Complexity of Rule Sets Induced from Data Sets with Many Lost and Attribute-Concept Values

Patrick G. Clark[1], Cheng Gao[1], and Jerzy W. Grzymala-Busse[1,2](✉)

[1] Department of Electrical Engineering and Computer Science,
University of Kansas, Lawrence, KS 66045, USA
patrick.g.clark@gmail.com, {cheng.gao,jerzy}@ku.edu
[2] Department of Expert Systems and Artificial Intelligence,
University of Information Technology and Management, 35-225 Rzeszow, Poland

Abstract. In this paper we present experimental results on rule sets induced from 12 data sets with many missing attribute values. We use two interpretations of missing attribute values: lost values and attribute-concept values. Our main objective is to check which interpretation of missing attribute values is better from the view point of complexity of rule sets induced from the data sets with many missing attribute values. The better interpretation is the attribute-value. Our secondary objective is to test which of the three probabilistic approximations used for the experiments provide the simplest rule sets: singleton, subset or concept. The subset probabilistic approximation is the best, with 5 % significance level.

Keywords: Incomplete data · Lost values · Attribute-concept values · Probabilistic approximations · MLEM2 rule induction algorithm

1 Introduction

The basic ideas of rough set theory are standard lower and upper approximations. A probabilistic approximation with a probability α is an extension of the standard approximation. For $\alpha = 1$, the probabilistic approximation is reduced to the lower approximation; for very small α, it is reduced to the upper approximation. Research on theoretical properties of probabilistic approximations was initiated in [1] and then was continued in, e.g., [2–5].

Incomplete data sets are analyzed using special approximations such as singleton, subset and concept [6,7]. Probabilistic approximations, for incomplete data sets and based on an arbitrary binary relation, were introduced in [8]. The first experimental results using probabilistic approximations were published in [9]. In experiments reported in this paper, we used three kinds of probabilistic approximations: singleton, subset and concept.

In this paper, we consider two interpretations of missing attribute values, lost values and attribute-concept values. Lost values indicate that the original values were erased, and as a result we should use only existing, specified attribute values for data mining. Attribute-concept values may be replaced by any specified attribute value for a given concept.

L. Rutkowski et al. (Eds.): ICAISC 2016, Part II, LNAI 9693, pp. 27–36, 2016.
DOI: 10.1007/978-3-319-39384-1_3

Experimental research on comparing different approaches to mining incomplete data was initiated in [10], where results of experiments on data sets with 35 % missing attribute values, using two interpretations of missing attribute values: lost values and “do not care” conditions, were presented.

Research on mining incomplete data with lost values and attribute-concept values, using different experimental setups, was presented in [11–14]. Results of initial research [10,12] show that the quality of rule sets, evaluated by an error rate computed by ten-fold cross validated, does not differ significantly with different combinations of missing attribute and probabilistic approximation type. On the other hand, for data sets with many lost values and attribute-concept values, experiments described in [13] show that the error rate was smaller for lost values.

In [11,14], complexity of rule sets induced from data with lost values and attribute-concept values was investigated. The results were not quite decisive, though the number of rules was always smaller for data sets with attribute-concept values, the results for the total number of rule conditions were not so conclusive.

Therefore the main objective of this paper is research on complexity of rule sets, in terms of the number of rules and total number of rule conditions, induced from data sets with many lost values and attribute-concept values using the Modified Learning from Examples Module version 2 (MLEM2) system for rule induction. The results of this paper show that the number of rules and the total number of conditions are always smaller for attribute-concept values than for lost values.

In our previous research [11,14], results on the best choice of probabilistic approximations (singleton, subset or concept) were not conclusive. So our secondary objective is to check which probabilistic approximation (singleton, subset or concept) is the best from the point of view of rule complexity. As results of our paper show, the best choice is the subset probabilistic approximation.

This paper starts with a discussion on incomplete data in Sect. 2 where we define approximations, attribute-value blocks and characteristic sets. In Sect. 3, we present singleton, subset and concept probabilistic approximations for incomplete data. Section 4 contains the details of our experiments. Finally, conclusions are presented in Sect. 5.

