Local and Global Approximations for Incomplete Data

Jerzy W. Grzymała-Busse[1,2] and Wojciech Rząsa[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] Department of Computer Science, University of Rzeszow, 35–310 Rzeszow, Poland

Abstract. For completely specified decision tables 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 generalizations of approximations based on relations that are not equivalence relations. 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. Additionally, we show that for decision tables with all missing attribute values being lost, local and global approximations are equal to one another and that they are unique.

1 Introduction

Development of appropriate methodology to incomplete data sets is crucial since many real-life data sets have missing attribute values. Mining incomplete data requires either a preprocessing (filling in missing attribute values before the main process of rule set induction, decision tree generation, etc.) or mining the data set taking into account that it is incomplete. In this paper we will use the latter approach.

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) [1,2,3,4,5,6,8,9,10,11,20,21,22,23]. We observe 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, in many attempts to study generalizations of the standard indiscernibility relation used to describe decision tables. In the latter area concerned relations are not equivalence relations. Our paper contributes to both research areas.

J.F. Peters and A. Skowron (Eds.): Transactions on Rough Sets VIII, LNCS 5084, pp. 21–34, 2008.
© Springer-Verlag Berlin Heidelberg 2008

In general, incomplete decision tables are described by characteristic relations, in a similar way as complete decision tables are described by indiscernibility relations [3,4,5,6].

In spite of the fact that input data are presented as decision tables in applications of rough set theory, in theory oriented research such information is frequently expressed as approximation spaces and neighborhood systems [12,17].

Our main objective is to study two novel kinds of approximations: local and global. Both of the two kinds of approximations are optimal in some sense. It means that lower approximations, local and global, are the largest sets that are locally and globally definable, respectively, and contained in the approximated set X. Similarly, upper approximations, local and global, are the smallest sets that are locally and globally definable, respectively, containing the approximated set X. As it will be shown the two kinds of approximations coincide for complete data, and they may differ for incomplete data sets.

A preliminary version of this paper was presented at the Fifth International Conference on Rough Sets and Current Trends in Computing, Kobe, Japan, November 6–8, 2006 [7].

2 Basic Notions

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, $U = \{1, 2, ..., 8\}$. Some variables are called *attributes* while one selected 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*.

For a complete decision table *indiscernibility* relation ind_B is defined according to the following formula

$$ind_B = \{(x, y) \in U^2 \mid \rho(x, a) = \rho(y, a), a \in B\}.$$

For any $B \subseteq A$, ind_B is an equivalence relation. Let $[x]_B$ denotes the equivalence class containing x with respect to the relation ind_B, let $I_A = \{[x]_A \mid x \in U\}$ and $I = \{[x]_B | x \in U, B \subseteq A\}$. It is known that every set $X \subseteq U$ may be presented as a union of some elements of the family I if and only if it can be presented as a union of some elements of the family I_A. Elements of the family I_A are called *elementary* sets. Every set $X \subseteq U$ that is a union of some elementary sets is called *definable*. We assume that the empty set is definable and we denote the family of all definable sets by D. It was observed in [16,17] that a pair (U, D) is a topological space with a topology of open-closed sets. This topology is equivalent to a topology defined by a base I_A as well as defined by a base I. A family I_A is a base of D with the smallest cardinality. The largest definable set $\underline{X}$ contained

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

in $X \subseteq U$ will be called a *lower approximation* of X. The smallest definable set $\overline{X}$ containing $X \subseteq U$ will be called an *upper approximation* of X. Therefore in the topological space (U, I), the lower approximation of the set is its *interior* and the upper approximation of the set is its *closure*. Thus in any discussion on definability and approximations of the set, we may restrict ourselves to elements of the space (U, I_A). This is an important property since the set I_A is easy to compute.

In the topological space (U, I), the set $X \subseteq U$ will be called *J-definable* if it can be presented as a union of some elements of the family J, where $J \subseteq I$. Obviously, if $J_1 \subseteq J_2$, then every set J_1-definable is also J_2-definable, [16,17].

