Proceedings of the IPMU'2004, the 10th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Perugia, Italy, July 4–9, 2004, vol. 2, 923–930

Rough Set Approaches to Rule Induction from Incomplete Data

Jerzy W. Grzymala-Busse
Department of Electrical Engineering
and Computer Science
University of Kansas,
Lawrence, KS 66045, USA
Jerzy@ku.edu
and
Institute of Computer Science
Polish Academy of Sciences, 01-237 Warsaw, Poland

Sachin Siddhaye
Department of Electrical Engineering
and Computer Science
University of Kansas,
Lawrence, KS 66045, USA
SSiddhaye@eecs.ku.edu

Abstract

In this paper we assume that data are presented in the form of decision tables, incomplete when some attribute values are missing. Two main cases of missing attribute values are considered: lost (the original value was erased) and "do not care" conditions (the original value was irrelevant). This paper uses, as the main tool, attribute-value pair blocks. These blocks are used to construct characteristic sets, characteristic relations, and lower and upper approximations for decision tables with missing attribute values. For such tables three different definitions of lower and upper approximations may be applied: singleton, subset, and concept.
A modified version of the LEM2 rule induction algorithm, accepting input data with both lost values and "do not care" conditions, is described. Results of experiments on some real-life incomplete data, in which all missing attribute values were considered to be either lost or "do not care" conditions are presented as well. A conclusion is that an error rate for classification is smaller when missing attribute values are considered to be lost.

Keywords. Rough set theory, incomplete data, missing attribute values, indiscernibility relation, characteristic relations.

1. Introduction

In real-life data some attribute values are frequently missing. There are two main reasons for attribute values to be missing: either they are lost (e.g., were erased) or they are "do not care" conditions (i.e., the original values were not recorded at all since they were irrelevant, and the decision to which concept a case belongs was taken without that information). Other interpretations of missing attribute values than lost and "do not care" conditions were presented in [4]. Decision tables with all missing attribute values that are lost were studied, within rough set theory, in [6], where two algorithms for rule induction from such data were presented. This approach was studied later, see, e.g., [11–13], where the indiscernibility relation was generalized to describe such incompletely specified decision tables. On the other hand, the first attempt to study "do not care" conditions using rough set theory was presented in [2], where a method for rule induction was introduced in which missing attribute values were replaced by all values from the domain of the attribute. "Do not care" conditions were also studied later, see, e.g., [8, 9], where the indiscernibility relation was again generalized, this time to describe incomplete decision tables with "do not care" conditions.

This paper uses, as the main tool, attribute-value pair blocks. These blocks are used to construct characteristic sets, characteristic relations, and lower and upper approximations

Table 1. An example of an incomplete decision table

	Attributes			Decision
	Age	Hypertension	Complications	Delivery
1	?	*	none	fullterm
2	20..29	yes	obesity	preterm
3	20..29	yes	none	preterm
4	20..29	no	none	fullterm
5	30..39	yes	?	fullterm
6	*	yes	alcoholism	preterm
7	40..50	no	?	fullterm

for decision tables with missing attribute values [4, 5]. We are assuming that the same decision table may contain both types of missing attribute values: lost and "do not care" conditions. A characteristic relation is a generalization of the indiscernibility relation. For such tables three different definitions of lower and upper approximations may be applied: singleton, subset, and concept [4, 5]. Similar three definitions of lower and upper approximations, though not for incomplete decision tables, were studied in [14–16].

A rule induction algorithm LEM2 (Learning from Examples Module, version 2), a component of LERS (Learning from Examples based on Rough Sets), [1, 3], is also based on the idea of attribute-value pair blocks, hence it was natural to modify LEM2 to accommodate decision tables with missing attribute values.

Some experiments with real-life incomplete data were conducted using the MLEM2 algorithm, a modified version of LEM2. Our objective was to compare different approaches to missing attribute values. The conclusion is that for these data lower error rate may be accomplished by assuming that missing attribute values are lost.

