Rough Sets as a Framework for Data mining

Prof. A.H. Butalia, Prof. M.L. Dhore, Computer Department, VIT, Pune, Maharashtra, India

Abstract The issues of Real World are:
- a) Very large data sets
- b) Mixed types of data (continuous valued, symbolic data)
- c) Uncertainty (noisy data)
- d) Incompleteness (missing, incomplete data)
- e) Data change
- f) Use of background knowledge

The main goal of the rough set analysis is induction of approximations of concepts.
Rough sets constitute a sound basis for KDD. It offers mathematical tools to discover patterns hidden in data.

It can be used for feature selection, feature extraction, data reduction, decision rule generation, and pattern extraction (templates, association rules) etc.
Recent extensions of rough set theory have developed new methods for decomposition of large data sets, data mining in distributed and multi-agent systems, and granular computing.

1. INTRODUCTION

Usually the primary considerations of traditional computing are precision, certainty, and rigor. We distinguish this as “hard” computing. In contrast, the principal notion in soft computing is that precision and certainty carry a cost and that computation, reasoning, and partial truth for obtaining low-cost solutions. This leads to the remarkable human ability of understanding distorted speech, deciphering sloppy handwriting, comprehending the nuances of natural language, summarizing text, recognizing and classifying images, driving a vehicle in dense traffic, and, more generally, making rational decisions in an environment of uncertainty and imprecision. The challenge, then, is to exploit the tolerance for imprecision by devising methods of computation that lead to an acceptable solution at low cost.

Example: if the car is at a distance of d ft and moving at a speed of s ft/s, then press the brake with p poundal for t seconds right now.

Precise solutions are not always feasible. In fact, we do not need a precise solution to such a problem. The exact position where the car stops is not important, but it should stop before the “red light” and should not hit any other care standing ahead of it. Hence an approximate idea about the distance of the car from the care or traffic signal ahead and the speed of the car should be enough. Under this situation X can control the car using rules of the form, “If the car is moving very fast and the ‘red light’ is close, then press the brake pretty hard.”

We can easily say that the action is purely guided by the intuition of an individual, the resultant decision being taken in imprecise terms.

Machine learning overlaps heavily with statistics. In fact, many machine learning algorithms have been found to have direct counterparts with statistics. As a broad subfield of artificial intelligence, Machine learning is concerned with the development of algorithms and techniques that allow computers to "learn". At a general level, there are two types of learning: inductive and deductive. Inductive machine learning methods create computer programs by extracting rules and patterns out of massive data sets. It should be noted that although pattern identification is important to Machine Learning, without rule extraction a process falls more accurately in the field of data mining.

Hence Rough Sets can be used as a framework for datamining specially in the areas of soft computing where exact data is not required and in some areas where approximate data can be of great help.

2. RELEVANCE

The rough sets theory can be used in various stages of information processing: 2.1 Organize a decision table representing an information system containing uncertain or imprecise data.
2.2 Analyze knowledge.
2.3 Analyze of consistency and conflicts in data sets.
2.4 Compute lower and upper approximation of sets.
2.5 Identify and evaluate a dependence of a set of attributes.
2.6 Calculate a discriminatory power of an attribute.
2.7 Calculate quality and accuracy of approximations.
2.8 Calculate reducts as sets of minimal number of attributes describing concepts.
2.9 Reduce data (with information preservation) by removing superfluous attributes.
2.10 Calculate the core and determine a subset of the most significant attributes.
2.11 Determine minimum attribute set from reducts.
2.12 Reason with presence of uncertainty.
2.13 Derive a decision algorithm as a set of production rules.