

EFFICIENT ALGORITHMS FOR ATTRIBUTES REDUCTION PROBLEM

SONGBO TAN

Graduate School, Chinese Academy of Sciences
Software Department, Institute of Computing Technology
Chinese Academy of Sciences, P. O. Box 2704, Beijing 100080, P. R. China
tansongbo@software.ict.ac.cn

HONGBO XU

Software Department, Institute of Computing Technology
Chinese Academy of Sciences, P. O. Box 2704, Beijing 100080, P. R. China
hbxu@software.ict.ac.cn

JUN GU

Software Department, Institute of Computing Technology
Chinese Academy of Sciences, P. O. Box 2704, Beijing 100080, P. R. China
Department of Computer Science
Hong Kong University of Science and Technology, Hong Kong
eecs@263.net

Received July 2004; revised February 2005

ABSTRACT. *The theory of rough set, proposed by Pawlak, provides a formal tool for knowledge discovery from imprecise and incomplete data. Attributes reduction is a crucial problem for rough set based data mining. Unfortunately, finding minimal reduct turns out to be a NP-hard problem. In this paper, we propose a heuristic reduction algorithm HeuriRed and a complete heuristic reduction algorithm HeuriComRed based on discernibility matrix. The time and space requirements of two proposed algorithms are both $O(|A| * |U|^2)$. In our experiment, in order to investigate the robustness, efficiency and completeness of proposed algorithms, we execute our algorithms HeuriRed, HeuriComRed and genetic reduct on some problems from UCI repository. The experimental results not only indicate the robustness and efficiency of our algorithms, but also provide proofs for the completeness of our algorithm HeuriComRed.*

Keywords: Rough Set, Attributes Reduction, Discernibility Matrix, Data Mining

1. **Introduction.** The rough set theory was developed by Pawlak in the early 1980's [1,2] and has been applied successfully in a lot of domains, such as machine learning, knowledge discovery, and expert systems. It provides powerful tools for data analysis and data mining from imprecise and ambiguous data.

Reduct is the most important concept in rough set-based data mining. Given a distinguished feature, called decision, the notion of decision reduct is constructed over, so called, discernibility matrix, where information about all pairs of objects with different decision values is stored. A reduct is any minimal (in the sense of inclusion) subset of non-decision features (conditions), which discern all such pairs, necessary to be considered, e.g., with