3 Incomplete Data Sets

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*.

For the rest of the paper we will discuss incomplete data sets such that for each case at least one attribute value is specified and all decision values are specified. In this paper 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 forgotten to enter into the data set. The original value existed but for a variety of reasons now it is not accessible.

The second type of missing attribute values, called *"do not care" conditions*, are based on an assumption that missing attribute values were initially, when the data set was created, irrelevant. For example, in a medical setup, patients were subjected to preliminary tests. Patients whose preliminary test results were negative were diagnosed as not affected by a disease. They were perfectly well

diagnosed in spite of the fact that not all tests were conducted on them. Thus some test results are missing because these tests were redundant. In different words, a missing attribute value of this type may be potentially replaced by any value typical for that attribute. This type of a missing attribute value will be called a "do not care" condition.

Note that both types of missing attribute values are universal (or standard), since they can be used for any incomplete data set. Obviously, if we are familiar with the reason why some attribute values are missing, we should apply the appropriate interpretation: lost values or "do not care" conditions.

For the rest of the paper we will denote lost values by "?" and "do not care" conditions by "*". An example of incomplete decision table is shown in Table 1.

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 [8], where two algorithms for rule induction, modified to handle lost attribute values, were presented. This approach was studied later, e.g., in [21] and [22], where the indiscernibility relation was generalized to describe such incomplete decision tables.

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 [10], [11], including extending the idea of the indiscernibility relation to describe such incomplete decision tables.

Other types of missing attribute values are possible as well, see, e.g., [6]. Moreover, note that some other rough-set approaches to missing attribute values were presented in, e.g., [1,2,15].

4 Blocks of Attribute-Value Pairs

An important tool to analyze 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 specified value of a for some case. A block of an attribute-value pair is defined in the following way:

- If for a specified 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 a specified 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.

Alternatively, a block of the pair (a, v) is defined according to the following formula:

$$[(a, v)] = \{x \in U | \rho(x, a) = v \text{ or } \rho(x, a) = *\}.$$

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 data sets with other types of missing attribute values, the definition of the attribute-value block is modified, see, e.g., [6].

5 Definability

As it was mentioned in Section 1, for complete data sets the family I_A is the set of all elementary sets. Additionally, the cardinality of the set I_A is smaller than or equal to the cardinality of the set U. Thus testing whether a set X is definable is—computationally—a simple task. For incomplete data the situation is different. 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 $K(x, a) = U$.

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 [25] $K_A(x)$ was called a successor neighborhood of x, see also [12,13,14,19,24,26,27].

Let $K = \{K_B(x) | x \in U, B \subseteq A\}$ and $K_B = \{K_B(x) \mid x \in U\}$, for $B \subseteq A$. For Table 1 members of the family K_A are:

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

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 [21], see also, e.g., [20,22].

For decision tables where all missing attribute values are "do not care" conditions a special characteristic relation was defined in [10], see also, e.g., [11].

For incomplete data sets, a set X will be called *B-globally definable* if it is K_B-definable, i.e., if X is a union of members of the family K_B. A set that is A-globally definable will be called *globally definable*. Obviously, the cardinality of the set K_A is smaller than or equal to the cardinality of U, so checking whether a set is globally definable is computationally easy. However, in general K_A is not a base of the approximation space (U, K). Thus X may be K_B-definable in this space (for some $B \subseteq A$) in spite of the fact that it is not K_A-definable. Moreover, a set may be K-definable and be not K_B-definable for any fixed $B \subseteq A$. The family K may have much greater cardinality than the family K_A. As a consequence, the problem of checking whether a set is definable in the space (U, K) is computationally complex. Similarly, searching for a base of the space (U, K) may be computationally complex.

