# Predicting incomplete data on the basis of non symmetric similarity relation

Ewa Adamus and Andrzej Piegat

Faculty of Computer Science and Information Technology, Szczecin University of Technology, Żołnierska 49, 71-210 Szczecin, Poland, eadamus@wi.ps.pl, apiegat@wi.ps.pl

**Abstract.** The rough sets theory was meant as a tool for imprecise and inconsistent information systems. Incomplete information can be also considered as a particular case of imprecise information. Because the rough sets theory makes the assumption of completeness of all attributes of input vector, many modifications of this theory were developed describing how to use the incomplete data. This article presents basic approaches: tolerance relation and non symmetric similarity relation. Furthermore, a new method of supplementing some incomplete objects from an information table has been proposed.

## Introduction

Rough sets theory is based on assumption that every considered object is connected with some kind of information. The objects with the same description (the same information) are indiscernible in the aspect of accessible information. The indiscernibility relation, defined in this way, is the mathematic base of rough sets theory.

The set of the indiscernible objects is called the *elementary class*. The set that is a union of certain elementary classes is a reference to precise set or an inexact one – a rough set. Each rough set has a boundary area, an area that contains objects that don't entirely belong to the class or to its complement. The objects which certainly belong to the class are its *lower approximation* and these which belong to the class only probably are its *upper approximation*.

Because the original definition of indiscernibility relation makes the assumption of completeness of input vectors of compared objects $x$ and $y$ in the input vector, the problem appears when at least one of the attributes doesn't have its defined value. This is the cause of creating modifications of indiscernibility relation. Some of the basics are described below. Additionally, a proposition of filling in certain incomplete objects based on non-symmetric similarity relation will be presented.

## Incomplete data and rough sets theory

Generally in the case when missing data occur, we can use either symmetric tolerance relation or non symmetric similarity relation. Basing on these relations we are going to make a comparative analysis.

The reason of introducing the tolerance relation to the ,,incomplete" input vector is treating incomplete objects inconsistently in the conventional indiscernibility relation. The indiscernibility relation treats an object which doesn't have a value for attribute $a \in A$ as an equivalent, although the real values for this attribute can be different [3]. In the case, when one of the objects has a value for the mentioned attribute *a*, the objects are treated as different, although the real value for attribute *a* may be identical with another object's value. We can acquire a consistent treatment of incomplete data by using the tolerance relation defined in [5, 6].

An information system is a pair $S = (U, A)$, where *U* - is a non-empty set of objects called the universe, and *A* – is a set of attributes. For attributes subset $B \subseteq A$, the tolerance relation is defined as follows:

$$\mathrm{TOL(B)} = \{(\mathrm{x}, \mathrm{y}) \in \mathrm{U} \times \mathrm{U} : \forall\, \mathrm{a} \in \mathrm{B}, \mathrm{f(a,x)} = \mathrm{f(a,y)} \text{ or } \mathrm{f(a,x)} = * \text{ or } \mathrm{f(a,y)} = *\}, \quad (1)$$

the relation is reflexive, symmetric but not necessarily transitive.

$T_B(x)$ is a set of objects that are indiscernible with *x* in regard to *B*:

$$\mathrm{T_B(x)} = \{\mathrm{y} \in \mathrm{U} : (\mathrm{x}, \mathrm{y}) \in \mathrm{TOL(B)}\}. \quad (2)$$

The similarity relation is an alternative for the indiscernibility relation for imprecise data or for precise data, which differ insignificantly for the whole analysis ([8]).

Basing on definitions in [8, 2] Stefanowski gives a definition of similarity relation that regards incomplete data with definitions of sets approximations [9, 10].

We say that object *y* is similar to *x* ( $yS_bx$ ), when:

$$\forall \mathrm{a} \in \mathrm{B} \text{ such that } \mathrm{f(a,y)} \neq *, \mathrm{f(a,x)} = \mathrm{f(a,y)}. \quad (3)$$

The relation isn't similar and reflexive. Additionally, two classes of similarity relations were defined:

- a set of objects similar to $x$:

$$\mathrm{S_B(x)} = \{\mathrm{y} \in \mathrm{U} : \mathrm{yS_Bx}\}, \quad (4)$$

- a set of objects, where *x* is similar to:

$$\mathrm{S_B^{-1}(x)} = \{\mathrm{y} \in \mathrm{U} : \mathrm{xS_By}\}. \quad (5)$$

