Waldemar W. Koczkodaj, Marian Orlowski, and Victor W. Marek

Myths about Rough Set Theory

The main concept of rough set theory (RST) is an indiscernibility relation normally associated with a set of attributes. For example, it may be the set consisting of attributes BloodPressure, Headache, and Temperature for an information table containing patient examination records. The key problem in this description is the informal term "normally associated." In real life, such an association does not exist until additional assumptions are made. In a similar situation (an information table resembles a relation) Codd assumes the existence of functional dependencies stressing that they cannot be found by just examining the data, and hence must be specified by the database designer.

Even simple values such as Yes or No cannot be considered objective for the attribute Headache since it may not be easy to decide if one does or does not have a headache (at least not to a high degree of accuracy). For example, is a minute headache still a headache or it not a headache at all? This problem is even more evident for another attribute, BloodPressure. Values Low, Normal, and High are a result of subjective judgements necessary for establishing boundaries for (more or less) objective measurements of blood pressure. Establishing these boundaries (say, <110,130> for normal), is as subjective as assuming a belief function in Dempster-Shafer theory or a membership function in fuzzy set theory. In fact the situation here is even more complex.

From a medical point of view, it is reasonable to assume that if <110,130> is a normal systolic blood pressure, then so is <109.999,130> since 0.001 is far below the precision of a regular sphygmomanometer. Moreover, we should also be prepared to accept 109.998; 109.997; 109.996; and eventually 0 (after 110,000 subtractions of 0.001 from the originally assumed value) unless an arbitrary decision is made about where to stop, since otherwise we may arrive at the rather risky medical conclusion that every deceased person has a normal blood pressure.

To those who may argue that RST does not need discretization but simply works for rough values, we reply that there aren't any polar bears roaming freely around the North Pole with their fur coats marked Small, Medium, and Large, and yet we are inundated with all kinds of measurements such as time, velocity, pressure, temperature, and so forth, that require discretization since RST is reduced to classic (crisp) set theory for continuous values.

The subjectivity issue in RST is more complicated than in other methods for managing uncertainty. As we stated, fuzzy set theory is based on a subjective membership function and Dempster-Shafer evidence theory requires a subjective belief (or plausibility) function. In the case of RST, however, it is easy to be misled by the illusion of objectivity since RST requires an information table often containing seemingly objective discrete (for instance, Yes, No, Low, or High) data. But an expert's subjective judgements are needed to form an information table from input data describing the real world. For example, the real data may contain blood pressure; these measurements need to be discretized (a process also known as roughing the data in RST terminology) into Low, Normal, and High according to certain criteria. These criteria may be per-

ONE MAY WONDER HOW MANY "SOMEHOWS" are needed to change subjectivity into objectivity.

ceived as objective but the truth is they are more often than not dependent on subjective expert opinions. For example, as mentioned previously, it is impossible to give precise boundaries for blood pressure although every physician can classify a measured blood pressure as Low, Normal, or High; if not, the physician is at least expected distinguish a dead person from a living being (otherwise he or she deserves a "rough doctor" nickname).

The mechanics of RST applied to data (after roughing them) is indeed deterministic. However, it is doubtful this makes the entire method objective since the preprocessed data is heavily contaminated with subjective judgements during the initial roughing stage. It is like the case of a geology professor who insists that technicians remove their eyeglasses fearing the metal from their frames may interfere with the spectrographic analysis of an ore sample meticulously cleaned with a rough metal brush. Roughing data involves so much initial subjective processing that there is no scientific basis for believing any kind of miraculous enhancement by further objective (or rather deterministic) processing would change them into a more precise or objective image of reality.

For an information table of modest size with 10 attributes with 20 values each, the number of possible instances is $5*10^{137}$ (the computations involve the Bell number which exhibits a superexponential growth). It is reasonable to request that RST rules obtained by examining an information table should be somehow representative and therefore that an input information table should be of some representative size. It is no surprise that RST tacitly assumes the information table is somehow given. (One may wonder how many "somehows" one needs to change subjectivity into objectivity.)

RST is, potentially, an important tool in the analysis of data with important applications in data mining and knowledge discovery. However, claims of its superiority (objectivity) over other approaches must be substantiated by scientific evidence. C

Waldemar W. Koczkodaj (waldemar@bethel.cs.laurentian.ca) is a full professor in the Department of Mathematics and Computer Science, Laurentian University, Ontario, Canada.
Marian Orlowski (orlowski@fit.qut.edu.au) is a senior lecturer in the School of Information Systems, Queensland, University of Technology, Australia.
Victor W. Marek (marek@cs.uky.edu) is a professor in the Department of Computer Science, University of Kentucky.