For incomplete data set it is advantageous to define a local definability. A set T of attribute-value pairs, where all attributes belong to set B and are distinct, will be called a *B-complex*. Any A-complex will be called—for simplicity—a *complex*. Obviously, any set containing a single attribute-value pair is a complex. For the rest of the paper we will discuss only *nontrivial complexes*, i.e., such complexes that the intersection of all attribute-value blocks from a given complex is not the empty set.

Set X *depends* on a complex T if and only if

$$\emptyset \neq [T] = \bigcap \{[t] \mid t \in T\} \subseteq X.$$

For an incomplete decision table and a subset B of the set A of all attributes, a union of intersections of attribute-value pair blocks of attribute-value pairs from some B-complexes, will be called a *B-locally definable* set. A-*locally definable* sets will be called locally definable. If (U, A, d, V, ρ) is an incomplete decision table, then the space (U, L) is an approximation space, where L is a family of all subsets of the set of all possible intersections of attribute-value blocks, members of complexes. Obviously, $L \supseteq K$ for any decision table. Thus, the computational complexity of the problems of looking for a minimal base of the space (U, L) and checking whether a set is locally definable is exponential.

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 A-locally definable since it is equal to the intersection of [(Temperature, normal)] and [(Nausea, yes)]. Nevertheless, the set {7, 8} is not A-globally definable.

The importance of the idea of local definability is a consequence of the following fact: A set is locally definable if and only if it can be expressed by decision rule sets. This is why it is so important to distinguish between locally definable sets and those that are not locally definable.

For decision tables in which all missing attribute values are lost, local definability is reduced to global definability. The proof of this fact will be given in Section 7.

Note that definability, introduced in [19], differs from our definitions. For example, the set {1, 2, 4, 6, 8}, globally definable according to our definition, is not definable in [19]. Additionally, sets that are definable in [19], are not even locally definable according to our definition.

6 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.

Let $B \subseteq A$. The *B-local lower* approximation of the concept X, denoted by $L\underline{B}X$, is defined as follows

$$\bigcup \{[T] \mid T \text{ is a complex of } X,\ [T] \subseteq X\}.$$

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

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

Obviously, the B-local lower approximation of X is unique and it is the largest 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, the smallest.

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)}, describes $L\underline{A}\{1,2,3,4,5,6\}$,

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

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)}, {(Nausea, no)}. Finally,

$$L\overline{A}\{7,8\} = L\underline{A}\{7,8\} = \{7,8\}.$$

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.

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

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, none)] = {4, 5, 6, 7}.

Moreover, for Table 2

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

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

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 [18].

7 Global Approximations

For incomplete decision tables global lower and upper approximations may be defined in a few different ways, see, e.g., [3,4,5]. In this paper we suggest yet another definition of global approximations. Note that our definition of global approximations is based on characteristic sets, as oppose to local approximations, introduced in the previous section, where attribute-value blocks were used.

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

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

Let us observe that the definition of global lower approximation is identical with the definition of subset (or concept) lower approximation [3,4,5]. The *B-global upper* approximation of the concept X, denoted by $G\overline{B}X$, is defined as the minimal set containing X and defined in the following way

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

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 approximation in polynomial time. Nevertheless, algorithms to compute global upper approximations are NP-hard as well.

On the other hand, determining local and global approximations is quite simple for incomplete data sets with all missing values being *lost*. For decision tables with all missing values being lost, the following results hold:

Lemma 1. Let the only missing attribute values in a decision table be lost. Let $x, y \in U$, let B be a subset of the attribute set A, and let $y \in K_B(x)$. Then $K_B(y) \subseteq K_B(x)$.