Basing on classes of similarity definitions, the lower and upper approximation for $\mathrm{X} \subseteq \mathrm{U}$ can be defined as:

$$\underline{B}_S(X) = \{x \in U : S_B^{-1}(x) \subseteq X\}, \quad (6)$$

$$\overline{B}_S(X) = \bigcup \{S_B(x) : x \subseteq X\}. \quad (7)$$

## Predicting incomplete data using non symmetric similarity relation

The literature distinguishes the following methods of dealing with the problem of incomplete data:

1. Data with missing values are not taken into consideration.
2. Estimating of the missing data – usually during the data preprocessing.
3. The methods are chosen accordingly to the possibilities of the missing value in some attributes of the sample's input vector. Only the defined values of the incomplete sample's input vector are taken into consideration in that case.

If the problem is applied to the rough set theory an explicit classification of this theory to one of the mentioned methods cannot be done. In this aspect the theory is a most distinguishable one if compared to all the other methods (see [1]). The incomplete sample, dependently on the context (of reciprocal relations with other objects in the base of knowledge) will certainly be accepted for the further analysis (it will be the lower approximation of the given deciding concept), or it will not be accepted as it will be found in the boundary area. Thus, when applying the rough set theory to solve the problem of incomplete data it is possible to divide incomplete data into two groups of objects: *certain* and *doubtful*. The first group deals with the data which contain additional information – they are either incomplete unique objects or objects which are similar to their own decision class only (in this case the incomplete sample may be treated as a simplified decision rule). In the case of the doubtful objects those elements will be taken into consideration with some probability in the further analysis. The aim of this work is to establish yet another scenario of dealing with incomplete data, i.e. supplementing the missing data.

On the basis of the interpretation of the approximations of set $X \subseteq U$ (def. 6., 8.) for an incomplete object $x \subseteq U$ one of the following scenarios can be adopted:
For the decision table $DT = (U, A \cup \{d\})$ where $d \notin A$ is the decision attribute and the subset of input attributes $B \subseteq A$ and object $x \subseteq U$ there is a defined class of objects and to which $x$ is similar $S_B^{-1}(x)$ (def. 5.) in order to simplify the recording of the further formulas, index $B$ will be omitted assuming that the whole analysis is done for the subset of attributes $B \subseteq A$. Then:

1. If $S^{-1}(x) \subseteq X$ then the incomplete object $x$ makes the lower approximation of set $X$. It is an equivalent of the third method for the dealing with incomplete data i.e. **the acceptance of only the defined information in the incomplete sample**.
2. If $S^{-1}(x) \cap X \neq \emptyset$ and $S^{-1}(x) \cap -X \neq \emptyset$ then **the incomplete object** $x$ **will not be a lower approximation of set** $X$ i.e. it will be found in the boundary area of the decision classes. It is impossible to classify it implicitly. The further part of the article will be devoted to this aspect.

Assuming that $S_r^{-1}(x) = S^{-1}(x) \setminus \{x\}$ i.e. we are interested in the class of objects to which $x$ is similar without the very object (i.e. $x$ is not similar to itself). Let us analyse the situation when set $S_r^{-1}(x)$ represents an opposite decision class, that is

$S_r^{-1}(x) \subseteq -X$ . If we possess information which is opposite to x , we know what values the incomplete object should not take, then we know the area of permitted values for x , which will comprise of the complementation of set $S_r^{-1}(x)$ . Following case 2. for the incomplete object $x \subseteq U$ for which condition: $S^{-1}(x) \cap X \neq \varnothing$ and $S^{-1}(x) \cap -X \neq \varnothing$ is fulfilled we can take one of the following scenarios:

1. If $S_r^{-1}(x) \subseteq -X$ , **then incomplete object** x **is filled with the complement of the similarity class** (see 10.)
2. Otherwise, if $S_r^{-1}(x) \cap X \neq \varnothing$ and $S_r^{-1}(x) \cap -X \neq \varnothing$ then in order to check whether we possess coherent opposite information we make an analysis of the so-called **directional classes of similarity**.

