A Comparison of Mining Incomplete and Inconsistent Data

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. We present experimental results on a comparison of incompleteness and inconsistency. Our experiments were conducted on 141 data sets, including 71 incomplete data and 62 inconsistent, created from eight original numerical data sets. We used the Modified Learning from Examples Module version 2 (MLEM2) rule induction algorithm for data mining. Among eight types of data sets combined with three kinds of probabilistic approximations used in experiments, in 12 out of 24 combinations the error rate, computed as a result of ten-fold cross validation, was smaller for inconsistent data (two-tailed test, 5 % significance level). For one data set, combined with all three probabilistic approximations, the error rate was smaller for incomplete data. For remaining nine combinations the difference in performance was statistically insignificant. Thus, we may claim that there is some experimental evidence that incompleteness is generally worse than inconsistency for data mining.

Keywords: Incomplete data · Inconsistent data · Rough set theory · Probabilistic approximations · MLEM2 rule induction algorithm

1 Introduction

A complete data set, i.e., a data set having all attribute values specified, is consistent if for any two cases with the same attribute values, both cases belong to the same concept (class). Another definition of consistency is based on rough set theory: a complete data set is consistent if for any concept its lower and upper approximations are equal [9,10]. However, in some situations the data set being mined is either incomplete, some of the attribute values are missing; or inconsistent, there are cases that are indiscernable but belong to different concepts.

The main objective of our paper is to compare mining incomplete and inconsistent data in terms of an error rate computed as a result of ten-fold cross validation. Using eight numerical data sets, we discretized each of them and then converted to a symbolic and consistent data set with intervals as attribute

G. Dregvaite and R. Damasevicius (Eds.): ICIST 2016, CCIS 639, pp. 414–425, 2016.
DOI: 10.1007/978-3-319-46254-7_33

values. We then randomly replaced some of the intervals with symbols representing missing attribute values. This process was conducted incrementally, starting by randomly replacing 5 % of the intervals with missing attribute values, and then an additional 5 %, until a case occurred with all attribute values missing. The process was then attempted twice more with the maximum percentage and if again a case occurred with all attribute values missing, the process was terminated for that data set. The new data sets, with missing attribute values, were as close as possible to the original data sets, having the same number of attributes, cases, and concepts.

Additionally, any original data set was discretized with a controlled level of inconsistency, starting from about 5 %, with the same increment of about 5 %. Due to the nature of discretization, the levels of inconsistency were only approximately equal to 5 %, 10 %, etc. Our way of generation of inconsistent data preserved as much as possible the original data set. Again, the number of attributes, cases and concepts were not changed.

All such incomplete and inconsistent data sets were validated using the same setup, based on rule induction by the MLEM2 rule induction algorithm and the same system for ten-fold cross validation.

To the best of our knowledge, no research comparing incompleteness with inconsistency was ever undertaken. However, our results should be taken with a grain of salt since the measures of incompleteness and inconsistency are different. We measure both of them in the most natural way: for a data set, incompleteness is measured by the percentage of missing attribute values, or percentage of missing attribute values to the total number of cases in the data set. Inconsistency is measured by the level of inconsistency, i.e., percentage of conflicting cases to the number of cases. Yet the first measure is local, it is associated with the attribute-value pairs, while the second is global, it is computed by comparing entire cases. On the other hand, if we want to compare incompleteness with inconsistency, there is no better way than using these two measures.

In our experiments we used the idea of a probabilistic approximation, with a probability α, as an extension of the standard approximation, well known in rough set theory. For $\alpha = 1$, the probabilistic approximation is identical with the lower approximation; for very small α, it is identical with the upper approximation. Research on properties of probabilistic approximations was first reported in [12] and then was continued in many other papers, for example, [11,14–16].

Incomplete data sets are usually analyzed using special approximations such as singleton, subset and concept [4,5]. For incomplete data sets probabilistic approximations were used for the first time in [6]. The first experimental results using probabilistic approximations were published in [2]. In experiments reported in this paper, we used concept probabilistic approximations.

2 Incomplete Data

Data sets may be presented in the form of a decision table. An example of such a decision table is shown in Table 1. Rows of the decision table represent cases

and columns represent 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, $A = \{Age, Cholesterol, Weight\}$. The value for a case x and an attribute a will be denoted by $a(x)$.

