On the Number of Rules and Conditions in Mining Data with Attribute-Concept Values and "Do Not Care" Conditions

Patrick G. Clark[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
[2] Department of Expert Systems and Artificial Intelligence, University of Information Technology and Management, 35-225 Rzeszow, Poland
jerzy@ku.edu

Abstract. In this paper we discuss two interpretations of missing attribute values: attribute-concept values and "do not care" conditions. Experiments were conducted on eight kinds of data sets, using three types of probabilistic approximations: singleton, subset and concept. Rules were induced by the MLEM2 rule induction system. Our main objective was to test which interpretation of missing attribute values provides simpler rule sets in terms of the number of rules and the total number of conditions. Our main result is that experimental evidence exists showing rule sets induced from data sets with attribute-concept values are simpler than the rule sets induced from "do not care" conditions.

1 Introduction

The most fundamental ideas of rough set theory are lower and upper approximations. In this paper we study probabilistic approximations. A probabilistic approximation, associated with a probability α, is a generalization of the standard approximation. For $\alpha = 1$, the probabilistic approximation becomes the lower approximation; for very small positive α, it becomes the upper approximation. Research on theoretical properties of probabilistic approximations started from [16] and then continued in many papers, see, e.g., [15–17,19–21].

Incomplete data sets may be analyzed using global approximations such as singleton, subset and concept [8–10]. Probabilistic approximations for incomplete data sets and based on an arbitrary binary relation were introduced in [12]. The first experimental results using probabilistic approximations were published in [1].

For our experiments we used eight incomplete data sets with two types of missing attribute values: attribute-concept values [11] and "do not care" conditions [4,13,18]. Additionally, in our experiments we used three types of probabilistic approximations: singleton, subset and concept.

In [3], the results indicate that rule set performance in terms of error rate is not significantly different for both missing attribute value interpretations. As a

M. Kryszkiewicz et al. (Eds.): PReMI 2015, LNCS 9124, pp. 13–22, 2015.
DOI: 10.1007/978-3-319-19941-2_2

result, given two rule sets with the same error rate, the more desirable would be the least complex, both for comprehension and computation performance. Therefore, the main objective of this paper is research on the complexity of rule sets induced from data sets with attribute-concept values and "do not care" conditions. Complexity is defined in terms of the number of rules and the number of rule conditions, with larger numbers indicating greater complexity.

Initially, the total number of rules and conditions in rule sets induced from incomplete data sets with attribute-concept values and "do not care" conditions were studied in [2]. However, in [2] only one type of probabilistic approximations was considered (concept) while in this paper we consider three types of probabilistic approximations (singleton, subset and concept). Additionally, in [2] only three values of α were discussed (0.001, 0.5 and 11.0) while in this paper we consider eleven values of α (0.001, 0.1, 0.2,..., 1.0).

Note that there are dramatic differences in complexity of rule sets induced from data sets with attribute-concept values and "do not care" conditions. For example, for the *bankruptcy* data set and concept approximation with $\alpha = 1.0$, the rule set induced from this data set in which missing attribute values were interpreted as attribute-concept values has four rules with seven conditions, while the rule set induced from the same data set in which missing attribute values were interpreted as "do not care" conditions has 13 rules with 31 conditions. The error rate, measured by ten-fold cross validation for the data set with attribute-concept values is 24.24 %, while the error rate for the same data set with "do not care" conditions is 37.88 %.

Our main result is that the simpler rule sets are induced from data sets in which missing attribute values are interpreted as attribute-concept values.

Our secondary objective was to identify the probabilistic approximation (singleton, subset or concept) that is associated with the lowest rule complexity. Our conclusion is that there is weak evidence that the best probabilistic approximation is subset.

2 Incomplete Data

We assume that the input data sets are presented in the form of a *decision table*. Rows of the decision table represent *cases*, while columns are labeled by *variables*. The set of all cases will be denoted by U. 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. 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: attribute-concept values and "do not care" conditions. *Attribute-concept values*, denoted by "−", indicate that the missing attribute value may be replaced by any of the values that have been specified for that attribute in a given concept. For example, if a patient is sick with flu, and if for other such patients the value of temperature is high or very-high, then we will replace the missing attribute values of temperature by values high and very-high, for details see [11].

"Do not care" conditions, denoted by "*", mean that the original attribute values are irrelevant, so we may replace them by any attribute value, for details see [4,13,18].

