Available online at www.sciencedirect.com
ScienceDirect
Information Sciences 177 (2007) 41–73
INFORMATION SCIENCES
AN INTERNATIONAL JOURNAL
www.elsevier.com/locate/ins

Rough sets and Boolean reasoning

Zdzisław Pawlak, Andrzej Skowron *

Institute of Mathematics, Warsaw University, ul. Banacha 2, 02-097 Warsaw, Poland

Received 24 February 2006; received in revised form 7 June 2006; accepted 7 June 2006

Commemorating the life and work of Zdzisław Pawlak.

Abstract

In this article, we discuss methods based on the combination of rough sets and Boolean reasoning with applications in pattern recognition, machine learning, data mining and conflict analysis.

Keywords: Boolean reasoning; Approximate Boolean reasoning; (In)discernibility; Rough sets; Reducts; Decision rules; Classifiers; Discretization; Symbolic value grouping; Association rules; Conflict analysis

Data! data! data!
Sir Artur Conan Doyle (1994). The Adventures of Sherlock Holmes
Penguin Books, London

1. Introduction: discernibility and Boolean reasoning

The reader is referred to [92] for the notation used in this survey article.

Tasks collected under the labels of data mining, knowledge discovery, decision support, pattern classification, and approximate reasoning require tools aimed at discovering *templates* (*patterns*) in data and classifying them into certain *decision classes*. Templates are in many cases most frequent sequences of events, most probable events, regular configurations of objects, the decision rules of high quality, standard reasoning schemes. Tools for discovery and classification of templates are based on *reasoning schemes* rooted in various paradigms [25]. Such patterns can be extracted from data by means of methods based, e.g., on Boolean reasoning and discernibility (see Section 2 and [15]).

Discernibility relations belong to the most important relations considered in rough set theory. The ability to discern between perceived objects is important for constructing many entities like reducts, decision rules or decision algorithms. In the classical rough set approach, a discernibility relation $DIS(B) \subseteq U \times U$, where

* Corresponding author.
E-mail address: skowron@mimuw.edu.pl (A. Skowron).

0020-0255/$ - see front matter
doi:10.1016/j.ins.2006.06.007

$B \subseteq A$ is a subset of attributes of an information system (U, A), is defined by $xDIS(B)y$ if and only if *non*-$(xI(B)y)$, where $I(B)$ is the B-indiscernibility relation [92]. However, this is, in general, not the case for the generalized approximation spaces. One can define indiscernibility by $x \in I(y)$ and discernibility by $I(x) \cap I(y) = \emptyset$ for any objects x, y, where $I(x) = B(x), I(y) = B(y)$ in the case of the indiscernibility relation, and $I(x), I(y)$ are neighborhoods of objects not necessarily defined by the equivalence relation in a more general case [91].

The idea of Boolean reasoning is based on construction for a given problem P of a corresponding Boolean function f_P with the following property: The solutions for the problem P can be decoded from prime implicants of the Boolean function f_P. Let us mention that to solve real-life problems it is necessary to deal with Boolean functions having large number of variables.

A successful methodology based on discernibility of objects and Boolean reasoning has been developed for computing of many entities important for applications, like reducts and their approximations, decision rules, association rules, discretization of real value attributes, symbolic value grouping, searching for new features defined by oblique hyperplanes or higher order surfaces, pattern extraction from data as well as conflict resolution or negotiation (see Section 2).

Most of the problems related to generation of the mentioned above entities are NP-complete or NP-hard. However, it was possible to develop efficient heuristics returning suboptimal solutions of the problems. The results of experiments on many data sets are very promising. They show very good quality of solutions generated by the heuristics in comparison with other methods reported in the literature (e.g., with respect to the classification quality of unseen objects). Moreover, these heuristics are very efficient from the point of view of time necessary for computing of solutions. Many of these methods are based on discernibility matrices (see Section 2). Note that it is possible to compute the necessary information about these matrices using[1] information encoded in decision systems (e.g., sorted in preprocessing [5,78,134]) directly, which significantly improves the efficiency of algorithms.

It is important to note that the methodology makes it possible to construct heuristics having a very important *approximation property* which can be formulated as follows: Expressions generated by heuristics, i.e., implicants *close* to prime implicants define approximate solutions for the problem.

In Section 2, we discuss applications of methods based on rough sets and Boolean reasoning in machine learning, pattern recognition, and data mining. Section 3 is dedicated to conflict analysis based on different aspects of discernibility and indiscernibility.

2. Rough set methods for machine learning, pattern recognition, and data mining

In supervised machine learning paradigm [28,47,62,63], a learning algorithm is given a training data set, usually in the form of a decision system $\mathscr{A} = (U, A, d)$,[2] prepared by an expert. Every such decision system classifies elements from U into decision classes. The purpose of the algorithm is to return a set of decision rules together with matching procedure and conflict resolution strategy, called a classifier, which makes it possible to classify unseen objects, i.e., objects that are not described in the original decision table. In this section, we provide a number of rough set methods that can be used in construction of classifiers. For more information the reader is referred, e.g., to [2,10,17,18,22,24,26,29–46,53–56,58–60,64,65,67,76,3,83–88,93,97–104,106,107,111,112,122,124–128,130,132,133,135], and for papers on hierarchical learning and ontology approximation, e.g., to [7,11–13,77,80,79,108,113,115,116].

Most of the techniques discussed below are based on computing prime implicants for computing different kinds of reducts. Unfortunately, they are computationally hard. However, many heuristics have been developed which turned out to be very promising. The results of experiments on many data sets, reported in the literature, show a very good quality of classification of unseen objects using these heuristics. A variety of methods for computing reducts and their applications can be found in [5,51,57,88,99,100,106,107,112,114,117,118,135,136]. The fact that the problem of finding a minimal reduct of a given information system is NP-hard was proved in [114].

[1] That is, without the necessity of generation and storing of the discernibility matrices.
[2] For simplicity, we consider decision systems with one decision.