Rough Set Strategies to Data with Missing Attribute Values

Jerzy W. Grzymala-Busse

Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS 66045, USA
and
Institute of Computer Science, Polish Academy of Sciences, 01-237 Warsaw, Poland
Jerzy@ku.edu
http://lightning.eecs.ku.edu/index.html

Summary. In this paper we assume that a data set is presented in the form of the incompletely specified decision table, i.e., some attribute values are missing. Our next basic assumption is that some of the missing attribute values are lost (e.g., erased) and some are "do not care" conditions (i.e., they were redundant or not necessary to make a decision or to classify a case). Incompletely specified decision tables are described by characteristic relations, which for completely specified decision tables are reduced to the indiscernibility relation. It is shown how to compute characteristic relations using an idea of block of attribute-value pairs, used in some rule induction algorithms, such as LEM2. Moreover, the set of all characteristic relations for a class of congruent incompletely specified decision tables, defined in the paper, is a lattice. Three definitions of lower and upper approximations are introduced. Finally, it is shown that the presented approach to missing attribute values may be used for other kind of missing attribute values than lost values and "do not care" conditions.

Key words: Data mining, rough set theory, incomplete data, missing attribute values

1 Introduction

Usually all ideas of rough set theory are explored using decision tables as a starting point [10], [11]. The decision table describes cases (also called examples or objects) using attribute values and a decision. Attributes are independent variables while the decision is a dependent variable. In the majority of papers on rough set theory it is assumed that the information is complete, i.e., that for all cases all attribute values and decision values are specified. Such a decision table is said to be completely specified.

In practice, however, input data, presented as decision tables, may have missing attribute and decision values, i.e., decision tables are incompletely

specified. Since our main concern is learning from examples, and an example with a missing decision value, (i.e., not classified) is useless, we will assume that only attribute values may be missing.

There are two main reasons why an attribute value is missing: either the value was lost (e.g., was erased) or the value was not important. In the former case attribute value was useful but currently we have no access to it. In the latter case the value does not matter, so such values are also called "do not care" conditions. In practice it means that originally the case was classified (the decision value was assigned) in spite of the fact that the attribute value was not given, since the remaining attribute values were sufficient for such a classification or to make a decision. For example, a test, represented by that attribute, was redundant.

The first rough set approach to missing attribute values, when all missing values were lost, was described in 1997 in [7], where two algorithms for rule induction, LEM1 and LEM2, modified to deal with such missing attribute values, were presented. In 1999 this approach was extensively described in [13], together with a modification of the original idea in the form of a valued tolerance based on a fuzzy set approach.

The second rough set approach to missing attribute values, in which the missing attribute value is interpreted as a "do not care" condition, was used for the first time in 1991 [4]. A method for rule induction was introduced in which each missing attribute value was replaced by all possible values. This idea was further developed and furnished with theoretical properties in 1995 [8].

In this paper a more general rough set approach to missing attribute values is presented. In this approach, in the same decision table, some missing attribute values are assumed to be lost and some are "do not care" conditions. A simple method for computing a characteristic relation describing the decision table with missing attribute values of either of these two types is presented. The characteristic relation for a completely specified decision table is reduced to the ordinary indiscernibility relation. It is shown that the set of all characteristic relations, defined by all possible decision tables with missing attribute values being one of the two types, together with two defined operations on relations, forms a lattice.

Furthermore, three different definitions of lower and upper approximations are introduced. Similar three definitions of approximations were studied in [15], [16], [17]. Some of these definitions are better suited for rule induction. Examples of rules induced from incompletely specified decision tables are provided. The paper ends up with a discussion of other approaches to missing attribute values.

A preliminary version of this paper was presented at the Workshop on Foundations and New Directions in Data Mining, associated with the third IEEE International Conference on Data Mining [6].