One of the most important ideas of rough set theory [14] 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 follows:

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

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 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

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

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

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

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 $\{x \in U \mid a(x) = v\}$ [5].

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 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 as follows

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

- If for an attribute a there exists a case x such that $a(x) = *$, i.e., the corresponding value is a "do not care" condition, then the case x should not be included in any blocks $[(a, v)]$ for all values v of attribute a.

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 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, $K(x, a) = U$,
- If $a(x) = *$ then the set $K(x, a) = U$, where U is the set of all cases.

3 Probabilistic Approximations

For incomplete data sets we may define approximations in many different ways [8]. For the lack of space, we are going to define only probabilistic approximations.

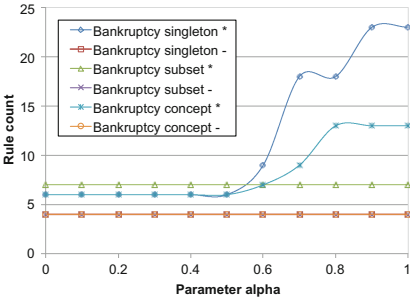

Fig. 1. Size of the rule set for the *Bankruptcy* data set

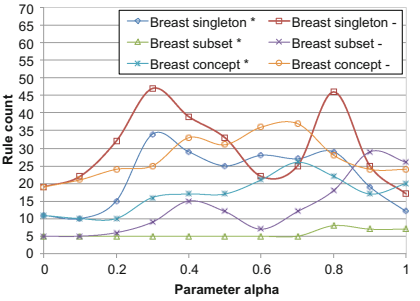

Fig. 2. Size of the rule set for the *Breast cancer* data set

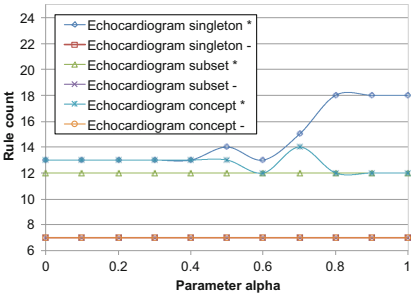

Fig. 3. Size of the rule set for the *Echocardiogram* data set

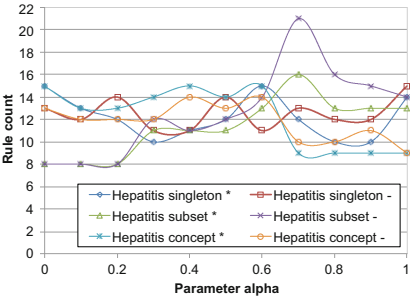

Fig. 4. Size of the rule set for the *Hepatitis* data set

A B-singleton probabilistic approximation of X with the threshold α, $0 < \alpha \leq 1$, denoted by $appr_{\alpha,B}^{singleton}(X)$, is defined as follows

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

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.