2 Incomplete Data

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, $U = \{1, 2, 3, 4, 5, 6, 7\}$. 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, $A = \{$ *Wind*, *Temperature*, *Humidity*$\}$. The value for a case x and an attribute a will be denoted by $a(x)$.

In this paper, we distinguish between two interpretations of missing attribute values: lost values, denoted by "?" and attribute-concept values, denoted by "–" [15,16]. Table 1 presents an incomplete data set affected by both lost values and attribute-concept values.

Table 1. A decision table

Case	Attributes			Decision
	Wind	Temperature	Humidity	Trip
1	low	?	low	yes
2	?	high	–	yes
3	high	–	low	yes
4	–	high	?	yes
5	high	low	–	no
6	low	high	?	no
7	?	?	high	no

One of the most important ideas of rough set theory [17] is an indiscernibility relation, defined for complete data sets. Let B be a nonempty subset of A. The indiscernibility relation $R(B)$ is a relation on U defined for $x, y \in U$ as defined by

$$(x, y) \in R(B) \text{ if and only if } \forall a \in B \ (a(x) = a(y))$$

The indiscernibility relation $R(B)$ is an equivalence relation. Equivalence classes of $R(B)$ are called *elementary sets* of B and are denoted by $[x]_B$. A subset of U is called *B-definable* if it is a union of elementary sets of B.

The set X of all cases defined by the same value of the decision d is called a *concept*. For example, a concept associated with the value *yes* of the decision *Trip* is the set $\{1, 2, 3, 4\}$. The largest B-definable set contained in X is called the *B-lower approximation* of X, denoted by $\underline{appr}_B(X)$, and defined as follows

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

The smallest B-definable set containing X, denoted by $\overline{appr}_B(X)$ is called the *B-upper approximation* of X, and is defined by

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

For a variable a and its value v, (a, v) is called a variable-value pair. A *block* of (a, v), denoted by $[(a, v)]$, is the set $\{x \in U \mid a(x) = v\}$ [18]. For incomplete decision tables the definition of a block of an attribute-value pair is modified in the following way.

- If for an attribute a there exists a case x such that $a(x) = ?$, 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 an attribute-concept value, i.e., $a(x) = -$, then the corresponding case x should be included in blocks $[(a, v)]$ for all specified values $v \in V(x, a)$ of attribute a, where $V(x, a)$ is defined by

$$\{a(y) \mid a(y) \textit{ is specified},\ y \in U,\ d(y) = d(x)\}$$

For the data set from Table 1, we have $V(2, Humidity) = \{low\}$, $V(3,$ *Temperature*$) = \{high\}$, $V(4, Wind) = \{low, high\}$ and $V(5, Humidity) = \{high\}$.

For the data set from Table 1 the blocks of attribute-value pairs are:

[(Wind, low)] = {1, 4, 6},
[(Wind, high)] = {3, 4, 5},
[(Temperature, low)] = {5}, and
[(Temperature, high)] = {2, 3, 4, 6},
[(Humidity, low)] = {1, 2, 3},
[(Humidity, high)] = {5, 7}.

For a case $x \in U$ and $B \subseteq A$, 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 $a(x)$ is specified, then $K(x, a)$ is the block $[(a, a(x))]$ of attribute a and its value $a(x)$,
- If $a(x) =?$ then the set $K(x, a) = U$, where U is the set of all cases,
- If $a(x) = -$, then the corresponding set $K(x, a)$ is equal to the union of all blocks of attribute-value pairs (a, v), where $v \in V(x, a)$ if $V(x, a)$ is nonempty. If $V(x, a)$ is empty, $K(x, a) = U$.

