Complexity of Rule Sets Induced by Characteristic Sets and Generalized Maximal Consistent Blocks

Patrick G. Clark[1], Cheng Gao[1], Jerzy W. Grzymala-Busse[1,2(✉)], Teresa Mroczek[2], and Rafal Niemiec[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
{tmroczek,rniemiec}@wsiz.rzeszow.pl

Abstract. We study mining incomplete data sets with two interpretations of missing attribute values, lost values and "do not care" conditions. For data mining we use characteristic sets and generalized maximal consistent blocks. Additionally, we use three types of probabilistic approximations, lower, middle and upper, so altogether we apply six approaches to data mining. Since it was shown that an error rate, associated with such data mining is not universally smaller for any approach, we decided to compare complexity of induced rule sets. Therefore, our objective is to compare six approaches to mining incomplete data sets in terms of complexity of induced rule sets. We conclude that there are statistically significant differences between these approaches.

Keywords: Incomplete data · Lost values · "do not care" conditions
Characteristic sets · Maximal consistent blocks
MLEM2 rule induction algorithm · Probabilistic approximations

1 Introduction

We study mining incomplete data sets with two interpretations of missing attribute values, lost values and "do not care" conditions. A missing attribute value is interpreted as lost if the original value existed but currently is unavailable, for example it is forgot or erased. A "do not care" condition means that the missing attribute value may be replaced by any value from the attribute domain. A "do not care" condition may occur as a result of a refusal to answer a question during the interview.

For data mining we use probabilistic approximations, a generalization of the lower and upper approximations, well known in rough set theory. A probabilistic approximation is associated with a parameter α, interpreted as a probability. When $\alpha = 1$, a probabilistic approximation becomes the lower approximation; if

L. Rutkowski et al. (Eds.): ICAISC 2018, LNAI 10842, pp. 301–310, 2018.
https://doi.org/10.1007/978-3-319-91262-2_27

α is small positive number, e.g., 0.001, a probabilistic approximation is the upper approximation. Initially, probabilistic approximations were applied to completely specified data sets [9,12–19]. Probabilistic approximations were generalized to incomplete data sets in [8].

Characteristic sets, for incomplete data sets with any interpretation of missing attribute values, were introduced in [7]. Maximal consistent blocks, restricted only to data sets with "do not care" conditions, were introduced in [11]. Additionally, in [11] maximal consistent blocks were used as granules to define only ordinary lower and upper approximations. A definition of the maximal consistent block was generalized to cover lost values and probabilistic approximations in [1]. The applicability of characteristic sets and maximal consistent blocks for mining incomplete data, from the view point of an error rate, was studied in [1]. As it happened, there is a small difference in quality of rule sets induced either way. Thus, we decided to compare characteristic sets with generalized maximal consistent blocks in terms of complexity of induced rule sets. In our experiments, the Modified Learning from Examples Module, version 2 (MLEM2) was used for rule induction [6].

2 Incomplete Data

In this paper, 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, $U = \{1, 2, 3, 4, 5, 6, 7, 8\}$. Independent variables are called *attributes* and a dependent variable is called a *decision* and is denoted by d. The set of all attributes is denoted by A. In Table 1, $A = \{$*Temperature*, *Headache*, *Cough*$\}$. The value for a case x and an attribute a is denoted by $a(x)$.

We distinguish between two interpretations of missing attribute values: lost values, denoted by "?" and "do not care" conditions, denoted by "$*$". Table 1 presents an incomplete data set with both lost values and "do not care" conditions.

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 *Flu* is the set $\{1, 2, 3, 4\}$.

For a completely specified data set, let a be an attribute and let v be a value of a. A *block* of (a, v), denoted by $[(a, v)]$, is the set $\{x \in U \mid a(x) = v\}$ [4].

For incomplete decision tables the definition of a block of an attribute-value pair (a, v) is modified in the following way.

- If for an attribute a and a case x we have $a(x) = ?$, the case x should not be included in any blocks $[(a, v)]$ for all values v of attribute a,
- If for an attribute a and a case x we have $a(x) = *$, the case x should be included in blocks $[(a, v)]$ for all specified values v of attribute a.

