Local and Global Approximations for Incomplete Data

Jerzy W. Grzymala-Busse[1,2] and Wojciech Rzasa[3] *

[1] Department of Electrical Engineering and Computer Science University of Kansas, Lawrence, KS 66045, USA
[2] Institute of Computer Science, Polish Academy of Sciences, 01–237 Warsaw, Poland
[3] Institute of Mathematics, University of Rzeszow, 35–310 Rzeszow, Poland

Abstract. For completely specified decision tables, where lower and upper approximations are unique, the lower approximation is the largest definable set contained in the approximated set X and the upper approximation of X is the smallest definable set containing X. For incomplete decision tables the existing definitions of upper approximations provide sets that, in general, are not minimal definable sets. The same is true for approximations based on relations that are generalizations of the equivalence relation. In this paper we introduce two definitions of approximations, local and global, such that the corresponding upper approximations are minimal. Local approximations are more precise than global approximations. Global lower approximations may be determined by a polynomial algorithm. However, algorithms to find both local approximations and global upper approximations are NP-hard.

1 Introduction

Recently we observed intensive research activity in two areas: rough set approaches to handle incomplete data, mostly in the form of decision tables with missing attribute values, and attempts to study generalizations of the standard indiscernibility relation. In the latter area concerned relations are not equivalence relations. Our paper contributes to both research areas.

Initially rough set theory was applied to complete data sets (with all attribute values specified). Recently rough set theory was extended to handle incomplete data sets (with missing attribute values) [1–9, 17–20].

We will distinguish two types of missing attribute values. The first type of missing attribute value will be called *lost*. A missing attribute value is lost when for some case (example, object) the corresponding attribute value was mistakenly erased or not entered into the data set.

The second type of missing attribute values, called *"do not care" conditions*,

* This research has been partially supported by the Ministry of Scientific Research and Information Technology of the Republic of Poland, grant 3 T11C 005 28

are based on an assumption that missing attribute values were initially, when the data set was created, irrelevant. The corresponding cases were classified even though the values of these attribute were not known. A missing attribute value of this type may be potentially replaced by any value typical for that attribute.

For incomplete decision tables there are two special cases: in the first case, all missing attribute values are lost, in the second case, all missing attribute values are "do not care" conditions. Incomplete decision tables in which all attribute values are lost, from the viewpoint of rough set theory, were studied for the first time in [6], where two algorithms for rule induction, modified to handle lost attribute values, were presented. This approach was studied later, e.g., in [18, 19], where the indiscernibility relation was generalized to describe such incomplete data. Furthermore, an approach to incomplete data based on relative frequencies was presented in [19]. Another approach, using fuzzy set ideas, was presented in [1].

On the other hand, incomplete decision tables in which all missing attribute values are "do not care" conditions, from the view point of rough set theory, were studied for the first time in [2], where a method for rule induction was introduced in which each missing attribute value was replaced by all values from the domain of the attribute. Originally such values were replaced by all values from the entire domain of the attribute, later, by attribute values restricted to the same concept to which a case with a missing attribute value belongs. Such incomplete decision tables, with all missing attribute values being "do not care conditions", were extensively studied in [8, 9], including extending the idea of the indiscernibility relation to describe such incomplete decision tables.

In general, incomplete decision tables are described by characteristic relations, in a similar way as complete decision tables are described by indiscernibility relations [3–5].

In rough set theory, one of the basic notions is the idea of lower and upper approximations. For complete decision tables, once the indiscernibility relation is fixed and the concept (a set of cases) is given, the lower and upper approximations are unique.

For incomplete decision tables, for a given characteristic relation and concept, there are three important and different possibilities to define lower and upper approximations, called singleton, subset, and concept approximations [3]. Singleton lower and upper approximations were studied in [8, 9, 16, 18, 19]. Note that similar three definitions of lower and upper approximations, though not for incomplete decision tables, were studied in [10–12, 21–24].

Our main objective is to study two novel kinds of approximations: local and global. The local approximations are defined using sets of attribute-value pairs called complexes, while the global approximations are formed from characteristic sets. Additionally, lower approximations, local and global, are the maximal sets that are locally and globally definable, respectively, and contained in the approximated set X. Similarly, upper approximations, local and global, are the minimal sets that are locally and globally definable, respectively, containing the approximated set X.

Note that some other rough-set approaches to missing attribute values were presented in [1, 2] as well.