For Table 1 and $B = A$,

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

First we will quote some definitions from [19]. Let X be a subset of U. The B-*singleton lower approximation* of X, denoted by $\underline{appr}_B^{singleton}(X)$, is defined by

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

The B-*singleton upper approximation* of X, denoted by $\overline{appr}_B^{singleton}(X)$, is defined by

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

The B-*subset lower approximation* of X, denoted by $\underline{appr}_B^{subset}(X)$, is defined by

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

The B-*subset upper approximation* of X, denoted by $\overline{appr}_B^{subset}(X)$, is defined by

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

The B-*concept lower approximation* of X, denoted by $\underline{appr}_B^{concept}(X)$, is defined by

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

The B-*concept upper approximation* of X, denoted by $\overline{appr}_B^{concept}(X)$, is defined by

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

For Table 1 and $X = \{5, 6, 7\}$, all A-singleton, A-subset and A-concept lower and upper approximations are:

$$\underline{appr}_A^{singleton}(X) = \{5, 7\},$$
$$\overline{appr}_A^{singleton}(X) = \{4, 5, 6, 7\},$$
$$\underline{appr}_A^{subset}(X) = \{5, 7\},$$
$$\overline{appr}_A^{subset}(X) = \{3, 4, 5, 6, 7\},$$
$$\underline{appr}_A^{concept}(X) = \{5, 7\},$$
$$\overline{appr}_A^{concept}(X) = \{4, 5, 6, 7\}.$$

