Rough Sets and Data Mining

Jerzy W. Grzymala-Busse
University of Kansas, USA

Wojciech Ziarko
University of Regina, Canada

INTRODUCTION

Discovering useful models capturing regularities of natural phenomena or complex systems until recently was almost entirely limited to finding formulae fitting empirical data. This worked relatively well in physics, theoretical mechanics, and other areas of science and engineering. However, in social sciences, market research, medicine, pharmacy, molecular biology, learning and perception, and in many other areas, the complexity of the natural processes and their common lack of analytical smoothness almost totally exclude the use of standard tools of mathematics for the purpose of data-based modeling. A fundamentally different approach is needed in those areas. The availability of fast data processors creates new possibilities in that respect. This need for alternative approaches to modeling from data was recognized some time ago by researchers working in the areas of neural nets, inductive learning, rough sets, and, more recently, data mining. The empirical models in the form of data-based structures of decision tables or rules play similar roles to formulas in classical analytical modeling. Such models can be analyzed, interpreted, and optimized using methods of rough set theory.

BACKGROUND

The theory of rough sets was originated by Pawlak (1982) as a formal mathematical theory, modeling knowledge about a universe of interest in terms of a collection of equivalence relations. Its main application areas are acquisition, analysis, and optimization of computer-processible models from data. The models can represent functional, partially functional, and probabilistic relations existing in data in the extended rough set approaches (Grzymala-Busse, 1998; Katzberg & Ziarko, 1996; Slezak & Ziarko, 2003; Ziarko, 1993). When deriving the models in the context of the rough set theory, there is no need for any additional information about data, such as, for example, probability distribution function in statistical theory, grade of membership in fuzzy set theory, and so forth (Grzymala-Busse, 1988).

The original rough set approach is concerned with investigating properties and limitations of knowledge. The main goal is forming discriminative descriptions of subsets of a universe of interest. The approach is also used to investigate and prove numerous useful algebraic and logical properties of knowledge and of approximately defined sets, called *rough sets*. The knowledge is modeled by an equivalence relation representing the ability to partition the universe into classes of indiscernible objects, referred to as *elementary sets*. The presence of the idea of approximately defined sets is a natural consequence of imperfections of existing knowledge, which may be incomplete, imprecise, or uncertain. Only an approximate description, in general, of a set (target set) can be formed. The approximate description consists of specification of lower and upper set approximations. The approximations are definable sets. The lower approximation is a union of all elementary sets contained in the target set. The upper approximation is a union of all elementary sets overlapping the target set. This ability to create approximations of non-definable, or rough, sets allows for development of approximate classification algorithms for prediction, machine learning, pattern recognition, data mining, and so forth. In these algorithms, the problem of classifying an observation into an undefinable category, which is not tractable, in the sense that the discriminating description of the category does not exist, is substituted by the problem of classifying the observation into an approximation of the category.

MAIN THRUST

The article is focused on data-mining-related extensions of the original rough set model. Based on the representative extensions, data mining techniques and applications are reviewed.

Extensions of Rough Set Theory

Developing practical applications of rough set theory revealed the limitations of this approach. For example, when dealing with market survey data, it was not possible to identify non-empty lower approximation of the target category of buyers of a product. Similarly, it often was not possible to identify non-trivial upper approximation of the target category, such as would not extend over the whole universe. These limitations follow from the fact that practical classification problems are often non-deterministic. When dealing with such problems, perfect prediction accuracy is not possible and not expected. The need to make rough set theory applicable to a more comprehensive class of practical problems inspired the development of extensions of the original approach to rough sets.

One such extension is the variable precision rough set model (VPRSM) (Ziarko, 1993). As in the original rough set theory, set approximations also are formed in VPRSM. The VPRSM criteria for forming the lower and upper approximations are relaxed, in particular by allowing a controlled degree of misclassification in the lower approximation of a target set. The resulting lower approximation represents an area of the universe where the correct classification can be made with desired probability of success, rather than deterministically. In this way, the VPRSM approach can handle a comprehensive class of problems requiring developing non-deterministic models from data. The VPRSM preserves all basic properties and algorithms of the Pawlak approach to rough sets. The algorithms are enhanced additionally with probabilistic information acquired from data (Katzberg & Ziarko, 1996; Ziarko, 1998, 2003, Ziarko & Xiao, 2004). The structures of decision tables and rules derived from data within the framework of VPRSM have probabilistic confidence factors to reflect the degree of uncertainty in classificatory decision making. The objective of such classifiers is to improve the probability of success rather than trying to guarantee 100% correct classification.

Another extension of rough set theory is implemented in the data mining system LERS (Grzymala-Busse, 1992, 1994), in which rules are equipped with three coefficients characterizing rule quality: specificity (i.e., the total number of attribute-value pairs on the left-hand side of the rule); strength (i.e., the total number of cases correctly classified by the rule during training; and the total number of training cases matching the left-hand side of the rule. For classification of unseen cases, the LERS incorporates the ideas of genetic learning, extended to use partial matching of rules and cases. The decision to which a case belongs is made on the basis of support, defined as the sum of scores of all matching rules from the class, where a score of the rule is the product of the first two coefficients associated with the rule. As indicated by experiments, partial matching is a valuable mechanism when complete matching fails (Grzymala-Busse, 1994). In the LERS classification system, the user may use 16 strategies for classification. In some of these strategies, the final decision is based on probabilities acquired from raw data (Grzymala-Busse & Zou, 1998).

Other extensions of rough set theory include generalizations of the basic concept of rough set theory—the indiscernibility relation. A survey of such methods was presented in Yao (2003).

From Data to Rough Decision Tables

When deriving models from data within the rough set framework, one of the primary constructs is a decision table derived from data referred to as *rough decision table* (Pawlak, 1991; Ziarko, 1999, 2002a). The rough decision table represents knowledge about the universe of interest and the relation between the knowledge and the target set or sets. The idea of the rough decision table was formulated in both the original framework of rough sets and in the extended VPRSM. In the latter case, the table is called *probabilistic decision table* (Ziarko, 2002a). In the table, some columns correspond to descriptive attributes used to classify objects of the domain of interest, while other columns represent target sets or rough approximations of the sets. The rows of the table represent the classes of the classification of the domain in terms of the descriptive attributes. If the decision table contains representatives of all or almost all classes of the domain, and if the relation with the prediction targets is completely or almost completely specified, then the table can be treated as a model of the domain. Such a model represents descriptions of all or almost all objects of the domain and their relationship to the prediction target. The specification of the relationship may include empirical assessments of conditional probabilities, if the VPRSM approach is used in model derivation. If the model is complete enough, and if the data-based estimates of probabilities are relatively close to real values, then the decision table can be used as a basis of a classifier system. To ensure relative completeness and generality of the decision table model, the values of the attributes used to construct the classification of the domain need to be sufficiently general. For example, in many practical problems, rather than using precise numeric measurements, value ranges often are used after preliminary discretization of original precise values. This conversion of original data values into secondary, less precise representation is one of the major pre-processing steps in rough set-based methodology. The acquired decision table can be further analyzed and optimized using classical algorithms for interattribute dependency computation and minimal nonredundant subset of attributes (attribute reduct) identification (Pawlak, 1991; Ziarko 2002b).