2 Blocks of Attribute-Value Pairs

We assume that the input data sets are presented in the form of a *decision table*. An example of a decision table is shown in Table 1. Rows of the decision table represent *cases*, while columns are labeled by *variables*. The set of all cases will be denoted by U. In Table 1, $U = \{1, 2, ..., 8\}$. 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, $A = \{Temperature, Headache, Nausea\}$. 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, $\rho(1, Temperature) = high$. A decision table with completely specified function ρ will be called *completely specified*, or, for the sake of simplicity, *complete*. In practice, input data for data mining are frequently affected by missing attribute values. In other words, the corresponding function ρ is incompletely specified (partial). A decision table with an incompletely specified function ρ will be called *incomplete*. Function ρ describing Table 1 is incompletely specified.

Table 1. An incomplete decision table

	Attributes			Decision
Case	Temperature	Headache	Nausea	Flu
1	high	?	no	yes
2	very_high	yes	yes	yes
3	?	no	no	yes
4	high	yes	yes	yes
5	high	?	yes	yes
6	normal	yes	no	yes
7	normal	no	yes	no
8	*	yes	*	no

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 lost values will be denoted by "?" and "do not care" conditions by "*". Additionally, we will assume that for each case at least one attribute value is specified.

An important tool to analyze complete decision tables is a block of the attribute-value pair. Let a be an attribute, i.e., $a \in A$ 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. For incomplete decision tables, a block of an attribute-value pair must be modified in the following way:

- 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 blocks$[(a, v)]$ 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 case x should be included in blocks $[(a, v)]$ for all specified values v of attribute a.

Thus,

[(Temperature, high)] = {1, 4, 5, 8},
[(Temperature, very_high)] = {2, 8},
[(Temperature, normal)] = {6, 7, 8},
[(Headache, yes)] = {2, 4, 6, 8},
[(Headache, no)] = {3, 7},
[(Nausea, no)] = {1, 3, 6, 8},
[(Nausea, yes)] = {2, 4, 5, 7, 8}.

For a case $x \in U$ the *characteristic set* $K_B(x)$ is defined as the intersection of the sets $K(x, a)$, for all $a \in B$, where the set $K(x, a)$ is defined in the following way:

- If $\rho(x, a)$ is specified, then $K(x, a)$ is the block $[(a, \rho(x, a)]$ of attribute a and its value $\rho(x, a)$,
- If $\rho(x, a) =?$ or $\rho(x, a) = *$ then the set $K(x, a) = U$.

For Table 1 and $B = A$,

$K_A(1) = \{1, 4, 5, 8\} \cap U \cap \{1, 3, 6, 8\} = \{1, 8\}$,
$K_A(2) = \{2, 8\} \cap \{2, 4, 6, 8\} \cap \{2, 4, 5, 7, 8\} = \{2, 8\}$,
$K_A(3) = U \cap \{3, 7\} \cap \{1, 3, 6, 8\} = \{3\}$,
$K_A(4) = \{1, 4, 5, 8\} \cap \{2, 4, 6, 8\} \cap \{2, 4, 5, 7, 8\} = \{4, 8\}$,
$K_A(5) = \{1, 4, 5, 8\} \cap U \cap \{2, 4, 5, 7, 8\} = \{4, 5, 8\}$,
$K_A(6) = \{6, 7, 8\} \cap \{2, 4, 6, 8\} \cap \{1, 3, 6, 8\} = \{6, 8\}$,
$K_A(7) = \{6, 7, 8\} \cap \{3, 7\} \cap \{2, 4, 5, 7, 8\} = \{7\}$, and
$K_A(8) = U \cap \{2, 4, 6, 8\} \cap U = \{2, 4, 6, 8\}$.

Characteristic set $K_B(x)$ may be interpreted as the set of cases that are indistinguishable from x using all attributes from B and 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. In [22] $K_A(x)$ was called a successor neighborhood of x, see also [10–12, 16, 21, 23, 24].

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

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

The characteristic relation $R(B)$ is reflexive but—in general—does not need to be symmetric or transitive. Also, the characteristic relation $R(B)$ is known if we know characteristic sets $K_B(x)$ for all $x \in U$. In our example, $R(A) = \{(1, 1), (1, 8), (2, 2), (2, 8), (3, 3), (4, 4), (4, 8), (5, 4), (5, 5), (5, 8), (6, 6), (6, 8), (7, 7), (8, 2), (8, 4), (8, 6), (8, 8)\}$. The most convenient way to define the characteristic relation is through the characteristic sets.