Table 1. A data set with numerical attributes

Case	Attributes			Decision
	Age	Cholesterol	Weight	Risk
1	20	180	140	Low
2	60	200	180	Low
3	40	220	160	Low
4	50	200	180	Low
5	60	220	180	High
6	40	220	180	High
7	50	180	220	High

Table 2 presents an example of the discretized and consistent data set. All attribute values are intervals and as such are considered symbolic.

Table 2. A discretized, consistent data set

Case	Attributes			Decision
	Age	Cholesterol	Weight	Risk
1	20–45	180–210	140–170	Low
2	45–60	180–210	170–210	Low
3	20–45	210–220	140–170	Low
4	45–60	180–210	170–210	Low
5	45–60	210–220	170–210	High
6	20–45	210–220	170–210	High
7	45–60	180–210	210–220	High

Table 3 presents an example of an incomplete data set. In this paper, we use only one interpretation of missing attribute values, a lost value, denoted by "?" [8,13]. The percentage of missing attribute values is the total number of missing attribute values, equal to eight, divided by the total number of attribute values, equal to 21, i.e., the percentage of missing attribute values is 38.1 %.

Table 4 represent an inconsistent data set. This data set was created from the data set from Table 1. The numerical data set from Table 1 was discretized with 30 % level of inconsistency. Cases 3 and 6 are conflicting, so the level of inconsistency is $2/7 \approx 30\,\%$.

Table 3. An incomplete data set

Case	Attributes			Decision
	Age	Cholesterol	Weight	Risk
1	?	180–210	140–170	Low
2	45–60	?	170–210	Low
3	20–45	?	?	Low
4	45–60	180–210	170–210	Low
5	45–60	?	170–210	High
6	?	210–220	?	High
7	45–60	180–210	?	High

Table 4. An inconsistent data set

Case	Attributes			Decision
	Age	Cholesterol	Weight	Risk
1	20–45	180–210	140–210	Low
2	45–60	180–210	140–210	Low
3	20–45	210–220	140–210	Low
4	45–60	180–210	140–210	Low
5	45–60	210–220	140–210	High
6	20–45	210–220	140–210	High
7	45–60	180–210	210–220	High

A fundamental idea of rough set theory [9] is an indiscernibility relation, defined for complete data sets. Let B be a nonempty subset of the set A of all attributes. 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*. The set of all concepts is denoted by $\{d\}^*$. For example, a concept associated with the value *low* of the decision *Risk* 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 Table 4,

$$\underline{appr}_A(\{1,2,3,4\}) = \{1,2,4\}$$

and

$$\overline{appr}_A(\{1,2,3,4\}) = \{1,2,3,4,6\}.$$

The level of inconsistency may be defined as follows

$$1 - \frac{\sum_{X \in \{d\}^*} |\underline{appr}_A(X)|}{|U|},$$

where $|S|$ denotes the cardinality of the set S.

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 $\{x \in U \mid a(x) = v\}$ [3]. 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.

For the data set from Table 3 the blocks of attribute-value pairs are:

[(Age, 20–45)] = {3},
[(Age, 45–60)] = {2, 4, 5, 7},
[(Cholesterol, 180–210)] = {1, 4, 7},
[(Cholesterol, 210–220)] = {6},
[(Weight, 180–210)] = {1}, and
[(Weight, 170–220)] = {2, 4, 5}.

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,

For Table 3 and $B = A$,

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

First we will quote some definitions from [7]. 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 3 and $X = \{5, 6, 7\}$, all A-singleton, A-subset and A-concept lower and upper approximations are:

$\underline{appr}_A^{singleton}(X) = \{6\}$,
$\overline{appr}_A^{singleton}(X) = \{2, 5, 6, 7\}$,
$\underline{appr}_A^{subset}(X) = \{6\}$,
$\overline{appr}_A^{subset}(X) = \{2, 4, 5, 6, 7\}$,
$\underline{appr}_A^{concept}(X) = \{6\}$,
$\overline{appr}_A^{concept}(X) = \{2, 4, 5, 6, 7\}$.

