Rule Set Complexity for Incomplete Data Sets with Many Attribute-Concept Values and "Do Not Care" Conditions

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 results of novel experiments conducted on 12 data sets with many missing attribute values interpreted as attribute-concept values and "do not care" conditions. In our experiments complexity of rule sets, in terms of the number of rules and the total number of conditions induced from such data, are evaluated. The simpler rule sets are considered better. Our first objective was to check which interpretation of missing attribute values should be used to induce simpler rule sets. There is some evidence that the "do not care" conditions are better. Our secondary objective was to test which of the three probabilistic approximations: singleton, subset or concept, used for rule induction should be used to induce simpler rule sets. The best choice is the subset probabilistic approximation and the singleton probabilistic approximation is the worst choice.

Keywords: Incomplete data · Attribute-concept values · "Do not care" conditions · Probabilistic approximations · MLEM2 rule induction algorithm

1 Introduction

In this paper data sets with missing attribute values are mined using probabilistic approximations. The probabilistic approximation, with a probability α, is an extension of a standard approximation, a basic idea of rough set theory. If $\alpha = 1$, the probabilistic approximation becomes the lower approximation, for very small and positive α, the probabilistic approximation is identical with the upper approximation. The idea of the probabilistic approximation was introduced in [20] and further developed in [19,22–24].

Data sets with missing attribute values need special kinds of approximations, called singleton, subset and concept [12,13]. Such approximations were generalized to singleton, subset and concept probabilistic approximations in [15]. The

V. Flores et al. (Eds.): IJCRS 2016, LNAI 9920, pp. 65–74, 2016.
DOI: 10.1007/978-3-319-47160-0_6

first experiments on probabilistic approximations were presented in [1]. In experiments reported in this paper, we used all three kinds of probabilistic approximations: singleton, subset and concept.

In this paper, missing attribute values may be interpreted in two different ways, as attribute-concept values or as "do not care" conditions. Attribute-concept values, introduced in [14], are typical values for a given concept. For example, if the concept is a set of people sick with flu, and a value of the attribute *Temperature* is missing for some person who is sick with flu, using this interpretation, we would consider typical values of *Temperature* for other people sick with flu, such as *high* and *very high*. A "do not care" condition is interpreted as if the original attribute value was irrelevant, we may replace it by any existing attribute value [8,17,21].

The first experiments on data sets with missing attribute values interpreted as lost values and "do not care" conditions, with 35 % of missing attribute values, were reported in [7]. Research on data with missing attribute values interpreted as attribute-concept values and "do not care" conditions was presented in [2–6]. In [6] two imputation methods for missing attribute values were compared with rough-set approaches based on two interpretations of missing attribute values, as lost values and "do not care" conditions, combined with using singleton, subset and concept probabilistic approximations. It was shown that the rough-set approaches were better than imputation for five out of six data sets. The smallest error rate was associated with data sets with lost values. In [3] experiments were related to the error rate computed by ten-fold cross validation for mining data sets with attribute-concept values and "do not care" conditions using only three probabilistic approximations: lower, middle (with $\alpha = 0.5$) and upper. Results were not conclusive, in four cases attribute-concept values were better, while in two cases "do not care" conditions were better, in remaining 18 cases differences between the two were statistically insignificant. In [4] the error rate was evaluated for data sets with many missing attribute-concept values and "do not care" conditions. In two cases "do not care" conditions were better, in one case attribute-concept values were better, in remaining three cases differences were statistically insignificant.

With inconclusive results of experiments on the error rate, the question is which interpretation of missing attribute values is associated with smaller complexity of rule sets. In [2], experiments on complexity of rule sets induced from data sets with attribute-concept values and "do not care" conditions using lower, middle and upper approximations were presented. For half of the cases the number of rules was smaller for attribute-concept values, similarly for the total number of rule conditions. Results on the choice of the best type of probabilistic approximation (singleton, subset or concept) were inconclusive. In [5] experiments were also focused on complexity of rules sets, this time for data sets with 35 % of attribute-concept values and "do not care" conditions. For 13 combinations (out of 24) the attribute-concept values were associated with simpler rules, for five combinations "do not care" conditions were better, similarly for the total number of rule conditions.