Proof. Let $a_1, a_2, ..., a_n$ be all attributes from B such that $\rho(x, a_i) = v_i$ is specified (i.e., $v_i \neq ?$) for all $i = 1, 2, ..., n$. Then $K_B(x)$ is equal to

$$[(a_1, v_1)] \cap [(a_2, v_2)] \cap ... \cap [(a_n, v_n)].$$

If $y \in K_B(x)$ then $y \in [(a_1, v_1)]$, $y \in [(a_2, v_2)]$,... $y \in [(a_n, v_n)]$. Moreover, $\rho(y, a_i)$ are all specified, for $i = 1, 2, ..., n$, since all missing attribute values are lost. Obviously, it is possible that for some $a \in A - B$, $\rho(y, a)$ is specified as well. Thus $K_B(y)$ is a subset of the following set $[(a_1, v_1)] \cap [(a_2, v_2)] \cap ... \cap [(a_n, v_n)]$, or,

$$K_B(y) \subseteq K_B(x).$$

Lemma 2. Let the only missing attribute values in a decision table be lost. Let T be a nontrivial complex. Let B be the set of all attributes involved in T. There exists $x \in U$ such that $[T] = K_B(x)$.

Proof. Let $\{a_1, a_2, ..., a_n\} = B$. Let $T = \{(a_1, v_1), (a_2, v_2), ..., (a_n, v_n)\}$, where $v_1, v_2, ...v_n$ are values of $a_1, a_2, ..., a_n$, respectively. Then for any $x \in [T]$, $\rho(x, a_1) = v_1, \rho(x, a_2) = v_2, ..., \rho(x, a_n) = v_n$, since all missing attribute values are lost. Therefore, $[T] = K_B(x)$.

Lemma 3. Let the only missing attribute values in a decision table be lost. Let B be a subset of the set A of all attributes. For every nontrivial complex T such that all attributes involved in T are in B there exists a subset X of U such that

$$[T] = \bigcup \{K_B(x) \mid x \in X\}.$$

Proof. Let C be the set of all attributes involved in T, i.e., if $T = \{(a_1, v_1), (a_2, v_2), ..., (a_n, v_n)\}$, where $v_1, v_2, ...v_n$ are values of $a_1, a_2, ..., a_n$, respectively, then $C = \{a_1, a_2, ..., a_n\}$. There exists $y \in U$ such that $[T] = K_C(y)$, by Lemma 2. Let $X = [T]$. Thus, $\cup\{K_B(y) \mid y \in X\} \subseteq [T]$ since $C \subseteq B$. On the other hand, $[T] \subseteq \cup\{K_B(y) \mid y \in X\}$ since $y \in K_B(y)$. Thus $\cup\{K_B(y) \mid y \in X\} = [T]$.

Due to our last result, we may observe that for data sets with the only missing attribute values being lost, sets $K_B(x)$, for $x \in U$ and $B \subseteq A$, are finer granules (subsets of U) than any nontrivial B-complexes, since any nontrivial B-complex is a union of some sets $K_B(x)$.

Theorem 1. Let the only missing attribute values in a decision table be lost and let B be a subset of the attribute set A. Then every subset $X \subseteq U$ is B-locally definable if and only if it is B-globally definable.

Proof. Straightforward, due to Lemma 3.

Theorem 2. Let the only missing attribute values in a decision table be lost and let B be a subset of the attribute set A. Then for every concept $X \subseteq U$, its B-local approximations are equal to its B-global approximations. Moreover, computing such approximations is of polynomial computational complexity.

Proof. For lower approximations, the proof is obvious since $L\underline{B}X \supseteq G\underline{B}X$ and Lemma 3. Similarly for upper approximations since $L\overline{B}X \subseteq G\overline{B}X$ and Lemma 3. Thus, we may compute the lower approximation (B-local or B-global) using the following formula

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

and, by analogy, the upper approximation (also, B-local or B-global) using the following formula

$$G\overline{B}X = \bigcup \{K_B(x) \mid x \in X\}.$$

The last formula needs some explanation. The B-global upper approximation $G\overline{B}X$ is defined as a minimal set satisfying the following formula: $\bigcup\{K_B(x) \mid \exists\, Y$

$\subseteq U \ \forall \ x \in Y,\ K_B(x) \cap X \neq \emptyset\}$. Any $y \in U - X$ such that $K_B(y) \cap X \neq \emptyset$ may be ignored as a member of Y since if $x \in K_B(y) \cap X$ then $K_B(x) \subseteq K_B(y)$ by Lemma 1, i.e., any element from $K_B(y) \cap X$ can be covered by some $K_B(x)$, where $x \in X$. Thus we may assume that $Y \subseteq X$.

