Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

A semantically sound approach to Pawlak rough sets and covering-based rough sets

Lynn D'eer [a,*], Chris Cornelis [a,b], Yiyu Yao [c]

[a] Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Belgium
[b] Department of Computer Science and Artificial Intelligence, Research Center on Information and Communications Technology (CITIC-UGR), University of Granada, Spain
[c] Department of Computer Science, University of Regina, S4S 0A2 Canada

ARTICLE INFO

Article history:
Received 2 February 2016
Received in revised form 13 May 2016
Accepted 27 June 2016
Available online xxxx

Keywords:
Pawlak
Covering
Rough sets
Semantics
Concepts

ABSTRACT

In this paper, we discuss a semantically sound approach to covering-based rough sets. We recall and elaborate on a conceptual approach to Pawlak's rough set model, in which we consider a two-part descriptive language. The first part of the language is used to describe conjunctive concepts, while in the second part disjunctions are allowed as well. Given the language, we discuss its elementary and definable sets, and we study how the approximation operators can be seen as derived notions of the family of definable sets, which is represented by a Boolean algebra over a partition. Furthermore, we generalize the two parts of the language in order to describe concepts of covering-based rough sets. Unfortunately, the family of definable sets will no longer be represented by a Boolean algebra over a partition, but by the union-closure of a covering. Therefore, only the derived covering-based lower approximations of sets are definable for the generalized language. In addition, it is discussed how the two-part languages are used to construct decision rules, which are used in data mining and machine learning.

1. Introduction

In a recent paper, Yao [36] argued that there are two sides to rough set theory: a conceptual and a computational one. In a conceptual approach it is studied how to define various notions and concepts of the theory, while in a computational approach it is studied how to compute them. Therefore, the former approach provides insights to the concepts of the theory, but may not supply computationally efficient algorithms, whereas the latter approach is very suitable for computations and applications, but the meaning of the concepts may be lost. Hence, both approaches are fundamental in the research on rough set theory.

A fundamental task of rough set theory is to analyze data representation in order to derive decision rules [8]. The left-hand-side (LHS) and the right-hand-side (RHS) of a rule are descriptions of two concepts and the rule is a linkage between the two concepts. In general, the left-hand-side consists of a conjunction of atomic formulas (atoms), where an atomic formula describes the smallest information block for a given attribute of the table and a possible value of that attribute. The right-hand-side of a rule consists of a disjunction of such atomic formulas. If an object satisfies all the atomic

[*] Corresponding author.
E-mail addresses: Lynn.Deer@UGent.be (L. D'eer), Chris.Cornelis@decsai.ugr.es (C. Cornelis), yyao@cs.uregina.ca (Y. Yao).

http://dx.doi.org/10.1016/j.ijar.2016.06.013
0888-613X/© 2016 Published by Elsevier Inc.

formulas in the left-hand-side of a decision rule, it will satisfy one of the atomic formulas in the right-hand-side of the rule. Hence, we can make a decision on this object.

It is important to have a formal way to represent and interpret those descriptors. To describe the semantics of a concept, we discuss its intension and its extension [36]. While the intension of a concept describes the properties that are characteristics of the concept, the extension of a concept contains all the objects satisfying the properties of the intension. Unfortunately, the intensions of concepts are barely discussed in the computational models, and if they are discussed, the intension and extension of a concept are not explicitly connected, as happens in Bonikowski et al. [1].

Such a semantically sound approach using both the intension and extension was already suggested by Pawlak [15] and Marek and Pawlak [12] prior to the introduction of rough set theory. However, except for a few articles by Marek and Truszczyński [13] and Yao [32,33,37], the conceptual formulation of rough sets is scarcely discussed. For three decades, the focus of the rough set research field has been on the computational approach of rough set approximation operators. Since Pawlak's seminal paper on rough sets in the early 1980s [16], the development of computational rough set models has flourished. Inter alia, a binary relation or neighborhood operator is often used to describe the indiscernibility relation between instances of the universe [22,25,29,30,41]. In addition, several covering-based rough set models are defined in literature [2,9,18,21,24,26–28,30,38–41]. More recently, the classification and comparison of these different rough set models have been discussed [3,19,20,35]. However, whereas in the model of Pawlak there is a clear semantical connection between the given data in the information or decision table, this connection is often absent in generalized models.