3 Probabilistic Approximations

Definitions of lower and upper approximations may be extended to the probabilistic approximations [6]. In our experiments we used only concept approximations, so we will cite the corresponding definition only for the concept approximation. A B-concept probabilistic approximation of the set X with the threshold α, $0 < \alpha \leq 1$, denoted by $appr_{\alpha,B}^{concept}(X)$, is defined by

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

where $Pr(X \mid K_B(x)) = \frac{|X \cap K_B(x)|}{|K_B(x)|}$ is the conditional probability of X given $K_B(x)$.

Since we are using only B-concept probabilistic approximations, for the sake of simplicity we will call them B-probabilistic approximations. Additionally, if $B = A$, B-probabilistic approximations will be called simply probabilistic approximations and will be denoted by $appr_\alpha(X)$.

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

For Table 3 and the concept $X = \{5, 6, 7\}$, there exist the following distinct probabilistic approximations:

$appr_{1.0}(X) = \{6\}$,
$appr_{0.5}(X) = \{4, 6, 7\}$,
$appr_{0.333}(X) = \{2, 4, 5, 6, 7\}$.

A special probabilistic approximations with $\alpha = 0.5$ will be called *middle* approximations.

4 Experiments

Our experiments are based on eight data sets, all taken from the University of California at Irvine *Machine Learning Repository*. Essential information about these data sets is presented in Table 5. All eight data sets are numerical.

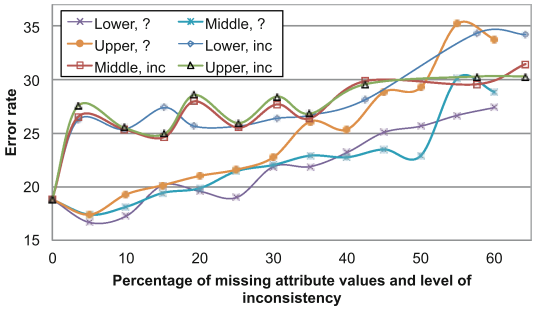

Fig. 1. Error rates for two series of data sets originated from the *Australian* data set. Incomplete data are denoted by "?", inconsistent data are denoted by "inc"

For any data set we created a series of incomplete data sets in the following way: first, the numerical data set was discretized using the agglomerative cluster analysis method [1]. Then we randomly replaced 5 % of specified attribute values by symbols of "?", denoting missing attribute values. After that, we replaced randomly and incrementally, with an increment equal to 5 %, new specified attribute values by symbols "?", preserving old ones. The process continued until we reached the point of having a case with all attribute values being "?"s. Then we returned to the one but last step and tried to add, randomly, 5 % of "?"s

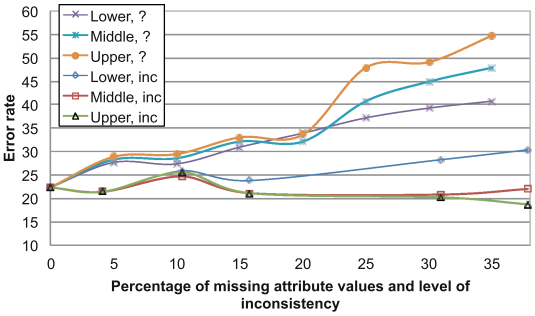

Fig. 2. Error rates for two series of data sets originated from the *Ecoli* data set. Incomplete data are denoted by "?", inconsistent data are denoted by "inc"

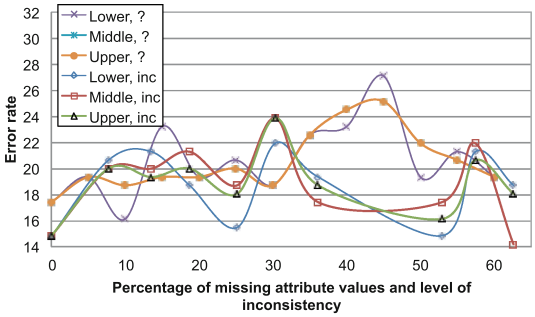

Fig. 3. Error rates for two series of data sets originated from the *Hepatitis* data set. Incomplete data are denoted by "?", inconsistent data are denoted by "inc"