Similarity relations „can be interpreted as representatives of inclusion relations as the similarity of y to x is equivalent to the notion that the description of object y is comprised in the description of object x [10]." However the relation of similarity is a relation of a partial order of set $S^{-1}(x)$ as not each two elements of this set are comparable. For instance the fig. 1. objects 8. and 9. are not comparable as they possess different values for the third attribute. In such a case we refer to set $S_r^{-1}(x)$ as **partially ordered**. In relation to this set $S_r^{-1}(x)$ can be presented as a family of disjointed pairs of sets and completely ordered by the inclusion relations (see fig. 1.). The authors of the present article have termed those sets as **directional classes of similarity** and marked with the symbol $S_{rK}^{-1}(y)$ where y fulfills the condition: $y \in S_r^{-1}(x) \wedge S^{-1}(y) = \{y\}$ that is belonging to a set of objects between which it is impossible to define the mutual inclusion relation of the objects' description.

$$S_{rK}^{-1}(y) = S_r^{-1}(x) \cap S(y), \tag{8}$$

Eventually set $S_r^{-1}(x)$ can be presented as a sum of the family of directional similarity sets of disjoint pairs:

$$S_r^{-1}(x) = \bigcup S_{rK}^{-1}(y). \tag{9}$$

Within the single directional set of similarity ( $S_{rK}^{-1}(y)$ ) we can define mutual inclusion relations between each two elements of this set. The direction of these relations is determined by the level of incompleteness of the objects.

The exemplary division of set $S^{-1}(x)$ into directional sets of similarity has been presented in a figure (fig. 1.) as set is a partially ordered set, a convenient form of graphic presentation of this type of a set is Hasse's diagram.

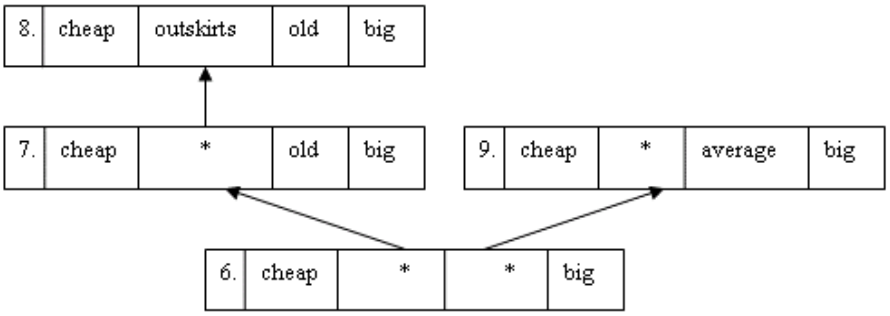


**Fig. 1.** The exemplary division of set $S^{-1}(x)$ into directional sets of similarity.

The particular branches of the diagram are of course directional sets of similarity, that is completely ordered subsets of a partially ordered set.

The analysis of directional sets of similarity allows to present the case $(S_r^{-1}(x) \cap X \neq \varnothing) \wedge (S_r^{-1}(x) \cap -X \neq \varnothing)$ as two instances:

1. If $\exists(S_{rK}^{-1}(y) \subseteq S^{-1}(x)) \wedge (S_{rK}^{-1}(y) \cap X) \wedge (S_{rK}^{-1}(y) \cap -X)$ - within a single directional set of similarity there is no unity as to the belonging to the decision concept. According to the classic definitions **object x will not be the lower approximation both for set X and -X**.
2. If $\forall(S_{rK}^{-1}(y) \subseteq S^{-1}(x)) \wedge ((S_{rK}^{-1}(y) \subseteq X) \vee (S_{rK}^{-1}(y) \subseteq -X)) \wedge (\left|S^{-1}(y)\right| = 1)$, where $\left|S^{-1}(y)\right|$ is the cardinality of the set $S^{-1}(y)$ - in all the directional sets of similarity there is unity as to the belonging to the decision concept and there is at least one set $S_{rK}^{-1}(y)$ representing an opposite decision concept: $S_{rK}^{-1}(y) \subseteq -X$ - in that case the incomplete object is **filled with a complement of a similarity class** (see 10.).

To sum up the above considerations an algorithm of dealing with an incomplete object x will be presented, with the assumption that for the decision table $DT = (\mathrm{U}, \mathrm{A} \cup \{\mathrm{d}\})$ and object $\mathrm{x} \subseteq \mathrm{U}$ there is a defined class of objects, to which x is similar: $S^{-1}(x)$ (def. 5.). Additionally $S_r^{-1}(x) = S^{-1}(x) \setminus \{x\}$ and $S_{rK}^{-1}(y)$ is a directional set of similarity for $y \in S_r^{-1}(x)$. Then:

**1.** If $S^{-1}(x) \subseteq X$ **- the acceptance of only the defined information in the incomplete sample.**

**2.** If $S^{-1}(x) \cap X \neq \varnothing$ and $S^{-1}(x) \cap -X \neq \varnothing$ then:

- If $S_r^{-1}(x) \subseteq -X$, then **the incomplete object x is filled with the complement of the similarity class** (see 10.)