For decision tables, in which all missing attribute values are lost, a special characteristic relation was defined in [18], see also, e.g., [17, 19].

For decision tables where all missing attribute values are "do not care" conditions a special characteristic relation was defined in [8], see also, e.g., [9].

3 Definability

Let $B \subseteq A$. For completely specified decision tables, any union of elementary sets of B is called a B-definable set [14]. Definability for completely specified decision tables should be modified to fit into incomplete decision tables. For incomplete decision tables, a union of some intersections of attribute-value pair blocks, in any such intersection all attributes should be different and attributes are members of B, will be called B-*locally definable* sets. A union of characteristic sets $K_B(x)$, where $x \in X \subseteq U$ will be called a B-*globally definable* set. Any set X that is B-globally definable is B-locally definable, the converse is not true. In the example of Table 1, the set {7, 8} is A-locally-definable since it is equal to the intersection of [(Temperature, normal)] and [(Nausea, yes)]. Nevertheless, {7, 8} is not A-globally-definable.

Obviously, if a set is not B-locally definable then it cannot be expressed by rule sets using attributes from B. This is why it is so important to distinguish between B-locally definable sets and those that are not B-locally definable.

4 Local Approximations

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, locally or globally. A set T of attribute-value pairs, where all attributes are distinct and in B, will be called a *B-complex*. For a set T of attribute-value pairs, the intersection of blocks for all t from T will be denoted by $[T]$.

For incomplete decision tables lower and upper approximations may be defined in a few different ways, see, e.g., [3–5]. In this paper we introduce a new idea of optimal approximations that are B-locally definable. Let $B \subseteq A$. The *B-local lower* approximation of the concept X, denoted by $L\underline{B}X$, is defined as follows

$$\cup\{[T] \mid T \textit{ is a } B\textit{–complex of } X,\ [T] \subseteq X\}.$$

The *B-local upper* approximation of the concept X, denoted by $L\overline{B}X$, is a set with the minimal cardinality containing X and defined in the following way

$$\cup\{[T] \mid \exists \text{ a family } \mathcal{T} \text{ of } B\text{–complexes } T \text{ of } X \text{ with } \forall\ T \in \mathcal{T},\ [T] \cap X \neq \emptyset\}.$$

Obviously, the B-local lower approximation of X is unique and it is the maximal B-locally definable set contained in X. Any B-local upper approximation of X is B-locally definable, it contains X, and is, by definition, minimal.

For Table 1

$$L\underline{A}\{1,2,3,4,5,6\} = [(Headache, no)] \cap [(Nausea, no)] = \{3\},$$

so one complex, {(Headache, no), (Nausea, no)}, is sufficient to describe $L\underline{A}\{1,2,3,4,5,6\}$,

$$L\underline{A}\{7,8\} = [(Temperature, normal)] \cap [(Nausea, yes)] = \{7,8\},$$

so again, one complex, {(Temperature, normal), (Nausea, yes)}, describes $L\underline{A}\{7,8\}$,

$$L\overline{A}\{1,2,3,4,5,6\} =$$
$$[(Temperature, high)] \cup [(Headache, yes)] \cup [(Nausea, no)] =$$
$$\{1,2,3,4,5,6,8\},$$

therefore, to describe $L\overline{A}\{1,2,3,4,5,6\}$ three complexes are necessary: {(Temperature, high)}, {(Headache, yes)}, and {(Nausea, no)}. Finally,

$$L\overline{A}\{7,8\} = [(Temperature, normal)] \cap [(Nausea, yes)] = \{7,8\}.$$

For the incomplete decision table from Table 1 the local lower approximations for both concepts, {1, 2, 3, 4, 5, 6} and {7, 8}, as well as the upper local approximations for these concepts, are unique. Though the local lower approximations are always unique, the local upper approximations, in general, are not unique. For example, let us consider an incomplete decision table from Table 2.

For Table 2

[(Age, <25)] = {1, 4, 6},
[(Age, 25..35)] = {1, 4, 7},
[(Age, >35)] = {1, 2, 3, 4, 5},
[(Complications, alcoholism)] = {1},
[(Complications, obesity)] = {2, 3},
[(Complications, none)] = {4, 5, 6, 7},
[(Hypertension, mild)] = {1}.
[(Hypertension, severe)] = {2}.
[(Hypertension, no)] = {4, 5, 6, 7}.