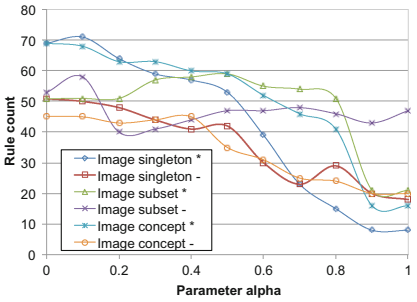

Fig. 5. Size of the rule set for the *Image segmentation* data set

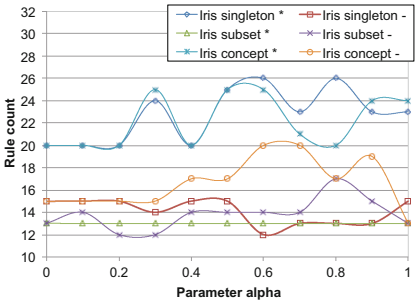

Fig. 6. Size of the rule set for the *Iris* data set

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 as follows

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

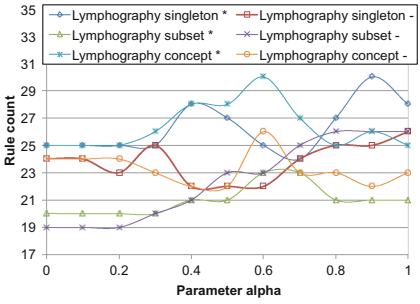

Fig. 7. Size of the rule set for the *Lymphography* data set

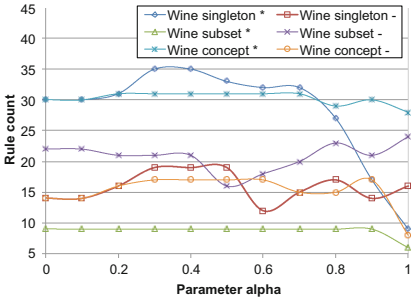

Fig. 8. Size of the rule set for the *Wine recognition* data set

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 as follows

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

For simplicity, the *A-singleton probabilistic approximation* will be called a *singleton probabilistic approximation*, *A-subset probabilistic approximation* will be called a *subset probabilistic approximation*, and *A-concept probabilistic approximation* will be called a *concept probabilistic approximation*.

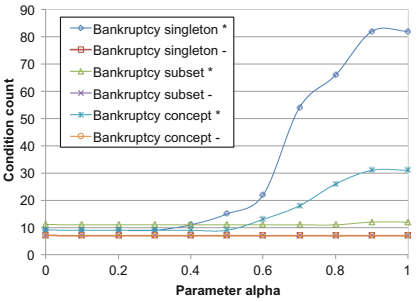

Fig. 9. Number of conditions for the *Bankruptcy* data set

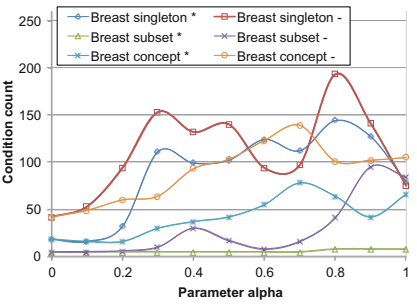

Fig. 10. Number of conditions for the *Breast cancer* data set

4 Experiments

Our experiments were conducted on eight data sets that are available from the University of California at Irvine *Machine Learning Repository*. For every data set a template was created by replacing randomly 35 % of existing specified attribute values by *attribute-concept values*. The same template was used for constructing a corresponding data set with *"do not care" conditions*, by replacing "–"s by "*"s. For two data sets, *bankruptcy* and *iris*, replacing more than 35 % of existing specified values by missing attribute values resulted in cases where all attribute values were missing. Hence we used for our experiments data sets with exactly 35 % missing attribute values.

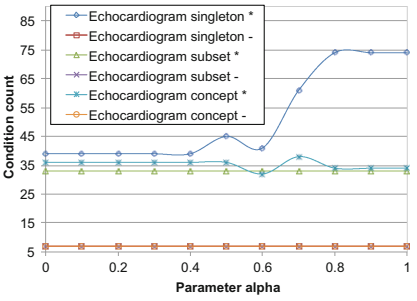

Fig. 11. Number of conditions for the *Echocardiogram* data set

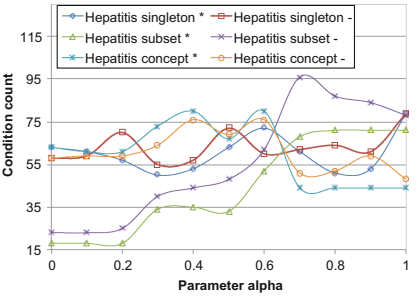

Fig. 12. Number of conditions for the *Hepatitis* data set

In our experiments, for any data set with given type of missing attribute values a rule set was induced using three types of probabilistic approximations: singleton, subset and concept, resulting in 24 combinations. For every such combination, rule sets induced from a data set with attribute-concept values

and the corresponding data set with "do not care" conditions were induced, for all eleven values of the parameter α, $\alpha = 0.001$, 01, 0.2,..., 1.0. Both the total number of rules and the total number of conditions in the rule set were compared using the Wilcoxon matched-pairs signed rank test with a 5 % level of significance, two-tailed test.

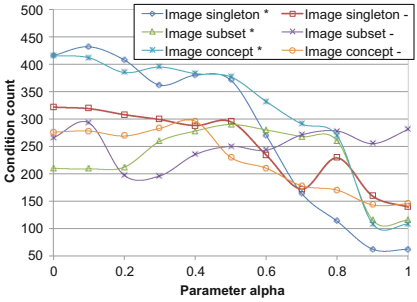

Fig. 13. Number of conditions for the *Image segmentation* data set

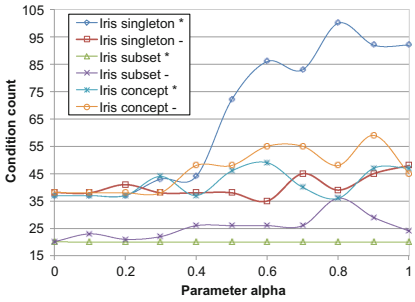

Fig. 14. Number of conditions for the *Iris* data set

In our experiments, we used the MLEM2 rule induction algorithm of the Learning from Examples using Rough Sets (LERS) data mining system [1,6,7]. Results of our experiments are presented in Figs. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 and 16.