2. Blocks of attribute-value pairs, characteristic sets, and characteristic relations

An example of an incomplete decision table, taken from [5], is presented in Table 1.

Rows of the decision table represent *cases*, while columns represent *variables*. The set of all cases is denoted by U. In Table 1, $U = \{1, 2, ..., 7\}$. Independent variables are called *attributes* and a dependent variable is called a *decision* and is denoted by d. The set of all attributes will be denoted by A. In Table 1, $A = \{Age, Hypertension, Complications\}$. Any decision table defines a function ρ that maps the direct product of U and A into the set of all values. For example, in Table 1, $\rho(1, Age) =$ 20..29. A decision table with an incompletely specified (partial) function ρ will be called *incompletely specified*, or *incomplete*. For the rest of the paper we will assume that all decision values are specified, i.e., they are not missing. Also, we will assume that all missing attribute values are denoted either by "?" or by "*", lost values will be denoted by "?", "do not care" conditions will be denoted by "*". Additionally, we will assume that for each case at least one attribute value is specified.

One of the fundamental ideas of rough set theory is an indiscernibility relation. For $B \subseteq A$ and $x, y \in U$, the indiscernibility relation

IND(B) is a relation on U defined as follows

$(x, y) \in$ IND(B) if and only if $\rho(x, a) = \rho(y, a)$ for all $a \in B$.

The indiscernibility relation IND(B) is an equivalence relation. Equivalence classes of IND(B) are called *elementary sets* and are denoted by $[x]_B$. Elementary sets may be computed by using attribute-value pair blocks. Let $a \in A$ and let v be a value of a for some case. For complete decision tables if $t = (a, v)$ is an attribute value pair, then a *block* of t, denoted $[t]$, is a set of all cases from U that for attribute a have value v.

Incomplete decision tables are described by characteristic relations instead of indiscernibility relations. Also, elementary sets are replaced by characteristic sets. An example of an incomplete table, again taken from [5], is presented in Table 1.

For incomplete decision tables the definition of a block of an attribute-value pair must be modified. If for an attribute a there exists a case x such that $\rho(x, a) = ?$, i.e., the corresponding value is lost, then the case x should not be included in any block $[(a, v)]$ for all values v of attribute a. If for an attribute a there exists a case x such that the corresponding value is a "do not care" condition, i.e., $\rho(x, a) = *$, then the corresponding case x should be included in all blocks $[(a, v)]$ for every possible value v of attribute a [4, 5]. This modification of the definition of the block of attribute-value pair is consistent with the interpretation of missing attribute values, lost and "do not care" condition. Thus, for Table 1

$$[(\text{Age}, 20..29)] = \{2, 3, 4, 6\},$$
$$[(\text{Age}, 30..39)] = \{5, 6\},$$
$$[(\text{Age}, 40..50)] = \{6, 7\},$$
$$[(\text{Hypertension}, \text{yes})] = \{1, 2, 3, 5, 6\},$$
$$[(\text{Hypertension}, \text{no})] = \{1, 4, 7\},$$
$$[(\text{Complications}, \text{none})] = \{1, 3, 4\},$$
$$[(\text{Complications}, \text{obesity})] = \{2\},$$
$$[(\text{Complications}, \text{alcoholism})] = \{6\}.$$

The characteristic set $K_B(x)$ is the intersection of blocks of attribute-value pairs (a, v) for all attributes a from B for which $\rho(x, a)$ is specified and $\rho(x, a) = v$ [4, 5]. For Table 1 and $B = A$,

$$K_A(1) = \{1, 3, 4\},$$
$$K_A(2) = \{2, 3, 4, 6\} \cap \{1, 2, 3, 5, 6\} \cap \{2\} = \{2\},$$
$$K_A(3) = \{2, 3, 4, 6\} \cap \{1, 2, 3, 5, 6\} \cap \{1, 3, 4\} = \{3\},$$
$$K_A(4) = \{2, 3, 4, 6\} \cap \{1, 4, 7\} \cap \{1, 3, 4\} = \{4\},$$
$$K_A(5) = \{5, 6\} \cap \{1, 2, 3, 5, 6\} = \{5, 6\},$$
$$K_A(6) = \{1, 2, 3, 5, 6\} \cap \{6\} = \{6\},$$

and

$$K_A(7) = \{6, 7\} \cap \{1, 4, 7\} = \{7\}.$$

Characteristic set $K_B(x)$ may be interpreted as the smallest set of cases that are indistinguishable from x using all attributes from B, using a given interpretation of missing attribute values. Thus, $K_A(x)$ is the set of all cases that cannot be distinguished from x using all attributes.