Moreover, for Table 2

$$L\underline{A}\{1,2,3,4\} =$$
$$[(Complications, alcoholism)] \cup [(Complications, obesity)] =$$
$$\{1,2,3\},$$

Table 2. An incomplete decision table

	Attributes			Decision
Case	Age	Complications	Hypertension	Delivery
1	*	alcoholism	mild	pre-term
2	>35	obesity	severe	pre-term
3	>35	obesity	?	pre-term
4	*	none	none	pre-term
5	>35	none	none	full-term
6	<25	none	none	full-term
7	25..35	none	none	full-term

$$L\underline{A}\{5,6,7\} = \emptyset,$$

However,

$$L\overline{A}\{1,2,3,4\}$$

is not unique, any of the following sets

$$[(Age, > 35)] = \{1,2,3,4,5\},$$

$$[(Age, < 25)] \cup [(Complications, obesity)] = \{1,2,3,4,6\},$$

or

$$[(Age, 26..35)] \cup [(Complications, obesity)] = \{1,2,3,4,7\}.$$

may serve as local upper approximations of {1, 2, 3, 4}.

Lastly,

$$L\overline{A}\{5,6,7\} = [(Complications, none)] = \{4,5,6,7\}.$$

Algorithms to compute local lower or upper approximations are NP-hard, since the corresponding problems may be presented in terms of prime implicants, monotone functions, and minimization. A similar result for reducts of complete decision tables is well known [15].

5 Global Approximations

Again, let $B \subseteq A$. Then *B-global lower* approximation of the concept X, denoted by $G\underline{B}X$, is defined as follows

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

Note that the definition of global lower approximation is identical with the definition of subset (or concept) lower approximation [3–5]. The *B-global upper* approximation of the concept X, denoted by $G\overline{B}X$, is a set with the minimal cardinality containing X and defined in the following way

$$\cup\{K_B(x) \mid \exists\ Y \subseteq U,\ x \in Y, K_B(x) \cap X \neq \emptyset\}.$$

Similarly as for local approximations, a global lower approximation for any concept X is unique. Additionally, both B-global approximations, lower and upper, are B-globally definable. On the other hand, global upper approximations do not need to be unique. For Table 1,

$$G\underline{A}\{1,2,3,4,5,6\} = K_A(3) = \{3\},$$

$$G\underline{A}\{7,8\} = K_A(7) = \{7\},$$

$$G\overline{A}\{1,2,3,4,5,6\} =$$
$$K_A(1) \cup K_A(2) \cup K_A(3) \cup K_A(5) \cup K_A(6) = \{1,2,3,4,5,6,8\}.$$

Furthermore,

$$G\overline{A}\{7,8\}$$

may be computed in four different ways:

(1) as $K_A(1) \cup K_A(7) = \{1,7,8\}$,

(2) as $K_A(2) \cup K_A(7) = \{2,7,8\}$,

(3) as $K_A(4) \cup K_A(7) = \{4,7,8\}$,

(4) or as $K_A(6) \cup K_A(7) = \{6,7,8\}$,

all four sets are global upper approximations of the concept {7, 8}.

In general, local approximations are more precise than global approximations. For any concept X and a subset B of A,

$$L\underline{B}X \supseteq G\underline{B}X$$

and

$$L\overline{B}X \subseteq G\overline{B}X.$$

It is not difficult to find a simple algorithm to compute global lower approximations in polynomial time. Nevertheless, algorithms to compute global upper approximations are NP-hard as well.

6 Conclusions

In this paper we introduced two new kinds of approximations: local and global. These approximations describe optimally approximated sets (lower approximations are maximal, upper approximations are minimal and, at the same time, local approximations are locally definable while global approximations are globally definable).

Note that our global approximations may be used to describe behavior of systems defined by relations that are not equivalence relations, as in [10–12, 16, 21–24].

As a final point, optimality comes with the price: algorithms to compute both local upper approximations and global upper approximations are NP-hard.