The difference in performance between the two interpretations of missing attribute values, as attribute-concept values or "do not care" conditions, is more clear for data sets with many missing attribute values. Results of this paper are more conclusive than in our previous research.

Thus, our first objective was to check which interpretation of missing attribute values should be used to induce simpler rule sets, in terms of the number of rules and total number of rule conditions, from data sets with many attribute-concept values and "do not care" conditions, using the Modified Learning from Examples Module version 2 (MLEM2) system for rule induction [11]. There is some evidence that the "do not care" conditions are better. Our secondary objective was to test which of the three probabilistic approximations: singleton, subset or concept, used for rule induction should be used to induce simpler rule sets. The best choice is the subset probabilistic approximation and the singleton probabilistic approximation is the worst choice.

2 Incomplete Data

In this paper the input data sets are in the form of a *decision table*. A decision table has rows representing *cases* and columns defining *variables* with the set of all cases denoted by U. The dependent variable d is called the *decision* and the independent variables are labeled *attributes*. The set of all attributes will be denoted by A. Additionally, the value for a specific case x and attribute a is denoted by $a(x)$.

There are multiple ways to represent missing attribute values, however in this paper we distinguish them with two interpretations. The first, attribute-concept values, are identified using $-$ and the second, denoted by $*$ are "do not care" conditions.

One of the most important ideas of rough set theory [18] 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. When considering a complete data set, the *block* of (a, v), denoted by $[(a, v)]$, is the set $\{x \in U \mid a(x) = v\}$ [9]. However, when representing missing information and incomplete data sets, the definition of a block of an attribute-value pair is modified in the following way.

– For an attribute a, where there exists a case x such that $a(x) = -$, the case x should be included in blocks $[(a, v)]$ for all specified values $v \in V(x, a)$ of attribute a, where

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

– For an attribute a, where there exists a case x such that $a(x) = *$, the case x should be included in blocks $[(a, v)]$ for all specified values v of the 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 Lower and Upper Approximations

We quote some definitions from [16]. Let X be a subset of U and let B be a subset of the set A of all attributes. The B-*singleton lower approximation* of X, denoted by $\underline{appr}_B^{singleton}(X)$, is defined as follows

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

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

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

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

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

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

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

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

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

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

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

4 Probabilistic Approximations

The 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

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

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

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

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

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

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

In general, all three probabilistic approximations are distinct, even for the same value of the parameter α. Additionally, if for a given set X a probabilistic approximation $appr_\beta(X)$ is not listed, then $appr_\beta(X)$ is equal to the closest probabilistic approximation $appr_\alpha(X)$ of the same type with α larger than or equal to β.

If a characteristic relation $R(B)$ is an equivalence relation, all three types of probabilistic approximation: singleton, subset and concept are reduced to the same probabilistic approximation.

5 Experiments

Our experimental data sets are based on six data sets available from the University of California at Irvine *Machine Learning Repository*. Basic information about these data sets are presented in Table 1.

Incomplete data sets were produced from the base data by creating a set of templates. To create the templates, existing specified attribute values are replaced at 5 % increments with a corresponding *attribute-concept* value. So the

