section, the evolution of Soft Computing techniques, its development & application as well as its impact in Data Mining have been highlighted.

I. INTRODUCTION

This paper provides a survey of the concept of data mining using Soft Computing techniques. A categorization has been provided based on the different soft computing tools and their hybridizations used, the data mining function implemented, and the preference criterion selected by the model. The utility of the different soft computing methodologies is highlighted. Generally fuzzy sets 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. Neural networks are nonparametric, robust, and exhibit good learning and generalization capabilities in data-rich environments. Genetic algorithms provide efficient search algorithms to select a model, from mixed media data, based on some preference criterion/objective function.

Rough sets are suitable for handling different types of uncertainty in data. Some challenges to data mining and the application of soft computing methodologies are indicated. The Digital revolution has made digitized information easy to capture and fairly inexpensive to store. With the development of computer hardware and software and the rapid computerization of business, huge amount of data have been collected and stored in databases. The rate at which such data is stored is growing at a phenomenal rate. As a result, traditional ad hoc mixtures of statistical techniques and data management tools are no longer adequate for analysing this vast collection of data.

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.

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, and knowledge acquisition for expert systems, data visualization, machine discovery and high-performance computing. KDD systems incorporate theories, algorithms, and methods from all these fields.
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.

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.

This paper gives an overview of the available literature on data mining that is scarce, in the soft computing framework. It also describes the basic notions of knowledge discovery in databases, and data mining. Some challenges are highlighted. This is followed by a survey explaining the role of the aforesaid soft computing tools and their hybridizations, categorized on the basis of the different data mining functions implemented and the preference criterion selected by the model. The utility and applicability of the different soft computing methodologies is highlighted. Some challenges to data mining and the application of soft computing methodologies are also indicated.

II. Objectives

Recently various soft computing methodologies have been applied to handle the different challenges posed by data mining. The main constituents of soft computing, at this juncture, 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.

Let us first describe the roles and significance of the individual soft computing tools and their hybridizations, followed by the various systems developed for handling the different functional aspects of data mining. A suitable preference criterion is often optimized during mining.

It may be mentioned that there is no universally best data mining method; choosing particular soft computing tool(s) or some combination with traditional methods is entirely dependent on the particular application and requires human interaction to decide on the suitability of an approach.

A. Fuzzy Sets

The modeling of imprecise and qualitative knowledge, as well as the transmission and handling of uncertainty at various stages are possible through the use of fuzzy sets. Fuzzy logic is capable of supporting, to a reasonable extent, human type reasoning in natural form. It is the earliest and most widely reported constituent of soft computing. The development of fuzzy logic has led to the emergence of soft computing. Fuzzy models can be said to represent a prudent and user-oriented sifting of data, qualitative observations and calibration of commonsense rules in an attempt to establish meaningful and useful relationships between system variables. Despite a growing versatility of knowledge discovery systems, there is an important component of human interaction that is inherent to any process of knowledge representation, manipulation and processing. Fuzzy sets are inherently inclined toward coping with linguistic domain knowledge and producing more interpretable solutions.

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 to discover dependencies between the data in qualitative/semi-qualitative format. In data mining, one is typically interested in a focused discovery of structure and an eventual quantification of functional dependencies existing therein. This helps to prevent searching for meaningless or trivial patterns in a database. 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.

2) Association Rules: An important area of data mining research deals with the discovery of association rules. An association rule describes an interesting association relationship among different attributes.
A Boolean association involves binary attributes, a
generalized association involves attributes that are
hierarchically related, and a quantitative association
involves attributes that can take on quantitative or
categorical values.

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.

3) Functional Dependencies: Fuzzy logic has been used for
analysing inference based on functional dependencies
(FDs), between variables, in database relations. Fuzzy
inference generalizes both imprecise (set-valued) and
precise inference. Similarly, fuzzy relational databases
generalize their classical and imprecise counterparts by
supporting fuzzy information storage and retrieval.
Inference analysis is performed using a special abstract
model which maintains vital links to classical, imprecise
and fuzzy relational database models. These links increase
the utility of the inference formalism in practical
applications involving “catalytic inference analysis,”
including knowledge discovery and database security.

4) Data Summarization: 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. Typically, fuzzy sets are used for an
interactive top-down summary discovery process which
utilizes fuzzy IS-A hierarchies as domain knowledge. Here
generalized tuples are used as a representational form of a
database summary including fuzzy concepts. By virtue of
fuzzy IS-A hierarchies, where fuzzy IS-A relationships
common in actual domains are naturally expressed, the
discovery process comes up with more accurate database
summaries.

5) Web Application: 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. Nasraoui et al.
have defined a user session as a temporally compact
sequence of Web accesses by a user and used a
dissimilarity measure between two Web sessions to capture
the organization of a Web site. Their goal is to categorize
these sessions using Web mining.

6) Image Retrieval: 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 [65].

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 [66]
has developed an interactive and iterative image retrieval
system that takes into account the subjectivity of human
perception of visual content. The feature relevance weights
are learned from the user’s positive and negative feedback,
and the Choquet integral is used as a dissimilarity measure.
The smooth transition in the user’s feedback is modeled by
continuous fuzzy membership functions.

B. Neural Networks

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 Unlike fuzzy sets, the main contribution of neural nets
toward data mining stems from rule extraction and
clustering.

1) 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.

2) 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 \( l \times l \) matrix whose
\((i,j)\)th element \( n_{ij} \) indicates the number of patterns
actually belonging to class \( i \), but classified as class \( j \).

• Accuracy: It is the correct classification percentage,

\[
\left( \frac{n_{ii}}{n_i} \right) \times 100.
\]

where \( n_{ii} \) is equal to the number of points in class \( i \) and
\( n_i \) of these points are correctly classified.

User’s accuracy: If \( n_i' \) points are found to be classified
into class \( i \),
Knowledge discovery systems have been developed using genetic programming concepts. The MASSON system, where intentional information is extracted for a given set of objects, is popular. The problem addressed is to find common characteristics of a set of objects in an object-oriented database. Genetic programming is used to automatically generate, evaluate, and select object-oriented queries. GAs are also used for several other purposes like fusion of multiple data types in multimedia databases, and automated program generation for mining multimedia data.

E. Rough Sets

The theory of rough sets 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.

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; and 2) the rule induction system based on learning from examples based on rough set theory (LERS).

F. Other Hybridizations

Banerjee et al. have used a rough-neuro-fuzzy integration to design a knowledge-based system, where the theory of rough sets is utilized for extracting domain knowledge. In the said rough-fuzzy MLP, the extracted crude domain knowledge is encoded among the connection weights. Rules are generated from a decision table by computing relative reducts. The network topology is automatically determined and the dependency factors of these rules are encoded as the initial connection weights. The hidden nodes model the conjuncts in the antecedent part of a rule, while the output nodes model the disjuncts.

A promising direction in mining a huge dataset is to 1) partition it; 2) develop classifiers for each module; and 3) combine the results. A modular approach has been pursued to combine the knowledge-based rough-fuzzy MLP sub-networks/modules generated for each class, using GAs. An \(l\)-class classification problem is split into \(l\) two-class problems.
III. CONCLUSION

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.

Recently, several commercial data mining tools have been developed based on soft computing methodologies. These include Data Mining Suite, using fuzzy logic; Braincell, Cognos Thought and IBM Intelligent Miners for Data, using neural networks; and Nuggets.

REFERENCES