The characteristic relation $R(B)$ is a relation on U defined for $x, y \in U$ as follows:

$(x, y) \in R(B)$ if and only if $y \in K_B(x)$.

We say that $R(B)$ is *implied* by its characteristic sets $K_B(x)$, $x \in U$. The characteristic relation $R(B)$ is reflexive but—in general—does not need to be symmetric or transitive.

3. Lower and upper approximations

For completely specified decision tables lower and upper approximations are defined on the basis of the indiscernibility relation. Any finite union of elementary sets, associated with B, will be called a *B-definable* set. Let X be any subset of the set U of all cases. The set X is called a *concept* and is usually defined as the set of all cases defined by a specific value of the decision. In general, X is not a B-definable set. However, set X may be approxi-

mated by two B-definable sets, the first one is called a *B-lower approximation of X*, denoted by $\underline{B}X$ and defined as follows

$$\{x \in U \mid [x]_B \subseteq X\}.$$

The second set is called a *B-upper approximation of X*, denoted by $\overline{B}X$ and defined as follows

$$\{x \in U \mid [x]_B \cap X \neq \emptyset\}.$$

The above shown way of computing lower and upper approximations, by constructing these approximations from singletons x, will be called the first method. The B-lower approximation of X is the greatest B-definable set, contained in X. The B-upper approximation of X is the smallest B-definable set containing X.

As it was observed in [10], for complete decision tables we may use a second method to define the B-lower approximation of X, by the following formula

$$\cup \{[x]_B \mid x \in U, [x]_B \subseteq X\},$$

and the B-upper approximation of x may de defined, using the second method, by

$$\cup\{[x]_B \mid x \in U, [x]_B \cap X \neq \emptyset).$$

For incompletely specified decision tables lower and upper approximations may be defined in a few different ways. First, the definition of definability should be modified. Any finite union of characteristic sets of B is called a *B-definable* set [5]. Three different definitions of lower and upper approximations may be used [4, 5]. Again, let X be a concept, let B be a subset of the set A of all attributes, and let $R(B)$ be the characteristic relation of the incomplete decision table with characteristic sets $K(x)$, where $x \in U$. Our first definition uses a similar idea as in the previous articles on incompletely specified decision tables [8, 9, 11–13], i.e., lower and upper approximations are sets of singletons from the universe U satisfying some properties. Thus, lower and upper approximations are defined by analogy with the above first method, by constructing both sets from singletons. We will call these definitions *singleton*. A singleton B-lower approximation of X is defined as follows:

$$\underline{B}X = \{x \in U \mid K_B(x) \subseteq X\}.$$

A singleton B-upper approximation of X is

$$\overline{B}X = \{x \in U \mid K_B(x) \cap X \neq \emptyset\}.$$

In our example of the decision presented in Table 1 let us say that $B = A$. Then the singleton A-lower and A-upper approximations of the two concepts: {1, 4, 5, 7} and {2, 3, 6} are:

$$\underline{A}\{1, 4, 5, 7\} = \{4, 7\},$$

$$\underline{A}\{2, 3, 6\} = \{2, 3, 6\},$$

$$\overline{A}\{1, 4, 5, 7\} = \{1, 4, 5, 7\},$$

$$\overline{A}\{2, 3, 6\} = \{1, 2, 3, 5, 6\}.$$

Note that the set {1, 4, 5, 7} is not A-definable (this set cannot be presented as a union of intersections of attribute-value pair blocks). Therefore singleton approximations are not useful. The second method of defining lower and upper approximations for complete decision tables uses another idea: lower and upper approximations are unions of elementary sets, subsets of U. Therefore we may define lower and upper approximations for incomplete decision tables by analogy with the second method, using characteristic sets instead of elementary sets. There are two ways to do this. Using the first way, a *subset B*-lower approximation of X is defined as follows:

