Available online at www.sciencedirect.com
ScienceDirect
Fuzzy Sets and Systems 274 (2015) 124–142
FUZZY sets and systems
www.elsevier.com/locate/fss

Generation of rough sets reducts and constructs based on inter-class and intra-class information

Robert Susmaga [a], Roman Słowiński [a,b,*]

[a] *Institute of Computing Science, Poznań University of Technology, Piotrowo 2, 60-965 Poznań, Poland*
[b] *Systems Research Institute, Polish Academy of Sciences, Newelska 6, 01-477 Warsaw, Poland*

Received 7 October 2013; received in revised form 10 June 2014; accepted 19 June 2014

Available online 25 June 2014

Abstract

The reduct, originating from the Classic Rough Set Approach (CRSA), is an inclusion minimal subset of attributes that provides discernibility between objects from different classes in at least the same degree as the set of all attributes. It can be thus referred to as being consistent and minimal, which are the two important characteristics of filter-based feature selection. These two characteristics have been also utilized to define reducts within the Dominance-based Rough Set Approach (DRSA). Further, the classic reduct, here referred to as an inter-class reduct, has evolved into what is known as intra-class reduct and construct in CRSA. The idea is that while inter-class reducts utilize only one part of information generated from all pairs of objects, intra-class reducts utilize the remaining part, while constructs utilize both. The paper delivers a final unification of inter-class reducts, intra-class reducts and constructs across CRSA and DRSA, showing how they can be both defined and computed uniformly, i.e. using basically the same concepts and algorithms. It also presents an exact algorithm, capable of generating all exact reduced subsets, but of considerable complexity, as well as a simple and fast heuristic, designed to generate a single reduced subset. Finally, it illustrates the computation process with examples and some experimental evaluation of CRSA constructs, which show how the use of both the inter-class and the intra-class information can assist the attribute reduction process and help obtaining useful insights into the analyzed data set.

Keywords: Rough sets; Reducts; Constructs; Inter-class and intra-class information; Indiscernibility; Similarity; Dominance

1. Introduction

In their paper [40], H. Tanaka, H. Ishibuchi and T. Shigenaga describe an expert system designed to create fuzzy if–then rules from a properly pre-processed decision table. Apart from applying fuzzyfication of the attribute value sets, an essential part of this pre-processing consists in reducing the attributes in the table, that is in selecting subsets of attributes that seem most relevant for the assumed task. Utilizing the concept of a superfluous attribute, the presented procedure allows to arrive at a reduced set of attributes, referred to as a reduct, a concept originating from the works

* Corresponding author at: Institute of Computing Science, Poznań University of Technology, Piotrowo 2, 60-965 Poznań, Poland.
E-mail address: roman.slowinski@cs.put.poznan.pl (R. Słowiński).

http://dx.doi.org/10.1016/j.fss.2014.06.012

of Z. Pawlak [17,19,20]. Summing up, two principal characteristics of the procedure can be distinguished: the need to achieve sufficient degree of object discernibility by the reduced set of attributes, and to keep this set as small (in terms of its cardinality) as possible. These two characteristics, which can be summarized as consistency and minimality, constitute two crucial aspects of the feature selection process based on the filter model and, as such, will be further exploited in the current paper.

The so-called ‘filter’ model, which may be viewed as a part of the Feature Selection family of methods [5,10], applies a collection of pre-defined, well-interpretable conditions to appoint sets of attributes that are presumed to form the best reduced sets. An alternative approach, the so-called ‘wrapper’ model, uses a pre-defined classifier, a pre-defined cross-validation procedure and a pre-defined classification measure, which are repeatedly looped in search for best classification results. Despite being very sensitive to the particular type of classifier, the particular cross-validation procedure and the particular classification measure used, the wrapper model may prove successful in many real-life applications, especially in those that do not need further justification. This is because this model does not specify why these and not other attributes have been selected, in result of which its outcomes are generally hardly interpretable.

The filter model, on the other hand, produces results that are more generalizable and interpretable, which means that they may be applied to a wider spectrum of data sets, potentially in different areas. In particular, the filter model is especially well suited to data analyses with descriptive goals, like data sets of historic character, i.e. data analyzed with the aim of revealing dependencies that are not intended to be generalized to other data sets or used for prediction. Finally, it remains a good alternative in the case of those data sets that are not large enough to provide statistically valid results for the various requirements of the ‘wrapper’ model.

The classic reducts, as well as all the concepts presented in this paper, represent the filter model. It is worth stressing that the classic reducts [20] have rich research history and have been the subject of many studies, in both classic [18,22,25] and extended forms [2,14–16,38,41]. They have also been considered in numerous papers on various aspects of information systems analysis [21,24,26,42], including the Dominance-based Rough Set Approach (DRSA) [7,21,28,29], a rapidly developing [4,12] branch of the theory, having a strong impact on the domain of multicriteria decision aiding [23].

The founding relation in CRSA is indiscernibility between objects [20], as induced by sets of discrete attributes. Indiscernibility induced by all condition attributes partitions the universe of discourse into subsets of objects referred to as atoms, while indiscernibility induced by the decision attribute partitions the same universe into subsets of objects referred to as classes. The main idea in CRSA is to examine how the latter partition may be approximated by the former, and to quantify this approximation with what is called the quality of the approximation, a measure of data consistency. A classic reduct is thus a subset of attributes that is minimal with respect to inclusion and ensures maximal consistency, i.e. the same level of consistency (quality of approximation) as that ensured by the whole set of condition attributes. When multiple reducts exist for a given data set, the intersection of all of them is referred to as the core of the attributes. Of course, only one (although potentially empty) core exists in each case.

A fully equivalent definition of CRSA reduct may be formulated using relations applied to pairs of objects belonging to different classes [25]. Because the consistency level decreases only when objects which belong to different classes and which are maximally discernible (i.e. discernible by the whole set of attributes) become indiscernible, the reduct is defined as such a subset of attributes that is minimal with respect to inclusion and ensures maximal discernibility between all objects that belong to different classes, i.e. discernibility equal to that ensured by the set of all attributes. Similar, relation-based definition exists for the CRSA core. Interestingly enough, the relation-based definitions are direct in the sense that the otherwise required computations, which include all the laborious steps leading to the calculation of the quality of approximation, can be spared in their case.

As far as DRSA is concerned, the founding relation is that of dominance implied by preference on the attribute value sets. It is thus assumed that the descriptions of objects in terms of the decision attribute as well as in terms of the condition attributes are preferentially ordered. In result, the founding relation determines what is called cones of dominance, both in condition space and in decision space (usually one-dimensional). Central to DRSA is the idea to examine how the decisional cone may be approximated with the conditional one, and to quantify this approximation with the quality of the approximation, a procedure similar to that used in CRSA. The reduct and the core have analogous definitions here, together with their relation-based versions.

The relation-based definition of reducts in CRSA and DRSA makes it clear that these concepts are built using only a part of all available information conveyed by inter-object (i.e. object–object) relations. The whole of this information