With the aim of refocusing our attention again on the earlier research, we recall the rough set framework of Pawlak [16] from a semantical point of view. Starting from the data, the definability of subsets of the universe is discussed before the notion of approximation operators [32]. In order to do this, a descriptive language is constructed in two parts. The formulas in the language are considered the intensions of the concepts. Corresponding to a formula, its meaning set, i.e., the set of objects satisfying the formula, is the extension of the concept. A set of objects is therefore definable, if it is the meaning set of a formula in the descriptive language, otherwise, it is undefinable. From this point of view, approximation operators are introduced in order to describe undefinable sets by means of definable sets. Given an undefinable set X, the greatest definable set contained in X is called the lower approximation of X, while the least definable set containing X is called the upper approximation of X. Therefore, these approximation operators are the only meaningful ones in this framework. Moreover, it will be discussed that the definable sets for Pawlak's model can be computed by a Boolean algebra over a partition related to the data.

Unfortunately, it is not always possible to define such a partition, for example, when the data is incomplete. Computational approaches for incomplete information tables were already discussed by Kryszkiewicz [10,11] and Grzymala-Busse [7]. Other examples include ordered information tables, in which the equivalence classes will mostly consist of only one object which is unreasonable for applications such as rule induction. For analyzing such information tables, Greco et al. used dominance-based rough sets [4,5,23]. Therefore, we extend the semantical approach of Pawlak's model to more general covering-based rough set models. However, the definable sets will no longer be computed by the use of a Boolean algebra over a partition, but by the union-closure of a covering.

The outline of this article is as follows. In Section 2, we situate our research and outline the main goal of this paper. We discuss the notion of an elementary and a definable set, and their connection with rule induction. In Section 3, we introduce a semantically sound approach to Pawlak's original rough set model, which is generalized for covering-based rough set models in Section 4. To end, we state concluding remarks and future research directions in Section 5.

2. A new conceptual understanding of rough set models

A possible application of Pawlak's rough set model and covering-based rough set models is rule induction. It is an important technique to extract knowledge from data represented in a decision table [8]. In this article, we assume the table to be complete. A complete *information table* is represented by the following tuple:

$$T = (U, At, \{V_a \mid a \in At\}, \{I_a \mid a \in At\}),$$

where U is a finite non-empty set of objects called the universe, At is a finite non-empty set of attributes, V_a is a non-empty set of values for $a \in At$, and $I_a : U \to V_a$ is a complete information function that maps an object of U to exactly one value in V_a. The table T is called a complete *decision table* if the set of attributes At is the union of two disjoint sets C and $\{d\}$, with C the set of conditional attributes and d the decision attribute.

Knowledge in a decision table may be described with a set of rules, where each rule consists of a condition part (left-hand-side) and decision part (right-hand-side), based on the conditional and decision attributes of the table, respectively. In general, the condition part of a rule can be written as a conjunction of atoms, while the decision part of the rule consists of a disjunction of atoms. For example, a rule can be represented as follows:

$$\begin{gathered}\text{If object } x \text{ satisfies condition}_1 \wedge \text{condition}_2 \wedge \ldots \wedge \text{condition}_n, \\ \text{then } x \text{ satisfies decision}_1 \vee \text{decision}_2 \vee \ldots \vee \text{decision}_m.\end{gathered}$$

Every atom is a formula related to an attribute a and one of its values v. It is therefore the intension related to the smallest indivisible block of information we can obtain from a decision table given the pair (a, v). Every object that satisfies