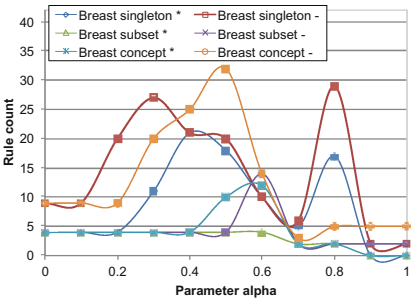

Fig. 1. Number of rules for the *breast cancer* data set

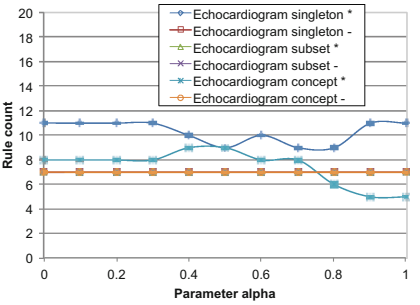

Fig. 2. Number of rules for the *echocardiogram* data set

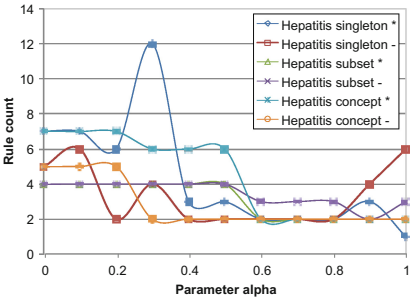

Fig. 3. Number of rules for the *hepatitis* data set

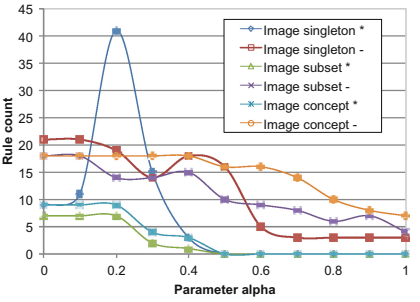

Fig. 4. Number of rules for the *image segmentation* data set

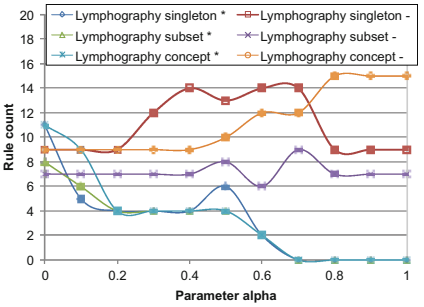

Fig. 5. Number of rules for the *lymphography* data set

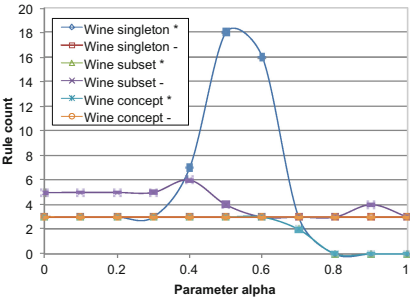

Fig. 6. Number of rules for the *wine recognition* data set

template creation begins with no missing values, then 5 % of the values are randomly replaced with *attribute-concept* values, then an additional 5 % are randomly replaced. The process continues with the data set until at least one row of the decision table attribute values are all missing values. Three attempts were made to randomly replace specified values with missing values where either a