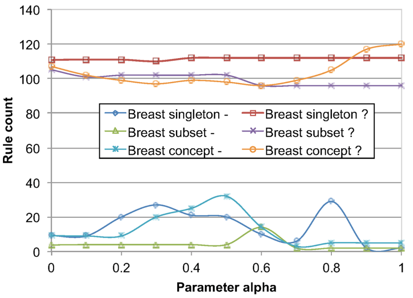

Fig. 1. Number of rules for the *breast cancer* data set

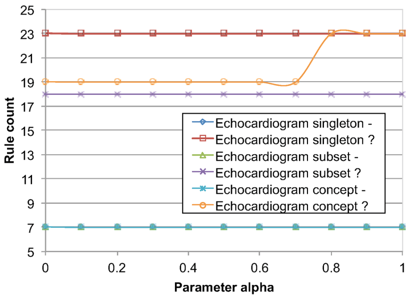

Fig. 2. Number of rules for the *echocardiogram* data set

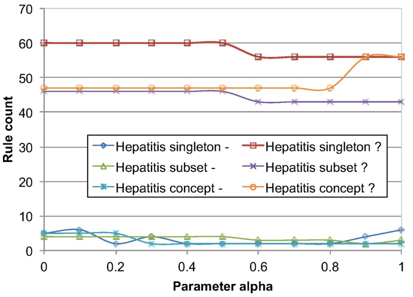

Fig. 3. Number of rules for the *hepatitis* data set

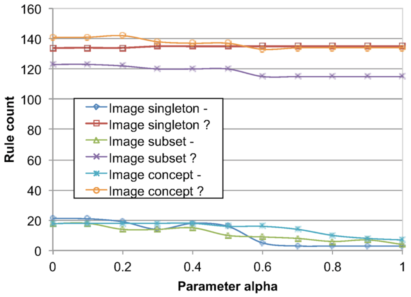

Fig. 4. Number of rules for the *image segmentation* data set

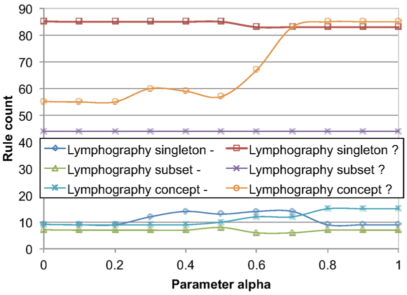

Fig. 5. Number of rules for the *lymphography* data set

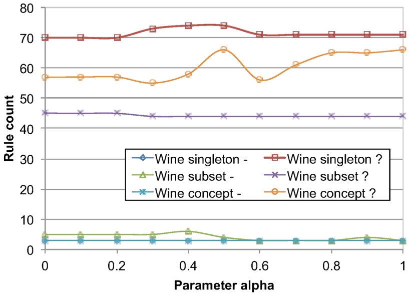

Fig. 6. Number of rules for the *wine recognition* data set

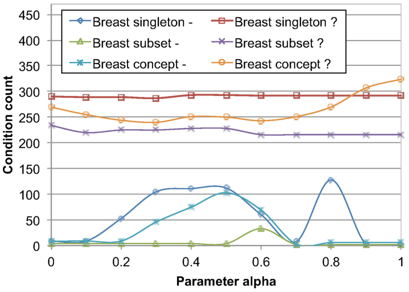

Fig. 7. Total number of conditions for the *breast cancer* data set

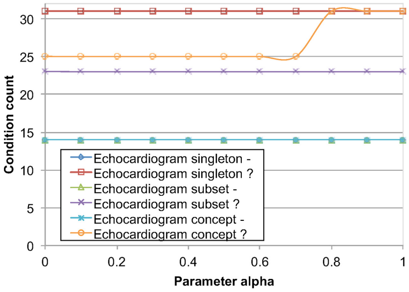

Fig. 8. Total number of conditions for the *echocardiogram* data set

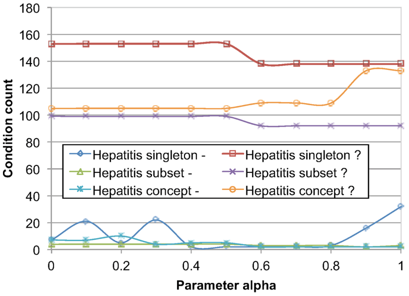

Fig. 9. Total number of conditions for the *hepatitis* data set

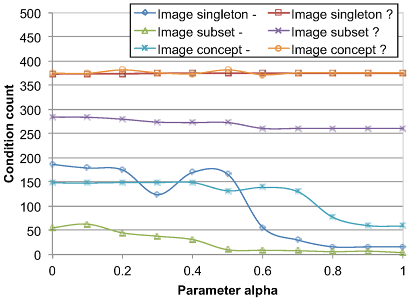

Fig. 10. Total number of conditions for the *image* data set

3 Probabilistic Approximations

In this section definitions of singleton, subset and concept approximations are extended to the corresponding probabilistic approximations. A B-singleton probabilistic approximation of X with the threshold α, $0 < \alpha \leq 1$, denoted by $appr_{\alpha,B}^{singleton}(X)$, is defined by

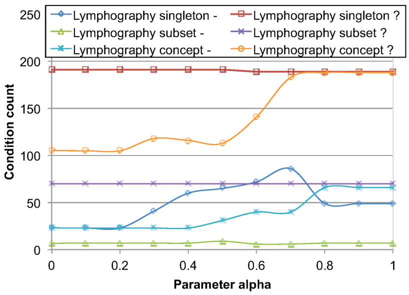

Fig. 11. Total number of conditions for the *lymphography* data set

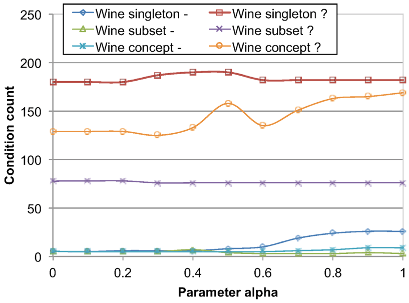

Fig. 12. Total number of conditions for the *wine recognition* data set

$$\{x \mid x \in U,\ Pr(X \mid K_B(x)) \geq \alpha\},$$

where $Pr(X \mid K_B(x)) = \frac{|X \cap K_B(x)|}{|K_B(x)|}$ is the conditional probability of X given $K_B(x)$ and $|Y|$ denotes the cardinality of set Y. A B-subset probabilistic approximation of the set X with the threshold α, $0 < \alpha \leq 1$, denoted by $appr^{subset}_{\alpha,B}(X)$, is defined by

$$\cup\{K_B(x) \mid x \in U,\ Pr(X \mid K_B(x)) \geq \alpha\}.$$

A B-concept probabilistic approximation of the set X with the threshold α, $0 < \alpha \leq 1$, denoted by $appr^{concept}_{\alpha,B}(X)$, is defined by

$$\cup\{K_B(x) \mid x \in X,\ Pr(X \mid K_B(x)) \geq \alpha\}.$$

Note that if $\alpha = 1$, the probabilistic approximation is the standard lower approximation and if α is small, close to 0, in our experiments it is 0.001, the same definition describes the standard upper approximation.