$$\underline{B}X = \cup\{K_B(x) \mid x \in U, K_B(x) \subseteq X\}.$$

A *subset B*-upper approximation of X is

$$\overline{B}X = \cup\{K_B(x) \mid x \in U, K_B(x) \cap X \neq \emptyset\}.$$

Since any characteristic relation $R(B)$ is reflexive, for any concept X, singleton B-lower and B-upper approximations of X are subsets of the subset B-lower and B-upper approxima-

tions of X, respectively. For the same decision table, presented in Table 1, the subset A-lower and A-upper approximations are

$$\underline{A}\{1, 4, 5, 7\} = \{4, 7\},$$

$$\underline{A}\{2, 3, 6\} = \{2, 3, 6\},$$

$$\overline{A}\{1, 4, 5, 7\} = \{1, 3, 4, 5, 6, 7\},$$

$$\overline{A}\{2, 3, 6\} = \{1, 2, 3, 4, 5, 6\}.$$

The second possibility is to modify the subset definition of lower and upper approximation by replacing the universe U from the subset definition by a concept X. A *concept* B-lower approximation of the concept X is defined as follows:

$$\underline{B}X = \cup\{K_B(x) \mid x \in X, K_B(x) \subseteq X \}.$$

Obviously, the subset B-lower approximation of X is the same set as the concept B-lower approximation of X. A *concept* B-upper approximation of the concept X is defined as follows:

$$\overline{B}X = \cup\{K_B(x) \mid x \in X, K_B(x) \cap X \neq \emptyset \} = \cup\{K_B(x) \mid x \in X\}.$$

The concept B-upper approximation of X is a subset of the subset B-upper approximation of X. Besides, the concept B-upper approximations are truly the smallest sets containing X. For the decision presented in Table 1, the concept A-lower and A-upper approximations are

$$\underline{A}\{1, 4, 5, 7\} = \{4, 7\},$$

$$\underline{A}\{2, 3, 6\} = \{2, 3, 6\},$$

$$\overline{A}\{1, 4, 5, 7\} = \{1, 3, 4, 5, 6, 7\},$$

$$\overline{A}\{2, 3, 6\} = \{2, 3, 6\}.$$

Note that for complete decision tables, all three definitions of lower approximations, singleton, subset and concept, coalesce to the same definition. Also, for complete decision tables, all three definitions of upper approximations coalesce to the same definition. This is not true for incomplete decision tables, as our example shows.

4. Rule induction

For the inconsistent input data, LERS computes lower and upper approximations of all concepts. Rules induced from the lower approximation of the concept *certainly* describe the concept, so they are called *certain*. On the other hand, rules induced from the upper approximation of the concept describe the concept only *possibly* (or *plausibly*), so they are called *possible* [3].

The same idea of blocks of attribute-value pairs is used in a rule induction algorithm LEM2 [1, 3]. LEM2 explores the search space of attribute-value pairs. Its input data file is a lower or upper approximation of a concept, so its input data file is always consistent. In general, LEM2 computes a local covering [3] and then converts it into a rule set.

In our experiments we used MLEM2, a modified version of the algorithm LEM2. The original algorithm LEM2 needs discretization, a preprocessing, to deal with numerical attributes. The algorithm MLEM2 can induce rules from incomplete decision tables with numerical attributes. Its previous version induced certain rules from incomplete decision tables with missing attribute values interpreted as lost and with numerical attributes. Recently, MLEM2 was further extended to induce both certain and possible rules from a decision table with some missing attribute values being lost and some missing attribute values being "do not care" conditions, while some attributes may be numerical.

