IDEA GROUP PUBLISHING
701 E. Chocolate Avenue, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com
ITB8925

Chapter VI

Data Mining Based on Rough Sets

Jerzy W. Grzymala-Busse
University of Kansas, USA

Wojciech Ziarko
University of Regina, Canada

ABSTRACT

The chapter is focused on the data mining aspect of the applications of rough set theory. Consequently, the theoretical part is minimized to emphasize the practical application side of the rough set approach in the context of data analysis and model-building applications. Initially, the original rough set approach is presented and illustrated with detailed examples showing how data can be analyzed with this approach. The next section illustrates the Variable Precision Rough Set Model (VPRSM) to expose similarities and differences between these two approaches. Then, the data mining system LERS, based on a different generalization of the original rough set theory than VPRSM, is presented. Brief descriptions of algorithms are also cited. Finally, some applications of the LERS data mining system are listed.

INTRODUCTION

Discovering useful models capturing regularities of natural phenomena or complex systems was, until recently, almost entirely limited to finding formulas fitting empirical data. This worked relatively well in physics, theoretical mechanics, and other classical and fundamental areas of Science and Engineering. However, in social sciences, market research, medical area, pharmacy, molecular biology, learning and perception in biology, and in many other areas, the complexities of the natural processes and their common lack of analytical "smoothness" almost totally exclude the possibility of using standard mathematical tools for the purpose of data-based modeling. To serve the modeling needs of all these areas, a fundamentally different approach is needed. The availability of fast data processors creates new possibilities in that respect. To take advantage of the possibilities, new mathematical theories oriented towards creating and manipulating empirical functions and relations need to be developed. In fact, 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, in the area of data mining. The models, in the form of data-based structures of decision tables or rules, play a similar role to formulas in classical analytical modeling. Such theories can be analyzed, interpreted, and optimized using the methods of rough set theory.

In this chapter, we are assuming that the reader is familiar with basic concepts of set theory and probability theory.

General Overview of Rough Set Theory

The theory of rough sets (RST) was originated by Pawlak in 1982 as a formal mathematical theory, modeling knowledge about the domain of interest in terms of a collection of equivalence relations (Pawlak, 1982). Its main application area is in acquisition, analysis, and optimization of computer-processible models from data. The models can represent functional, partial functional, and probabilistic relations existing in data in the extended rough set approaches (Katzberg & Ziarko, 1996; Ziarko, 1993, 1999). The main advantage of rough set theory is that it does not need any preliminary or additional information about data (like probability in probability theory, grade of membership in fuzzy set theory, etc.) (Grzymala-Busse, 1988).

The Original Rough Set Model

The original rough set model is concerned with investigating the properties and the limitations of knowledge with respect to being able to form discriminative descriptions of subsets of the domain. The model is also used to investigate and prove numerous useful algebraic and logical properties of the knowledge and approximately defined sets, called rough sets. The inclusion of the approximately defined sets in the rough set model is a consequence of the knowledge imperfections in practical situations. In general, only an approximate description of a set can be formed. The approximate description consists of definitions of lower approximation and upper approximation. The approximations are definable sets, that is, having a discriminative description. The upper approximation is the smallest definable set containing the target set. The lower approximation is the largest definable set included in 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, etc. In these algorithms, the problem of classifying an observation into an indefinable category, which is not tractable in the sense that the discriminating description of the category does not exist, is replaced by the problem of classifying the observation into a definable approximation of the category that is tractable. If the approximations are "tight" enough, then the likelihood of an error of decisionmaking or prediction based on such an approximate classifier is minimal.