Abstract

The amount of information in the world is increasing exponentially due to the onset of internet and high powered computing. Data mining is a field which is dealing with the challenges of information overload. Rough set theory may be used as a tool for data mining. Rough Set methodology is concerned with the classification and analysis of imprecise, uncertain or incomplete information. This research has generated a path breaking approach to decision making using rough set methodology and aims to discover meaningful and important knowledge from databases using a rule based knowledge discovery system. The knowledge extracted from the databases using rough set theory is usually represented in the form of rules. Since the numbers of rules generated by rough set theory are excessive for knowledge processing therefore, a ‘reduct based rule generation measure’ is proposed to facilitate the knowledge understanding process. An important outcome of this research is that the numbers of rules generated by the proposed method are significantly less in number but are sufficient to understand a particular domain.

The performance of data analysis may be influenced by many factors, and one important factor is the size of the data set. Dimension reduction is needed when the dataset has a large number of features. Most data sets that are collected from real world applications contain noisy data, which may distract the analyst. Thus the original data need to be cleaned for reducing the size of the dataset and removing the noise as well. This data cleaning is usually done by data reduction. Reduct is a proper term in rough set methodology that is used for data reduction. It means a minimal attribute subset with the same approximating power as the whole set. The attributes in a reduct should not only be strongly relevant to the learning task, but also not redundant with each other. The process of searching reducts, called attribute reduction, and is a feature subset selection process. Core contains the indispensable attributes of the dataset. Reduct and Core are the two most important concepts of Rough set theory. Therefore an algorithm for simultaneously finding the reduct and core of the information system has also been proposed which proves to be efficient while retaining the significant knowledge in features of the dataset.
By using the proposed algorithm the redundant attributes are eliminated in order to generate the effective reduct set or to construct the core of the attribute set.

In real life applications, some input data is usually missing in the information table or in other words we can say that decision tables are incompletely specified and some attribute values are frequently absent. Handling missing attribute values in rough set theory is a big challenge. There are many different approaches for handling missing attribute values. Rules are generated from incomplete decision table with the help of the Method of ‘Assigning All Possible Values of the Attribute’ that is the missing attribute values are replaced by all possible values of attribute from the domain of the attribute.

Feature Selection (FS) refers to the problem of selecting those input attributes that are most predictive of a given outcome which can be used to solve problems in many areas such as machine learning, pattern recognition and signal processing. FS preserves the original meaning of the features after reduction, unlike other dimensionality reduction methods. Feature selection techniques aim at reducing the number of insignificant features in the classification rules. Significant and relevant features are found by using three methodologies namely Rough Set, Artificial Neural Networks and Logistic Regression. Features selected by these three techniques have been found to be the same for the examined dataset.

The present research proposes rough sets as an efficient and simplistic decision making tool. Empirical studies on artificial data sets and on real world data sets have been conducted to demonstrate the use of proposed technique such as the medical diagnostic system. Rough set approach has demonstrated its supremacy as better decision making approach in comparison to Artificial Neural Networks and Logistic Regression. It has been found that the proposed technique provide better results.