Since all characteristic sets $K_B(x)$, where $x \in U$, are intersections of blocks of attribute-value pairs, for attributes from B, and for subset and concept definitions of B-lower and B-upper approximations are unions of sets of the type $K_B(x)$, it is the most natural to use an algorithm based on blocks of attribute-value pairs, such as MLEM2 [1, 2] for rule induction.

The set of certain rules, induced from Table 1

Table 2. Error rates

	Approaches					
Data set	1	2	3	4	5	6
Breast cancer	29.02	30.42	30.42	29.37	30.42	30.42
Hepatitis	16.79	17.43	17.43	20.63	17.43	17.43
House	7.35	5.75	5.29	35.11	12.17	6.44
Primary tumor	68.45	60.19	63.14	72.87	63.44	63.44

for concept lower approximations is

(Hypertension, no) & (Age, 40..50) ->
(Delivery, fullterm)

(Hypertension, no) & (Age, 20..29) ->
(Delivery, fullterm)

(Age, 20..29) & (Hypertension, yes) ->
(Delivery, preterm)

and the corresponding possible rule set, induced form concept upper approximations is:

(Age, 30..39) -> (Delivery, fullterm)

(Hypertension, no) -> (Delivery, fullterm)

(Complications, none) -> (Delivery, fullterm)

(Age, 20..29) & (Hypertension, yes) ->
(Delivery, preterm)

5. Experiments

Different rough set approaches to rule induction from incomplete data were tested experimentally on real-life data. Four data sets were selected.

The *breast cancer* data set was obtained from the University Medical Center, Institute of Oncology, Ljubljana, Yugoslavia, due to donations from M. Zwitter and M. Soklic. Breast cancer is one of three data sets provided by the Oncology Institute that has repeatedly appeared in the machine learning literature. There are nine out of 286 examples containing unknown attribute values.

The *hepatitis* data set was donated by G. Gong, Carnegie-Mellon University, via Bojan Cestnik of Jozef Stefan Institute. There were 75 out of 155 examples that contain unknown attribute values in this data set.

The *house* data set, which has 203 examples that contain unknown attribute values, consists of votes of 435 congressmen in 1984 on 16 key-issues (yes or no).

The *primary-tumor* data set was obtained from the University Medical Center, Institute of Oncology, Ljubljana, Yugoslavia. The data set primary-tumor has 21 concepts and 17 attributes, and 207 out of 339 examples contain at least one missing value.

In our experiments we used two interpretations of missing attribute values: either we assumed that all missing attribute values were lost (denoted by "?") or that all of them were "do not care" conditions (denoted by "*"). Moreover, we tested two approaches to approximations: subset and concept. Finally, both certain rules (from lower approximations) and possible rules (from upper approximations) were induced. Since subset lower approximations, for any concept,

are equal to concept lower approximations, we ended up with six approaches:

1) missing attribute values interpreted as lost, concept definition of lower approximations,

2) missing attribute values interpreted as lost, subset definition of upper approximations,

3) missing attribute values interpreted as lost, concept definition of upper approximations,

4) missing attribute values interpreted as "do not care" conditions, concept definition of lower approximations,

5) missing attribute values interpreted as "do not care" conditions, subset definition of upper approximations,

6) missing attribute values interpreted as "do not care" conditions, concept definition of upper approximations.

The algorithm MLEM2 was used for rule induction and the LERS classification system was used to classify testing data against rules induced by MLEM2. For computing the error rate we used two-fold cross validation. All four data sets were divided into two halves and kept the same through testing all six approaches. Two-fold cross validation may be not the best tool to estimate the real error rate, but our objective was only to compare different approaches to missing attribute values. Results are presented in Table 2.

6. Conclusions

The idea of an attribute-value pair block, the main tool for data mining used in this paper, is both simple and useful. It is especially useful for incomplete decision tables, since it is used to determine characteristic sets, characteristic relations, lower and upper approximations, and, finally, it is used in rule induction.