Moreover, we may also assume that $Y = X$. Indeed, let us suppose that Y should be a proper subset of X and let $x \in X - Y$. Then $x \in K_B(y)$ for some $y \in Y$. However, $K_B(x) \subseteq K_B(y)$, by Lemma 1. Therefore, if we assume that $Y = X$, the set $G\overline{B}X$ will be not affected.

Computing both $G\underline{B}X$ and $G\overline{B}X$ using such formulas requires an algorithm with time computational complexity, in the worst case, of $O(n^2 \cdot m)$, where n is the cardinality of U and m is the cardinality of A.

Corollary. Let the only missing attribute values in a decision table be lost and let B be a subset of the attribute set A. For every concept $X \subseteq U$, all B-local upper approximations of X are unique, all B-global upper approximations of X are unique, and are equal to one another.

8 Conclusions

In this paper we introduced two new kinds of approximations: local and global. These approximations describe optimally approximated sets (lower approximations are largest, upper approximations are smallest 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 [12,13,14,19,24,25,26,27].

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

References

1. Greco, S., Matarazzo, B., 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, pp. 295–316. Kluwer Academic Publishers, Dordrecht (2000)
2. Grzymala-Busse, J.W.: On the unknown attribute values in learning from examples. In: Raś, Z.W., Zemankova, M. (eds.) ISMIS 1991. LNCS, vol. 542, pp. 368–377. Springer, Heidelberg (1991)
3. Grzymala-Busse, J.W.: Rough set strategies to data with missing attribute values. In: Workshop Notes, Foundations and New Directions of Data Mining, the 3-rd International Conference on Data Mining, Melbourne, FL, USA, November 19–22, pp. 56–63 (2003)

