Similarity Based Rough Sets on the Iris Data Set

Dávid Nagy[a]

[a]University of Debrecen
nagy.david@inf.unideb.hu

Abstract

Nowadays the amount of data is growing exponentially. However, this data is often incomplete or inconsistent. There can be many reasons if a value is missing. For example, it can be unknown, unassigned or even applicable. Inconsistency occurs when the data are contradictory. These issues can cause some undesirable events (bad prediction, inappropriate decision making etc). In computer science, there are numerous ways to handle these kinds of inaccuracies. Rough set theory can be considered as a rather new field in computer science. Its fundamentals was proposed by professor Pawlak in the 80's The pawlakian systems handle the uncertainty among the data with a relation which is based on the indiscernibility of objects. In many cases, based on the available knowledge, two objects cannot be distinguished from each other. Two arbitrary objects can be treated as indiscernible if all of their known properties are the same. This indiscernibility can be modeled by an equivalence relation which represent our background knowledge or its limits. It can affect the membership relation by making the judgment on this relation uncertain. It makes a set vague because a decision about a certain object has an effect on the decisions about all the objects that are indiscernible from the given object. This uncertainty can be represented by set-approximation tools. If one wants to extract as much useful information as possible from large-scale information systems, then it is inevitable to handle the indiscernibility. Rough set theory tries to answer how certain sets can be characterized or if a given object belongs to a set generated by some property. The original Pawlakian approximation space is based on the indiscernibility of objects. This can sometimes be too strict as the objects must have the same properties. In many cases, only the similarity of objects is sufficient. Similarity can be represented by tolerance relations in formal systems from mathematical point of view. The author of this article proposed a new approximation space which is based on the similarity of objects and it is different from the existing approximation spaces that are also based on similarity. This new system is called similarity based rough sets. It is generated by the correlation clustering which is a data mining technique that uses a tolerance relation. In this article, the author investigates how the similarity based rough sets technique can be applied to the famous iris data set and how the different set of flowers can be approximated.

Keywords: Rough set theory, Iris data set, Correlation clustering