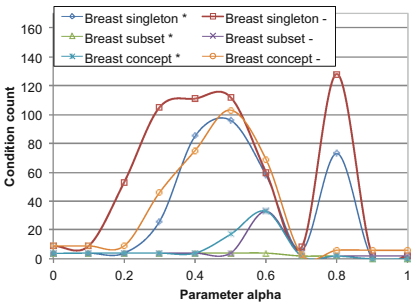

Fig. 7. Total number of conditions for the *breast cancer* data set

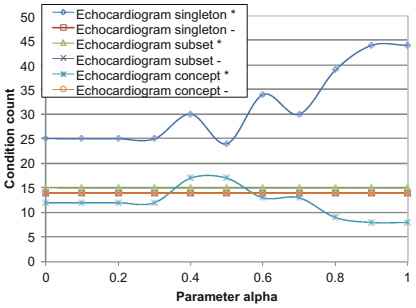

Fig. 8. Total number of conditions for the *echocardiogram* data set

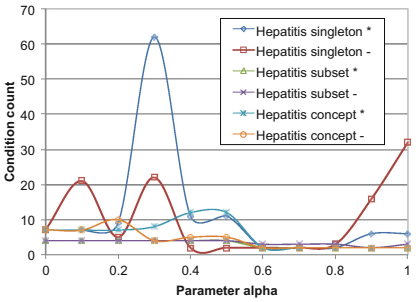

Fig. 9. Total number of conditions for the *hepatitis* data set

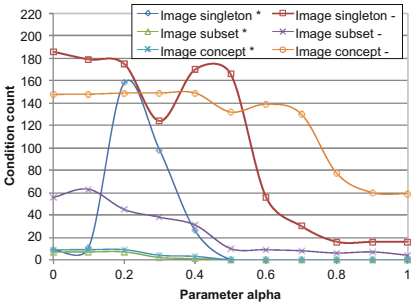

Fig. 10. Total number of conditions for the *image segmentation* data set

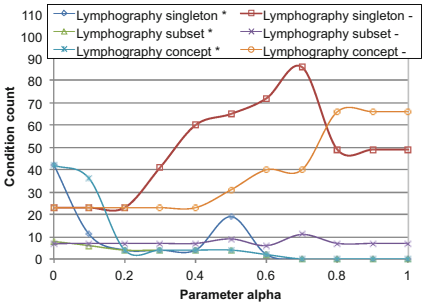

Fig. 11. Total number of conditions for the *lymphography* data set

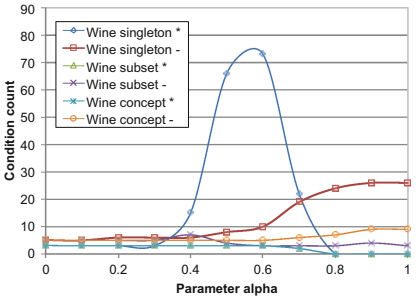

Fig. 12. Total number of conditions for the *wine recognition* data set

new data set with an extra 5 % is created or the process stops. To produce the "do not care" condition data sets, the same templates are used, replacing − with ∗.

In this paper, data sets with many missing attribute values are studied. We chose the maximum number of missing values that could be synthesized and for

Table 1. Data sets used for experiments

Data set	Number of			Percentage of
	Cases	Attributes	Concepts	Missing attribute values
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

this research, has been defined as more than 40 % of the values being replaced. As shown in Table 1, the maximum percentage of missing values ranges between 40.15 % and 69.89 %.

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

First we compared two interpretations of missing attribute values, attribute-concept values and "do not care" conditions with respect to the number of rules in a rule set. For every data set type, separately for singleton, subset and concept probabilistic approximations, the Wilcoxon matched-pairs signed rank test was used with a 5 % level of significance two-tailed test. With six data set types and three approximation types, the total number of combinations was 18.

For the number of rules in a rule set, for five combinations the "do not care" condition interpretation of missing attribute values was the best. For two combinations the attribute-concept values were the best. For the remaining 11 combinations the difference was not statistically significant. Similarly, for the total number of conditions in a rule set, for 11 combinations this number was smaller for "do not care" conditions, for two combinations attribute-concept values were the best, for the remaining five combinations the difference was not statistically significant.

Next, for a given interpretation of missing attribute values we compared all three types of probabilistic approximations in terms of the number of rules and the total number of conditions in a rule set using multiple comparisons based on Friedman's nonparametric test. Here, with six types of data sets and two interpretations of missing attribute values, the total number of combinations was 12. For the number of rules, the smallest number was associated with the subset probabilistic approximations for three combinations, with one tie between subset and concept probabilistic approximations. For remaining combinations the difference was not statistically significant. The singleton probabilistic approximation was never a winner. For the total number of rule conditions, the smallest

number was also associated with the subset probabilistic approximations for six combinations, with one tie between subset and concept probabilistic approximations. For remaining combinations the difference was not statistically significant. Again, the singleton probabilistic approximation was never a winner.