– Otherwise, if $S_r^{-1}(x) \cap X \neq \varnothing$ and $S_r^{-1}(x) \cap -X \neq \varnothing$ then in order to check whether we possess coherent opposite information we make an analysis of the **directional classes of similarity:**
  - If $\exists(S_{rK}^{-1}(y) \subseteq S^{-1}(x)) \wedge (S_{rK}^{-1}(y) \cap X) \wedge (S_{rK}^{-1}(y) \cap -X)$ - **object x will not be the lower approximation both for set X and -X.**
  - If $\forall(S_{rK}^{-1}(y) \subseteq S^{-1}(x)) \wedge ((S_{rK}^{-1}(y) \subseteq X) \vee (S_{rK}^{-1}(y) \subseteq -X)) \wedge (\left|S^{-1}(y)\right| = 1)$ - **the incomplete object x is filled with a complement of a similarity class** (see 10.).

The supplementation of the incomplete object x, which qualifies to be supplemented, consists of complement of the class of objects to which the incomplete one is similar: $\boldsymbol{C(S^{-1}(x))}$ except for such objects which belong to tolerance relation class $T(x)$ on the basis of which we can make a explicit classification: $S^{-1}(y) \subseteq X \vee S^{-1}(y) \subseteq -X$ :

$$C(S^{-1}(x)) \setminus \{y : y \in T(x) \wedge (S^{-1}(y) \subseteq X \vee S^{-1}(y) \subseteq -X)\}. \qquad \textbf{(10)}$$

## Experiments

The presented method has been verified on the basis of a set with real measurement data. The experiments were made with the use of the diabetes set (for diagnosing diabetes of Pima indians) from the popular benchmark dataset [7]. Each sample has 8 inputs and 1 output which takes the value 0 or 1. The whole dataset includes 768 complete samples. The data set was modified by introducing various degrees of incompleteness for the following basic kinds of incompleteness [4]:

- MCAR (*missing completely at random*);
- MAR (*missing at random*);
- NI *(non ignorable*).

Fig. 2. presents the dependence of the incompleteness degree of the data on the average number of rules making a proper classification ( *the number of proper rules/ the number of all the rules*). In order to generate the set of rules, the LEM2 algorithm [3] was used. The similarity relation makes the basis for the method of conditioned supplementing; when there are no samples qualifying to be completed the method comes down to the non-symmetric similarity relation. That is why fig. 2. presents the results of testing for the decision rules generated on the basis of the similarity relation with and without the conditional supplementation. The verification of the rules was made with the use of the *k-fold validation* technique. Fig.2. presents the results of the calculations for the testing stage.