References

1. Greco, S., Matarazzo, B., and Slowinski, R.: Dealing with missing data in rough set analysis of multi-attribute and multi-criteria decision problems. In *Decision Making: Recent Developments and Worldwide Applications*, ed. by S. H. Zanakis, G. Doukidis, and Z. Zopounidis, Kluwer Academic Publishers, Dordrecht, Boston, London, 2000, 295–316.
2. Grzymala-Busse, J.W.: 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. Lecture Notes in Artificial Intelligence, vol. 542, Springer-Verlag, Berlin, Heidelberg, New York (1991) 368–377.
3. Grzymala-Busse, J.W.: 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, Melbourne, FL, USA, November 19–22, 2003, 56–63.
4. Grzymala-Busse, J.W.: Data with missing attribute values: Generalization of idiscernibility relation and rule induction. *Transactions on Rough Sets*, Lecture Notes in Computer Science Journal Subline, Springer-Verlag, vol. **1** (2004) 78–95.
5. Grzymala-Busse, J.W.: Characteristic relations for incomplete data: A generalization of the indiscernibility relation. Proc. of the RSCTC'2004, the Fourth International Conference on Rough Sets and Current Trends in Computing, Uppsala, Sweden, June 1–5, 2004. Lecture Notes in Artificial Intelligence 3066, Springer-Verlag 2004, 244–253.
6. Grzymala-Busse, J.W. and Wang A.Y.: 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. Hong, T.P., Tseng L.H. and Chien, B.C.: Learning coverage rules from incomplete data based on rough sets. Proc. of the IEEE International Conference on Systems, Man and Cybernetics, Hague, the Netherlands, October 10–13, 2004, 3226–3231.
8. Kryszkiewicz, M.: Rough set approach to incomplete information systems. Proc. of the Second Annual Joint Conference on Information Sciences, Wrightsville Beach, NC, September 28–October 1, 1995, 194–197.

9. Kryszkiewicz, M.: Rules in incomplete information systems. *Information Sciences* **113** (1999) 271–292.
10. Lin, T.Y.: Neighborhood systems and approximation in database and knowledge base systems. Fourth International Symposium on Methodologies of Intelligent Systems (Poster Sessions), Charlotte, North Carolina, October 12–14, 1989, 75–86.
11. Lin, T.Y.: Chinese Wall security policy—An aggressive model. Proc. of the Fifth Aerospace Computer Security Application Conference, Tucson, Arizona, December 4–8, 1989, 286–293.
12. Lin, T.Y.: Topological and fuzzy 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) 287–304.
13. Pawlak, Z.: Rough Sets. *International Journal of Computer and Information Sciences* **11** (1982) 341–356.
14. Pawlak, Z.: *Rough Sets. Theoretical Aspects of Reasoning about Data*. Kluwer Academic Publishers, Dordrecht, Boston, London (1991).
15. Skowron, A. and Rauszer, C.: The discernibility matrices and functions in information systems. In *Handbook of Applications and Advances of the Rough Sets Theory*, ed. by R. Slowinski, Kluwer Academic Publishers, Dordrecht, Boston, London (1992) 331–362.
16. Slowinski, R. and Vanderpooten, D.: A generalized definition of rough approximations based on similarity. *IEEE Transactions on Knowledge and Data Engineering* **12** (2000) 331–336.
17. Stefanowski, J.: *Algorithms of Decision Rule Induction in Data Mining*. Poznan University of Technology Press, Poznan, Poland (2001).
18. Stefanowski, J. and Tsoukias, A.: On the extension of rough sets under incomplete information. Proc. of the 7th International Workshop on New Directions in Rough Sets, Data Mining, and Granular-Soft Computing, RSFDGrC'1999, Ube, Yamaguchi, Japan, November 8–10, 1999, 73–81.
19. Stefanowski, J. and Tsoukias, A.: Incomplete information tables and rough classification. *Computational Intelligence* **17** (2001) 545–566.
20. Wang, G.: Extension of rough set under incomplete information systems. Proc. of the IEEE International Conference on Fuzzy Systems (FUZZ_IEEE'2002), vol. 2, Honolulu, HI, May 12–17, 2002, 1098–1103.
21. Yao, Y.Y.: Two views of the theory of rough sets in finite universes. *International J. of Approximate Reasoning* **15** (1996) 291–317.
22. Yao, Y.Y.: Relational interpretations of neighborhood operators and rough set approximation operators. *Information Sciences* **111** (1998) 239–259.
23. Yao, Y.Y.: 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, October 19–22, 2003, 44–51.
24. Yao, Y.Y. and Lin, T.Y.: Generalization of rough sets using modal logics. *Intelligent Automation and Soft Computing* **2** (1996) 103–119.