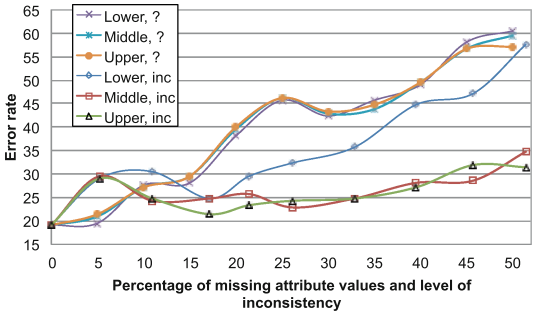

Fig. 4. Error rates for two series of data sets originated from the *Image Segmentation* data set. Incomplete data are denoted by "?", inconsistent data are denoted by "inc"

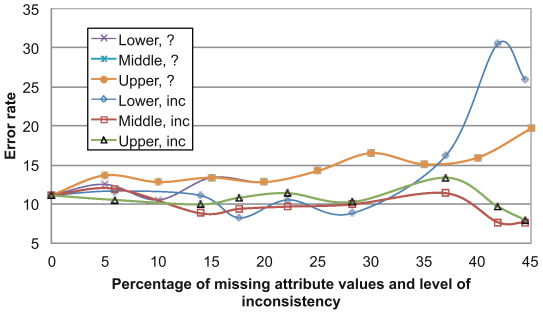

Fig. 5. Error rates for two series of data sets originated from the *Ionosphere* data set. Incomplete data are denoted by "?", inconsistent data are denoted by "inc"

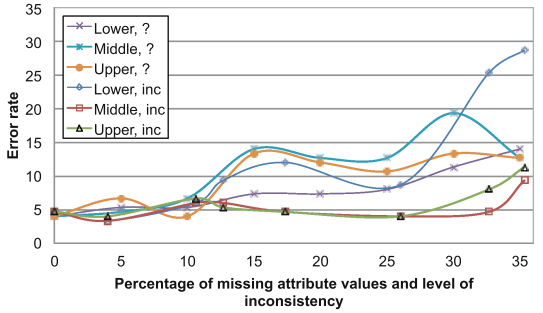

Fig. 6. Error rates for two series of data sets originated from the *Iris* data set. Incomplete data are denoted by "?", inconsistent data are denoted by "inc"

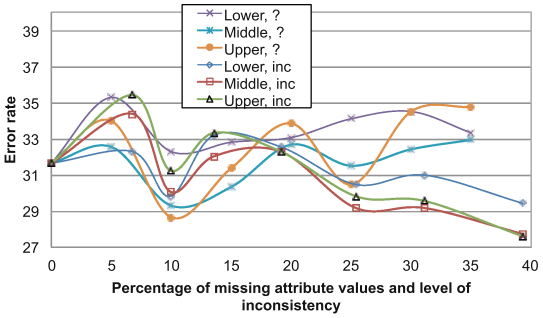

Fig. 7. Error rates for two series of data sets originated from the *Pima* data set. Incomplete data are denoted by "?", inconsistent data are denoted by "inc"

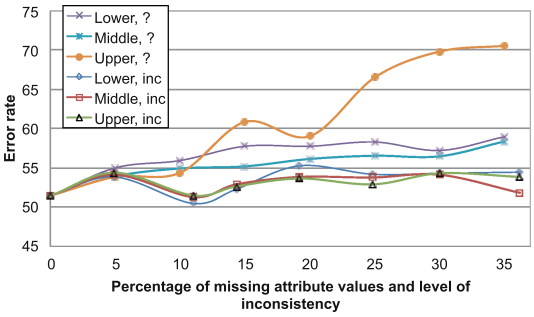

Fig. 8. Error rates for two series of data sets originated from the *Yeast* data set. Incomplete data are denoted by "?", inconsistent data are denoted by "inc"

again. If after three such attempts the result was still a case with "?"s as values for all attributes, the process was terminated. For example, for the *australian* data set such maximum for missing attribute values is 60 %.

For each original numerical data set, a series of inconsistent data sets was created by discretization, using the same agglomerative cluster analysis method as for the missing data sets. However, different levels of inconsistency were used as a stopping condition for discretization. Note that due to the nature of discretization, only some levels of inconsistency were possible to accomplish, so the levels of inconsistency are not as regular as percentage of missing attribute values. For example, for the *australian* data set these levels are 3.48, 9.71, 15.22 etc. instead of 5, 10, 15, as for the percentage of missing attribute values, though we tried to keep both series as close as possible.

