Extracting Rules from Incomplete Decision Systems: System ERID

Agnieszka Dardzińska[1] and Zbigniew W. Raś[2,3]

[1] Bialystok Technical University, Department of Mathematics, ul. Wiejska 45A, 15-351, Bialystok, Poland adardzin@uncc.edu
[2] University of North Carolina, Department of Computer Science, Charlotte, N.C. 28223, USA
[3] Polish Academy of Sciences, Institute of Computer Science, ul. Ordona 21, 01-237 Warsaw, Poland ras@uncc.edu

Summary. We present a new method for extracting rules from incomplete Information Systems (IS) which are generalizations of information systems introduced by Pawlak [7]. Namely, we allow to use a set of weighted attribute values instead of a single value to describe objects in IS. The proposed strategy has some similarities with system LERS [3]. It is a bottom-up strategy, guided by two thresholds values (minimum support and minimum confidence) and generating sets of weighted objects with descriptions of minimal length. The algorithm starts with identifying sets of objects having descriptions of length one (values of attributes). Some of these sets satisfy both thresholds values and they are used for constructing rules. They are marked as successful. All sets having a number of supporting objects below the threshold value are marked as unsuccessful. Pairs of descriptions of all remaining sets (unmarked) are used to construct new sets of weighted objects having descriptions of length 2. This process is continued recursively by moving to sets of weighted objects having k-value properties. In [10], [1], ERID is used as a null value imputation toll for knowledge discovery based chase algorithm.

1 Introduction

There is a number of strategies which can be used to extract rules describing values of one attribute (called decision attribute) in terms of other attributes (called classification attributes) available in the system. For instance, we can mention here such systems like $LERS$ [3], $AQ19$ [5], *Rosetta* [6] and, $C4.5$ [8]. In spite of the fact that the number of rule discovery methods is still increasing, most of them are developed under the assumption that the information about objects in information systems is either precisely known or not known at all. This implies that either one value of an attribute is assigned to an object as its property or no value is assigned to it (instead of no value we use the term *null value*). Problem of inducing rules from information systems with attribute

values represented as sets of possible values was discussed for instance by Kryszkiewicz and Rybinski [4], Greco, Matarazzo, and Slowinski [2], and by Ras and Joshi [9].

In this paper, we present a new strategy for discovering rules in information systems when data are partially incomplete. Namely, we allow to use a set of weighted attribute values as a value of an attribute. A weight assigned to an attribute value represents user confidence in that value. For instance, by assigning a value $\{(brown, \frac{1}{3}), (black, \frac{2}{3})\}$ of the attribute *Color of Hair* to an object x we say that the confidence in object x having brown hair is $\frac{1}{3}$ whereas the confidence that x has black hair is $\frac{2}{3}$. Similar assumption was used, for instance, in papers by Greco [2] or Slowinski [11] but their approach to rule extraction from incomplete information systems is different than ours.

2 Discovering Rules in Incomplete IS

In this section, we give the definition of an incomplete information system which can be called *probabilistic* because of the requirement placed on values of attributes assigned to its objects. Namely, we assume that value of an attribute for a given object is a set of weighted attribute values and the sum of these weights has to be equal to one. Next, we present an informal strategy for extracting rules from incomplete information systems. Finally, we propose a method of how to compute confidence and support of such rules.

We begin with a definition of an incomplete information system which is a generalization of an information system given by Pawlak [7].

By an incomplete Information System we mean $S = (X, A, V)$, where X is a finite set of objects, A is a finite set of attributes and $V = \bigcup\{V_a : a \in A\}$ is a finite set of values of attributes from A. The set V_a is a domain of attribute a, for any $a \in A$.

We assume that for each attribute $a \in A$ and $x \in X$, $a(x) = \{(a_i, p_i) : i \in J_{a(x)} \wedge (\forall i \in J_{a(x)})[a_i \in V_a] \wedge p_i = 1\}$.

Null value is interpreted as the set all possible values of an attribute with equal confidence assigned to all of them. Table 1 gives an example of an incomplete information system $S = (\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}, \{a, b, c, d, e\}, V\}$.

Clearly, $a(x_1) = \{(a_1, \frac{1}{3}), (a_2, \frac{2}{3})\}, a(x_2) = \{(a_2, \frac{1}{4}), (a_3, \frac{3}{4})\}, ..$

Let us begin to extract rules from S describing attribute e in terms of attributes $\{a, b, c, d\}$ following a strategy similar to $LERS$ [3]. We start with identifying sets of objects in X having properties $a_1, a_2, a_3, b_1, b_2, c_1, c_2, c_3, d_1, d_2$ and next for each of these sets we check in what relationship it is with a set of objects in X having property e_1, next property e_2, and finally property e_3. Attribute value a_1 is interpreted as a set $a_1^\star$, equal to $\{(x_1, \frac{1}{3}), (x_3, 1), (x_5, \frac{2}{3})\}$. The justification of this interpretation is quite simple. Only 3 objects in S