6 Conclusions

As follows from our experiments, there is some evidence that the number of rules and the total number of conditions are smaller for "do not care" conditions than for attribute-concept values. Additionally, the best probabilistic approximation that should be used for rule induction from data with many attribute-concept values and "do not care" conditions is the subset probabilistic approximation. On the other hand, the singleton probabilistic approximation is the worst.

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 (LNAI), vol. 8480, pp. 156–167. Springer, Heidelberg (2014). doi:10.1007/978-3-319-07617-1_14
4. Clark, P.G., Grzymala-Busse, J.W.: Mining incomplete data with many attribute-concept values and do not care conditions. In: Proceedings of the IEEE International Conference on Big Data, pp. 1597–1602 (2015)
5. Clark, P.G., Grzymala-Busse, J.W.: On the number of rules and conditions in mining data with attribute-concept values and "do not care" conditions. In: Kryszkiewicz, M., Bandyopadhyay, S., Rybinski, H., Pal, S.K. (eds.) PReMI 2015. LNCS, vol. 9124, pp. 13–22. Springer, Heidelberg (2015). doi:10.1007/978-3-319-19941-2_2
6. Clark, P.G., Grzymala-Busse, J.W., Kuehnhausen, M.: Mining incomplete data with many missing attribute values. a comparison of probabilistic and rough set approaches. In: Proceedings of the Second International Conference on Intelligent Systems and Applications, pp. 12–17 (2013)
7. Clark, P.G., Grzymala-Busse, J.W., Rzasa, W.: Mining incomplete data with singleton, subset and concept approximations. Inf. Sci. **280**, 368–384 (2014)
8. Grzymala-Busse, J.W.: On the unknown attribute values in learning from examples. In: Proceedings of the 6th International Symposium on Methodologies for Intelligent Systems, pp. 368–377 (1991)
9. 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)

10. Grzymala-Busse, J.W.: A new version of the rule induction system LERS. Fundamenta Informaticae **31**, 27–39 (1997)
11. 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)
12. 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)
13. Grzymala-Busse, J.W.: Data with missing attribute values: generalization of indiscernibility relation and rule induction. Trans. Rough Sets **1**, 78–95 (2004)
14. 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)
15. Grzymala-Busse, J.W.: Generalized parameterized approximations. In: Yao, J.T., Ramanna, S., Wang, G., Suraj, Z. (eds.) RSKT 2011. LNCS (LNAI), vol. 6954, pp. 136–145. Springer, Heidelberg (2011). doi:10.1007/978-3-642-24425-4_20
16. Grzymala-Busse, J.W., Rzasa, W.: Definability and other properties of approximations for generalized indiscernibility relations. Trans. Rough Sets **11**, 14–39 (2010)
17. Kryszkiewicz, M.: Rules in incomplete information systems. Inf. Sci. **113**(3–4), 271–292 (1999)
18. Pawlak, Z.: Rough sets. Int. J. Comput. Inform. Sci. **11**, 341–356 (1982)
19. Pawlak, Z., Skowron, A.: Rough sets: some extensions. Inf. Sci. **177**, 28–40 (2007)
20. Pawlak, Z., Wong, S.K.M., Ziarko, W.: Rough sets: probabilistic versus deterministic approach. Int. J. Man Mach. Stud. **29**, 81–95 (1988)
21. Stefanowski, J., Tsoukias, A.: Incomplete information tables and rough classification. Comput. Intell. **17**(3), 545–566 (2001)
22. Yao, Y.Y.: Probabilistic rough set approximations. Int. J. Approximate Reasoning **49**, 255–271 (2008)
23. Yao, Y.Y., Wong, S.K.M.: A decision theoretic framework for approximate concepts. Int. J. Man Mach. Stud. **37**, 793–809 (1992)
24. Ziarko, W.: Probabilistic approach to rough sets. Int. J. Approximate Reasoning **49**, 272–284 (2008)