Our experiments were conducted on 141 data sets, 71 among them were incomplete and 62 were inconsistent, 8 discretized and consistent data sets were used as special cases for both incomplete and inconsistent data sets.

For every data set we used three different probabilistic approximations for rule induction (lower, middle and upper). Thus we had 24 different approaches to rule induction. For rule induction we used the MLEM2 rule induction algorithm, a part of the Learning from Examples based on Rough Sets (LERS) data mining system [3].

For these 24 approaches we compared incomplete data with inconsistent ones for the same type of probabilistic approximations, using the Wilcoxon matched-pairs signed rank test, with 5 % level of significance, two-tailed test. Since we had 71 incomplete data sets and 62 inconsistent data sets, missing pairs were constructed by interpolation. Results of experiments rates for which there were no matching results, either incomplete or inconsistent, are not depicted in Figs. 1, 2, 3, 4, 5, 6, 7 and 8.

Results of our experiments, presented in Figs. 1, 2, 3, 4, 5, 6, 7 and 8, are: among 24 approaches, in 12 inconsistency was better (the error rate was smaller for inconsistent data). The *australian* data set was an exception, for all three probabilistic approximations the error rate was significantly smaller for incom-

Table 5. Data sets

Data set	Cases	Number of attributes	Concepts
Australian	690	14	2
Ecoli	336	8	8
Hepatitis	155	19	2
Image Segmentation	210	19	7
Ionosphere	351	34	2
Iris	150	4	3
Pima	768	8	2
Yeast	1484	8	9

plete data sets. For remaining nine approaches the difference between incompleteness and inconsistency was statistically insignificant.

In summary, there is evidence that inconsistency in data sets is less harmful for mining data than incompleteness, though more research is required.

5 Conclusions

As a results of our experiments, conducted on 141 data sets, including 71 incomplete data and 62 inconsistent, in 12 out of 24 combinations of the type of the original data set and a type of approximation, the error rate was smaller for inconsistent data. For one data set, combined with all three probabilistic approximations, the error rate was smaller for incomplete data. For remaining nine combinations the difference in performance was statistically insignificant. Thus, we may claim that there is some experimental evidence that incompleteness is generally worse than inconsistency for data mining.

References

1. Chmielewski, M.R., Grzymala-Busse, J.W.: Global discretization of continuous attributes as preprocessing for machine learning. Int. J. Approximate Reasoning **15**(4), 319–331 (1996)
2. 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)
3. Grzymala-Busse, J.W.: A new version of the rule induction system LERS. Fundamenta Informaticae **31**, 27–39 (1997)
4. 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)
5. Grzymala-Busse, J.W.: Data with missing attribute values: Generalization of indiscernibility relation and rule induction. Trans. Rough Sets **1**, 78–95 (2004)

6. Grzymala-Busse, J.W.: Generalized parameterized approximations. In: Proceedings of the 6-th International Conference on Rough Sets and Knowledge Technology, pp. 136–145 (2011)
7. Grzymala-Busse, J.W., Rzasa, W.: Definability and other properties of approximations for generalized indiscernibility relations. Trans. Rough Sets **11**, 14–39 (2010)
8. 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)
9. Pawlak, Z.: Rough sets. Int. J. Comput. Inform. Sci. **11**, 341–356 (1982)
10. Pawlak, Z.: Rough Sets. Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Dordrecht (1991)
11. Pawlak, Z., Skowron, A.: Rough sets: some extensions. Inf. Sci. **177**, 28–40 (2007)
12. Pawlak, Z., Wong, S.K.M., Ziarko, W.: Rough sets: probabilistic versus deterministic approach. Int. J. Man Mach. Stud. **29**, 81–95 (1988)
13. Stefanowski, J., Tsoukias, A.: Incomplete information tables and rough classification. Comput. Intell. **17**(3), 545–566 (2001)
14. Yao, Y.Y.: Probabilistic rough set approximations. Int. J. Approximate Reasoning **49**, 255–271 (2008)
15. Yao, Y.Y., Wong, S.K.M.: A decision theoretic framework for approximate concepts. Int. J. Man Mach. Stud. **37**, 793–809 (1992)
16. Ziarko, W.: Probabilistic approach to rough sets. Int. J. Approximate Reasoning **49**, 272–284 (2008)