The total number of rules was smaller for attribute-concept values than for "do not care" conditions for 13 combinations: for the *bankruptcy* and *echocardiogram* data sets with all three types of probabilistic approximations, for the *image* data set with concept probabilistic approximations, and for the *iris* and *lymphography* data sets for singleton and concept probabilistic approximations. On the other hand, the total number of rules was smaller for "do not care" conditions than for attribute-concept values for five combinations: for the *breast cancer* data set with singleton, subset and concept approximations and for the *hepatitis* and *wine recognition* data sets with subset probabilistic approximations. For the remaining six combinations of the data set and probabilistic approximation type the difference between the number of rules induced from the attribute-concept values and "do not care" conditions was statistically insignificant.

Similarly, for the same 24 combinations we compared the total number of conditions in rule sets. For 13 combinations the total number of conditions was smaller for data sets with attribute-concept values than for "do not care" conditions: for the *bankruptcy* and *echocardiogram* data sets with all three types of probabilistic approximations, for the *image* data set and concept probabilistic approximations and for the *iris* and *lymphography* data sets with singleton and subset probabilistic approximations and for the and *wine recognition* data set with singleton and subset approximations. However, for 5 combinations the

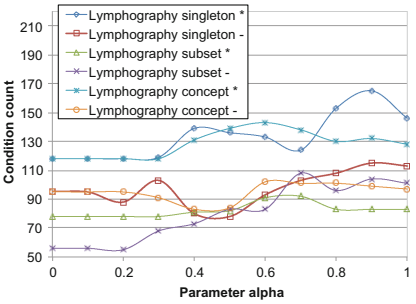

Fig. 15. Number of conditions for the *Lymphography* data set

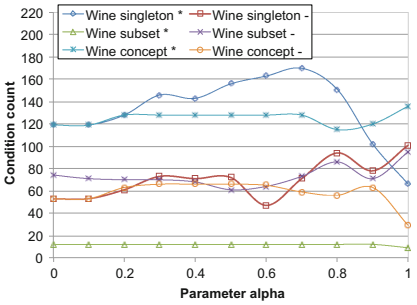

Fig. 16. Number of conditions for the *Wine recognition* data set

total number of conditions was smaller for "do not care" conditions than for attribute-concept values: for the *breast cancer* data set with all three types of probabilistic approximations, for the *hepatitis* data set with subset probabilistic approximations and for the *wine recognition* data set with concept approximations.

We may conclude that there is some evidence to support the idea that rule sets induced from data sets with attribute-concept values are simpler than rule sets induced from data sets with "do not care" conditions.

Our secondary objective was to select a type of probabilistic approximation that should be used for induction the simplest rules. Results of our experiments were divided into four groups, based on the type of the missing attribute values (attribute-concept values and "do not care" conditions) and whether the number of rules or the total number of conditions was used as a criterion of quality. Within each group we had 24 combinations (eight data sets and three types of probabilistic approximations). The Friedman multiple comparison rank sum test was applied, with 5 % significance level.

In our first group, where attribute-concept values were concerned with the number of rules, in one combination, associated with the *breast cancer* data set, the subset probabilistic approximations were better than the singleton probabilistic approximations and for another combination (for the *iris* data set) the subset probabilistic approximations were better than the concept probabilistic approximations. For the *wine recognition* data set, in two combinations, the concept probabilistic approximations were better than the remaining two probabilistic approximations. For the remaining 20 combinations results were statistically inconclusive.

For a group associated with "do not care" conditions and the number of rules, for nine combinations the subset approximations were better than other probabilistic approximations (for the *breast cancer*, *iris*, *lymphography* and *wine recognition* the subset probabilistic approximations were better than the remaining two probabilistic approximations and for the *echocardiogram* data set the subset probabilistic approximations were better than the singleton probabilistic

approximations). For the 15 other combinations the results were statistically inconclusive.