For Table 1 and the concept $X = [(Trip, no)] = \{4, 5, 6\}$, there exist the following distinct probabilistic approximations:

$appr^{singleton}_{1.0,A}(X)= \{5, 7\}$,
$appr^{singleton}_{0.5,A}(X)= \{5, 6, 7\}$,
$appr^{singleton}_{0.333,A}(X)= \{4, 5, 6, 7\}$,
$appr^{subset}_{1.0,A}(X)= \{5, 7\}$,
$appr^{subset}_{0.5,A}(X)= \{4, 5, 6, 7\}$,
$appr^{subset}_{0.333,A}(X)= \{3, 4, 5, 6, 7\}$,
$appr^{concept}_{1.0,A}(X)= \{5, 7\}$,
$appr^{concept}_{0.5,A}(X)= \{4, 5, 6, 7\}$.

4 Experiments

Our experiments are based on six data sets that are available on the University of California at Irvine *Machine Learning Repository*. Basic information about these data sets is presented in Table 2.

Table 2. Data sets used for experiments

Data set	Number of			Percentage of missing attribute values
	Cases	Attributes	Concepts	
Breast cancer	277	9	2	44.81
Echocardiogram	74	7	2	40.15
Hepatitis	155	19	2	60.27
Image segmentation	210	19	7	69.85
Lymphography	148	18	4	69.89
Wine recognition	178	13	3	64.65

For every data set a set of templates was created. Templates were formed by replacing incrementally (with 5 % increment) existing specified attribute values by *lost* values. Thus, we started each series of experiments with no *lost* values, then we added 5 % of *lost* values, then we added additional 5 % of *lost* values, etc., until at least one entire row of the data sets was full of *lost* values. Then three attempts were made to change configuration of new *lost* values and either a new data set with extra 5 % of *lost* values were created or the process was terminated. Additionally, the same templates were edited for further experiments by replacing question marks, representing *lost* values by "–"s, representing *attribute-concept* values.

For any data set there was some maximum for the percentage of missing attribute values. For example, for the *Breast cancer* data set, it was 44.81 %. In our experiments we used only such incomplete data sets, with as many missing attribute values as possible. Note that for some data sets the maximum of the number of missing attribute values was less than 40 %, we have not used such data for our experiments.

For rule induction we used the Modified Learning from Examples Module version 2 (MLEM2) rule induction algorithm, a component of the Learning from Examples based on Rough Sets (LERS) data mining system [18]. Results of our experiments are presented in Figs. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 and 12.

Our main objective was to select the better interpretation of missing attribute values: lost values or attribute-concept values in terms of complexity measured by the number of rules and the total number of conditions in rule sets. For any data set we compared the size of the rule set and the total number of conditions in the rule set for two interpretations of missing attribute values with the same type of probabilistic approximation. Our results show that the number of rules was always smaller for attribute-concept values than for lost values. Similarly, the total number of conditions was always smaller for attribute-concept values than for lost values.

Our secondary objective was to find the best kind of probabilistic approximations (singleton, subset or concept). Here the answer is more complicated.

For any data set we compared all three kinds of probabilistic approximations assuming the same type of missing attribute values using multiple comparisons based on Friedman's nonparametric test. As a result, the smallest number of rules is accomplished by subset approximations for eight out of 12 data sets (5 % significance level). For four data sets (*echocardiogram*, *hepatitis*, *image segmentation* and *wine recognition*, all with attribute-concept values), the difference is not statistically significant. The total number of conditions is also the smallest for subset approximations except two data sets (*echocardiogram* and *hepatitis*, both with attribute-concept values).