We tested experimentally six different rough set approaches to missing attribute values. As follows from Table 2, the obvious conclusion is that interpreting missing attribute values as lost provides better results (smaller error rate) then interpreting missing attribute values as "do not care" conditions. However, is not that clear whether better are certain or possible rules and whether better are subset and concept approximations. There are other possible strategies to use certain and possible rule sets, e.g., using certain rules first and then possible rules, using both rule sets in parallel, etc. [7], so further research is required to explore these possibilities.

References

[1] C.-C. Chan and J. W. Grzymala-Busse. On the attribute redundancy and the learning programs ID3, PRISM, and LEM2. Department of Computer Science, University of Kansas, TR-91-14, December 1991, 20 pp.

[2] J. W. Grzymala-Busse. On the unknown attribute values in learning from examples. Proc. of the ISMIS-91, 6th International Symposium on Methodologies for Intelligent Systems, Charlotte, North Carolina, October 16–19, 1991, 368–377, *Lecture Notes in Artificial Intelligence*, vol. 542, Springer-Verlag, Berlin, Heidelberg, New York, 1991.

[3] J. W. Grzymala-Busse. LERS—A system for learning from examples based on rough sets. In *Intelligent Decision Support. Handbook of Applications and Advances of the Rough Sets Theory*, ed. by R. Slowinski, Kluwer Academic Publishers, Dordrecht, Boston, London, 1992, 3–18.

[4] J. W. Grzymala-Busse. Rough set strategies to data with missing attribute values. Workshop Notes, Foundations and New Directions of Data Mining, the 3-rd International Conference on Data Mining November 19–22, Melbourne, FL, USA, 56–63.

[5] J. W. Grzymala-Busse. Characteristic relations for incomplete data: a generalization of the indiscernibility relation. *Accepted for the RSCTC'2004, the Fourth International Conference on Rough Sets and Current Trends in Computing, Uppsala, Sweden, June 1–5, 2004.*

[6] J. W. Grzymala-Busse and A. Y. Wang.

Modified algorithms LEM1 and LEM2 for rule induction from data with missing attribute values. Proc. of the Fifth International Workshop on Rough Sets and Soft Computing (RSSC'97) at the Third Joint Conference on Information Sciences (JCIS'97), Research Triangle Park, NC, March 2–5, 1997, 69–72.

[7] J. W. Grzymala-Busse and C. P. B. Wang. Classification and rule induction based on rough sets. Proc. of the 5th IEEE International Conference on Fuzzy Systems FUZZ-IEEE'96, New Orleans, Louisiana, September 8–11, 1996, 744–747.

[8] M. Kryszkiewicz. Rough set approach to incomplete information systems. Proceedings of the Second Annual Joint Conference on Information Sciences, September 28–October 1, 1995, Wrightsville Beach, NC, 194–197.

[9] M. Kryszkiewicz. Rules in incomplete information systems. *Information Sciences* 113 (1999) 271–292.

[10] Z. Pawlak. *Rough Sets. Theoretical Aspects of Reasoning about Data*. Kluwer Academic Publishers, Dordrecht, Boston, London, 1991.

[11] J. Stefanowski. *Algorithms of Decision Rule Induction in Data Mining*. Poznan University of Technology Press, Poznan, Poland, 2001.

[12] J. Stefanowski and A. Tsoukias. On the extension of rough sets under incomplete information. Proceedings of the 7th International Workshop on New Directions in Rough Sets, Data Mining, and Granular-Soft Computing, RSFDGrC'1999, Yamaguchi, Japan, 73–81.

[13] J. Stefanowski and A. Tsoukias. Incomplete information tables and rough classification. *Computational Intelligence* 17 (2001) 545–566.

[14] Y. Y. Yao. Two views of the theory of rough sets in finite universes. *International J. of Approximate Reasoning* 15 (1996) 291–317.

[15] Y. Y. Yao. Relational interpretations of neighborhood operators and rough set approximation operators. *Information Sciences* 111 (1998) 239–259.

[16] Y. Y. Yao. On the generalizing rough set theory. Proc. of the 9th Int. Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing (RSFDGrC'2003), Chongqing, China, Oct. 19–22, 2003, 44–51.