4. Grzymala-Busse, J.W.: Data with missing attribute values: Generalization of idiscernibility relation and rule induction. In: Peters, J.F., Skowron, A., Grzymała-Busse, J.W., Kostek, B.z., Świniarski, R.W., Szczuka, M.S. (eds.) Transactions on Rough Sets I. LNCS, vol. 3100, pp. 78–95. Springer, Heidelberg (2004)
5. Grzymala-Busse, J.W.: Characteristic relations for incomplete data: A generalization of the indiscernibility relation. In: Tsumoto, S., Słowiński, R., Komorowski, J., Grzymała-Busse, J.W. (eds.) RSCTC 2004. LNCS (LNAI), vol. 3066, pp. 244–253. Springer, Heidelberg (2004)
6. Grzymala-Busse, J.W.: Incomplete data and generalization of indiscernibility relation, definability, and approximations. In: Ślęzak, D., Wang, G., Szczuka, M., Düntsch, I., Yao, Y. (eds.) RSFDGrC 2005. LNCS (LNAI), vol. 3641, pp. 244–253. Springer, Heidelberg (2005)
7. Grzymala-Busse, J.W., Rzasa, W.: Local and global approximations for incomplete data. In: Greco, S., Hata, Y., Hirano, S., Inuiguchi, M., Miyamoto, S., Nguyen, H.S., Słowiński, R. (eds.) RSCTC 2006. LNCS (LNAI), vol. 4259, pp. 244–253. Springer, Heidelberg (2006)
8. Grzymala-Busse, J.W., Wang, A.Y.: Modified algorithms LEM1 and LEM2 for rule induction from data with missing attribute values. In: Proc. of the Fifth International Workshop on Rough Sets and Soft Computing (RSSC 1997) at the Third Joint Conference on Information Sciences (JCIS 1997), Research Triangle Park, NC, March 2–5, 1997, vol. 5, pp. 69–72 (1997)
9. Hong, T.P., Tseng, L.H., Chien, B.C.: Learning cevrage rules from incomplete data based on rough sets. In: Proc. of the IEEE International Conference on Systems, Man and Cybernetics, Hague, the Netherlands, October 10–13, 2004, pp. 3226–3231 (2004)
10. Kryszkiewicz, M.: Rough set approach to incomplete information systems. In: Proceedings of the Second Annual Joint Conference on Information Sciences, Wrightsville Beach, NC, September 28–October 1, 1995, pp. 194–197 (1995)
11. Kryszkiewicz, M.: Rules in incomplete information systems. Information Sciences 113, 271–292 (1999)
12. Lin, T.Y.: Neighborhood systems and approximation in database and knowledge base systems. In: Fourth International Symposium on Methodologies of Intelligent Systems (Poster Sessions), Charlotte, North Carolina, October 12–14, pp. 75–86 (1989)
13. Lin, T.Y.: Chinese Wall security policy—An aggressive model. In: Proceedings of the Fifth Aerospace Computer Security Application Conference, Tucson, Arizona, December 4–8, 1989, vol. 8, pp. 286–293 (1989)
14. Lin, T.Y.: Topological and fuzzy rough sets. In: Slowinski, R. (ed.) Intelligent Decision Support. Handbook of Applications and Advances of the Rough Sets Theory, pp. 287–304. Kluwer Academic Publishers, Dordrecht (1992)
15. Nakata, M., Sakai, H.: Rough sets handling missing values probabilistically interpreted. In: Ślęzak, D., Wang, G., Szczuka, M., Düntsch, I., Yao, Y. (eds.) RSFDGrC 2005. LNCS (LNAI), vol. 3641, pp. 325–334. Springer, Heidelberg (2005)
16. Pawlak, Z.: Rough Sets. International Journal of Computer and Information Sciences 11, 341–356 (1982)
17. Pawlak, Z.: Rough Sets. In: Theoretical Aspects of Reasoning about Data, Kluwer Academic Publishers, Dordrecht (1991)
18. Skowron, A., Rauszer, C.: The discernibility matrices and functions in information systems. In: Slowinski, R. (ed.) Handbook of Applications and Advances of the Rough Sets Theory, pp. 331–362. Kluwer Academic Publishers, Dordrecht (1992)

19. Slowinski, R., Vanderpooten, D.: A generalized definition of rough approximations based on similarity. IEEE Transactions on Knowledge and Data Engineering 12, 331–336 (2000)
20. Stefanowski, J.: Algorithms of Decision Rule Induction in Data Mining. Poznan University of Technology Press, Poznan (2001)
21. Stefanowski, J., Tsoukias, A.: On the extension of rough sets under incomplete information. In: Zhong, N., Skowron, A., Ohsuga, S. (eds.) RSFDGrC 1999. LNCS (LNAI), vol. 1711, pp. 73–81. Springer, Heidelberg (1999)
22. Stefanowski, J., Tsoukias, A.: Incomplete information tables and rough classification. Computational Intelligence 17, 545–566 (2001)
23. Wang, G.: Extension of rough set under incomplete information systems. In: Proc. of the IEEE International Conference on Fuzzy Systems (FUZZ_IEEE 2002), Honolulu, HI, May 12–17, 2002, vol. 2, pp. 1098–1103 (2002)
24. Yao, Y.Y.: Two views of the theory of rough sets in finite universes. International J. of Approximate Reasoning 15, 291–317 (1996)
25. Yao, Y.Y.: Relational interpretations of neighborhood operators and rough set approximation operators. Information Sciences 111, 239–259 (1998)
26. Yao, Y.Y.: On the generalizing rough set theory. In: Wang, G., Liu, Q., Yao, Y., Skowron, A. (eds.) RSFDGrC 2003. LNCS (LNAI), vol. 2639, pp. 44–51. Springer, Heidelberg (2003)
27. Yao, Y.Y., Lin, T.Y.: Generalization of rough sets using modal logics. Intelligent Automation and Soft Computing 2, 103–119 (1996)