For the remaining two groups, both associated with the total number of conditions, the results were similar. In four combinations of attribute-concept values, the subset approximations were the best. For the remaining 15 combinations of attribute-concept values, the results were statistically inconclusive. For nine combinations of "do not care" conditions, the subset probabilistic approximations were the best. In the remaining 15 combinations of "do not care" conditions, the results were inconclusive. In summary, there is weak evidence that the subset probabilistic approximations are the best to be used for inducing the simplest rule sets.

5 Conclusions

As follows from our experiments, there is evidence that the rule set size is smaller for the attribute-concept interpretation of missing attribute values than for the "do not care" condition interpretation. The total number of conditions in rule sets is also smaller for attribute-concept interpretation of missing attribute values than for "do not care" condition interpretation. Thus we may claim attribute-concept values are better than "do not care" conditions as an interpretation of a missing attribute value in terms of rule complexity.

Furthermore, all three kinds of probabilistic approximations (singleton, subset and concept) do not differ significantly with respect to the complexity of induced rule sets. However, there exists some weak evidence that the subset probabilistic approximations are better than the remaining two: singleton and concept.

References

1. 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)
2. Clark, P.G., Grzymala-Busse, J.W.: Complexity of rule sets induced from incomplete data sets with attribute-concept values and and "do not care" conditions. In: Proceedings of the Third International Conference on Data Management Technologies and Applications, pp. 56–63 (2014)
3. Clark, P.G., Grzymala-Busse, J.W.: Mining incomplete data with attribute-concept values and "do not care" conditions. In: Polycarpou, M., de Carvalho, A.C.P.L.F., Pan, J.-S., Woźniak, M., Quintian, H., Corchado, E. (eds.) HAIS 2014. LNCS, vol. 8480, pp. 156–167. Springer, Heidelberg (2014)
4. Grzymala-Busse, J.W.: On the unknown attribute values in learning from examples. In: Raś, Zbigniew W., Zemankova, M. (eds.) ISMIS 1991. LNCS, vol. 542, pp. 368–377. Springer, Heidelberg (1991)
5. Grzymala-Busse, J.W.: LERS—a system for learning from examples based on rough sets. In: Slowinski, R. (ed.) Intelligent Decision Support. Handbook of Applications and Advances of the Rough Set Theory, pp. 3–18. Kluwer Academic Publishers, Dordrecht (1992)

6. Grzymala-Busse, J.W.: A new version of the rule induction system LERS. Fundamenta Informaticae **31**, 27–39 (1997)
7. Grzymala-Busse, J.W.: MLEM2: a new algorithm for rule induction from imperfect data. In: Proceedings of the 9th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, pp. 243–250 (2002)
8. 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)
9. Grzymała-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)
10. Grzymala-Busse, J.W.: Data with missing attribute values: generalization of indiscernibility relation and rule induction. Trans. Rough Sets **1**, 78–95 (2004)
11. Grzymala-Busse, J.W.: Three approaches to missing attribute values—a rough set perspective. In: Proceedings of the Workshop on Foundation of Data Mining, in conjunction with the Fourth IEEE International Conference on Data Mining, pp. 55–62 (2004)
12. 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)
13. Kryszkiewicz, M.: Rough set approach to incomplete information systems. In: Proceedings of the Second Annual Joint Conference on Information Sciences, pp. 194–197 (1995)
14. Pawlak, Z.: Rough sets. Int. J. Comput. Inf. Sci. **11**, 341–356 (1982)
15. Pawlak, Z., Skowron, A.: Rough sets: some extensions. Inf. Sci. **177**, 28–40 (2007)
16. Pawlak, Z., Wong, S.K.M., Ziarko, W.: Rough sets: probabilistic versus deterministic approach. Int. J. Man Mach. Stud. **29**, 81–95 (1988)
17. Ślęzak, D., Ziarko, W.: The investigation of the bayesian rough set model. Int. J. Approximate Reasoning **40**, 81–91 (2005)
18. 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. 73–82. Springer, Heidelberg (1999)
19. Yao, Y.Y.: Probabilistic rough set approximations. Int. J. Approximate Reasoning **49**, 255–271 (2008)
20. Yao, Y.Y., Wong, S.K.M.: A decision theoretic framework for approximate concepts. Int. J. Man Mach. Stud. **37**, 793–809 (1992)
21. Ziarko, W.: Probabilistic approach to rough sets. Int. J. Approximate Reasoning **49**, 272–284 (2008)