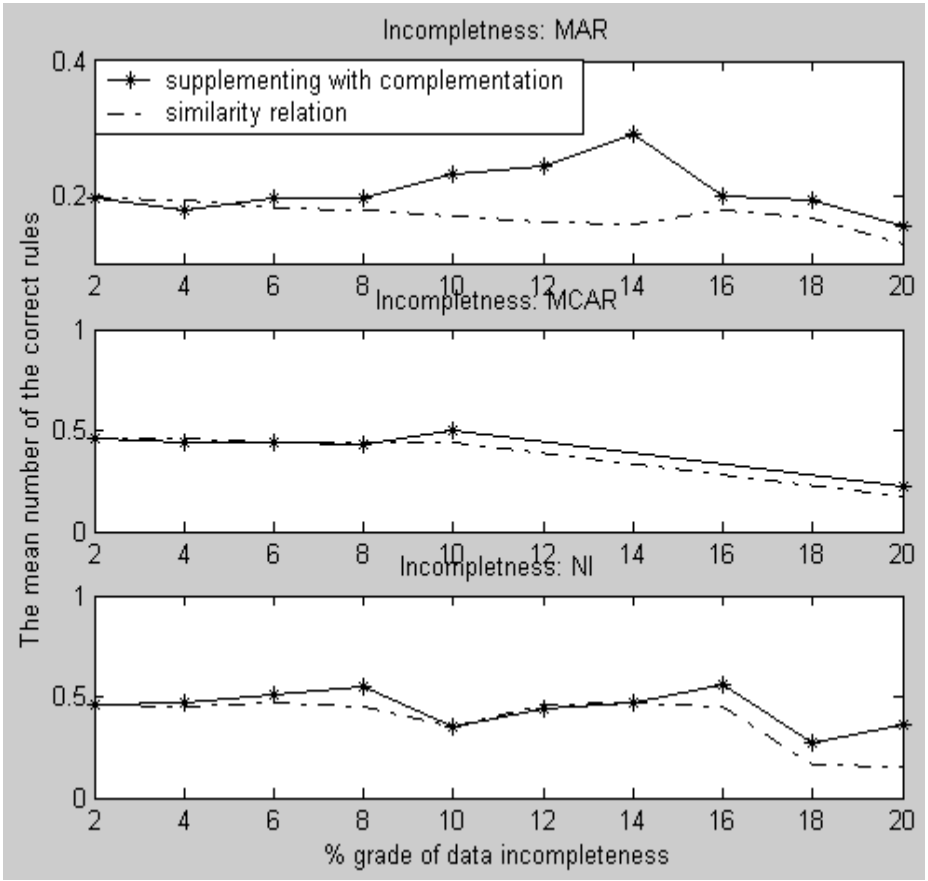


**Fig. 2.** The relationship between the level of incompleteness of the mean number of the correct rules for testing samples.

The presented results imply that the application of the conditioned supplementation for the induction of rules for incomplete data improves the quality of modelling. Completing of certain samples on the basis of the information provided by the non-symmetric similarity relation can bring some beneficial effects for the application of the similarity relation itself.

## Conclusions

The authors of the article have attempted to prove that supplementing incomplete data (especially in the cases when this prediction brings into the analysis additional information) has some benefits.

1. Using this method, a consistent decision table isn't converted into inconsistent one. However, in the case of the RS theory, we have good tool to deal with inconsistencies, but the supplemented new objects may bring inconsistencies and be indiscernible with the existing objects.
2. In consequence of this way of supplementing, we don't lose significant information included in the unique, incomplete objects.
3. Supplementing on the basis of the non-symmetric similarity relation increases the accuracy approximation of a certain class. This is because the set of the newly supplemented objects is entirely included in the lower approximation of the current set.

## References


1. E. Adamus Przegląd metod stosowanych do badań nad niekompletnymi danymi pomiarowymi. VIII *Sesja Informatyki*, II:387-395, Szczecin, 2003.
2. Greco S., Matarazzo B., Słowiński R. Dealing with missing data In rough sets analysis of multi-attribute and multi-criteria decision problems. *Kluwer Academic Publisher*, p. 295-316, 2002.
3. J. Grzymala-Busse and A. Y. Wang. Modified algorithms lem1 and lem2 for rule induction from with missing attribute values. Proc.of theFifth International Workshop on Rough Sets and Soft Computing (RSSC'97) at the Third Join Conference on Information Sciences (JCIS'97), Research Triangle Park, NC, pages 69-72, March 2-5 1997.
4. G. King, J. Honaker and A. Joseph Analyzing Incomplete Political Science Data: An Alternative Algorithm for multiple imputation. *American Political Science Review*, (95) p. 49-69, 2001.
5. M. Kryszkiewicz. Rough set approach to rules generation from incomplete information systems. *ICS Research Report 55/95, Warsaw University of Technology; also in: International Journal of Information Sciences, 1995.*
6. M. Kryszkiewicz Rough set approach to incomplete information system. *International Journal of Information Sciences*, (112): p. 39-49, 1998.
7. Prechelt L.: Proben1 – A set of neural network benchmark problems and bench-marking rules, *Technical Report*, 1994
8. R. Słowiński and D. Vanderpooten. A generalized definition of rough Approximations based on similarity. *IEEE Transactions on Knowledge and Data Engineering*, 12: 331-336, march/april 2000.
9. J. Stefanowski and A. Tsoukias. On the extension of rough sets under incomplete information. *New Directions in Rough Sets, Data Mining and Granular-Soft Computing, LNAI 1711, Springer-Verlag, Berlin,* 1999.
10. J. Stefanowski and A. Tsoukias. Incomplete information tables and rough classification. *Int. Journal of Computational Intelligence*. 2001.