5 Conclusions

As follows from our experiments, the number of rules and the total number of conditions is always smaller for attribute-concept values than for lost values. Additionally, the best probabilistic approximation that should be used for rule induction from data with many missing attribute values is subset probabilistic approximation.

References

1. Pawlak, Z., Wong, S.K.M., Ziarko, W.: Rough sets: probabilistic versus deterministic approach. Int. J. Man Mach. Stud. **29**, 81–95 (1988)
2. Pawlak, Z., Skowron, A.: Rough sets: Some extensions. Inf. Sci. **177**, 28–40 (2007)
3. Yao, Y.Y.: Probabilistic rough set approximations. Int. J. Approximate Reasoning **49**, 255–271 (2008)
4. Yao, Y.Y., Wong, S.K.M.: A decision theoretic framework for approximate concepts. Int. J. Man Mach. Stud. **37**, 793–809 (1992)
5. Ziarko, W.: Probabilistic approach to rough sets. Int. J. Approximate Reasoning **49**, 272–284 (2008)
6. Grzymala-Busse, J.W.: Rough set strategies to data with missing attribute values. In: Notes of the Workshop on Foundations and New Directions of Data Mining, in conjunction with the Third International Conference on Data Mining, pp. 56–63 (2003)
7. Grzymala-Busse, J.W.: Data with missing attribute values: generalization of indiscernibility relation and rule induction. Trans. Rough Sets **1**, 78–95 (2004)
8. Grzymała-Busse, J.W.: Generalized parameterized approximations. In: Yao, J.T., Ramanna, S., Wang, G., Suraj, Z. (eds.) RSKT 2011. LNCS, vol. 6954, pp. 136–145. Springer, Heidelberg (2011)
9. Clark, P.G., Grzymala-Busse, J.W.: Experiments on probabilistic approximations. In: Proceedings of the 2011 IEEE International Conference on Granular Computing, pp. 144–149 (2011)
10. Clark, P.G., Grzymala-Busse, J.W., Rzasa, W.: Mining incomplete data with singleton, subset and concept approximations. Inf. Sci. **280**, 368–384 (2014)
11. Clark, P.G., Grzymala-Busse, J.W.: Complexity of rule sets induced from incomplete data with lost values and attribute-concept values. In: Proceedings of the Third International Conference on Intelligent Systems and Applications, pp. 91–96 (2014)

12. Clark, P.G., Grzymala-Busse, J.W.: Mining incomplete data with lost values and attribute-concept values. In: Proceedings of the IEEE International Conference on Granular Computing, pp. 49–54 (2014)
13. Clark, P.G., Grzymala-Busse, J.W.: Mining incomplete data with many lost and attribute-concept values. In: Ciucci, D., Wang, G., Mitra, S., Wu, W.-Z. (eds.) RSKT 2015. LNCS, vol. 9436, pp. 100–109. Springer, Heidelberg (2015)
14. Clark, P.G., Grzymala-Busse, J.W.: On the number of rules and conditions in mining incomplete data with lost values and attribute-concept values. In: Proceedings of the DBKDA 7-th International Conference on Advances in Databases, Knowledge, and Data Applications, pp. 121–126 (2015)
15. Grzymala-Busse, J.W., Wang, A.Y.: Modified algorithms LEM1 and LEM2 for rule induction from data with missing attribute values. In: Proceedings of the 5-th International Workshop on Rough Sets and Soft Computing in Conjunction with the Third Joint Conference on Information Sciences, pp. 69–72 (1997)
16. Stefanowski, J., Tsoukias, A.: Incomplete information tables and rough classification. Comput. Intell. **17**(3), 545–566 (2001)
17. Pawlak, Z.: Rough sets. Int. J. Comput. Inform. Sci. **11**, 341–356 (1982)
18. Grzymala-Busse, J.W.: A new version of the rule induction system LERS. Fundamenta Informaticae **31**, 27–39 (1997)
19. Grzymala-Busse, J.W., Rzasa, W.: Definability and other properties of approximations for generalized indiscernibility relations. Trans. Rough Sets **11**, 14–39 (2010)