For the data set from Table 1 the blocks of attribute-value pairs are:
[(Temperature, normal)] = {3, 6, 8},

Table 1. A decision Table

	Attributes			Decision
Case	Temperature	Headache	Cough	Flu
1	high	yes	?	yes
2	high	no	*	yes
3	*	?	yes	yes
4	high	no	?	yes
5	?	no	*	no
6	normal	*	no	no
7	high	no	yes	no
8	*	no	?	no

[(Temperature, high)] = {1, 2, 3, 4, 7, 8},
[(Headache, no)] = {2, 4, 5, 6, 7, 8},
[(Headache, yes)] = {1, 6},
[(Cough, no)] = {2, 5, 6},
[(Cough, yes)] = {2, 3, 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 $a(x) =$? or $a(x) = *$, then $K(x, a) = U$.

For Table 1 and $B = A$,
$K_A(1) = \{1\}$,
$K_A(2) = \{2, 4, 7, 8\}$,
$K_A(3) = \{2, 3, 5, 7\}$,
$K_A(4) = \{2, 4, 7, 8\}$,
$K_A(5) = \{2, 4, 5, 6, 7, 8\}$,
$K_A(6) = \{6\}$,
$K_A(7) = \{2, 7\}$,
$K_A(8) = \{2, 4, 5, 6, 7, 8\}$.

A binary relation $R(B)$ on U, defined for $x, y \in U$ in the following way

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

will be called the *characteristic relation*. In our example $R(A) = \{(1, 1), (2, 2), (2, 4), (2, 7), (2, 8), (3, 2), (3, 3), (3, 5), (3, 7), (4, 2), (4, 4), (4, 7), (4, 8), (5, 2), (5, 4), (5, 5), (5, 6), (5, 7), (5, 8), (6, 6), (7, 2), (7, 7), (8, 2), (8, 4), (8, 5), (8, 6), (8, 7), (8, 8)\}$.

We quote some definitions from [1]. Let X be a subset of U. The set X is *B-consistent* if $(x, y) \in R(B)$ for any $x, y \in X$. If there does not exist a B-consistent subset Y of U such that X is a proper subset of Y, the set X is called a *generalized maximal B-consistent block*. The set of all generalized maximal B-consistent blocks will be denoted by $\mathscr{C}(B)$. In our example, $\mathscr{C}(A) = \{\{1\}, \{2, 4, 8\}, \{2, 7\}, \{3\} \{5, 8\}, \{6\}\}$.

Let $B \subseteq A$ and $Y \in \mathscr{C}(B)$. The set of all generalized maximal B-consistent blocks which include an element x of the set U, i.e. the set

$$\{Y | Y \in \mathscr{C}(B), x \in Y\}$$

will be denoted by $\mathscr{C}_B(x)$.

For data sets in which all missing attribute values are "do not care" conditions, an idea of a maximal consistent block of B was defined in [10]. Note that in our definition, the generalized maximal consistent blocks of B are defined for arbitrary interpretations of missing attribute values. For Table 1, the generalized maximal A-consistent blocks $\mathscr{C}_A(x)$ are
$\mathscr{C}_A(1) = \{\{1\}\}$,
$\mathscr{C}_A(2) = \{\{2, 4, 8\}, \{2, 7\}\}$,
$\mathscr{C}_A(3) = \{\{3\}\}$,
$\mathscr{C}_A(4) = \{\{2, 4, 8\}\}$,
$\mathscr{C}_A(5) = \{\{5, 8\}\}$,
$\mathscr{C}_A(6) = \{\{6\}\}$,
$\mathscr{C}_A(7) = \{\{2, 7\}$,
$\mathscr{C}_A(8) = \{\{2, 4, 8\}, \{5, 8\}\}$.

3 Probabilistic Approximations

In this section, we will discuss two types of probabilistic approximations: based on characteristic sets and on generalized maximal consistent blocks.

3.1 Probabilistic Approximations Based on Characteristic Sets

In general, probabilistic approximations based on characteristic sets may be categorized as singleton, subset and concept [3,7]. In this paper we restrict our attention only to concept probabilistic approximations, for simplicity calling them probabilistic approximations based on characteristic sets.

A *probabilistic approximation based on characteristic sets* of the set X with the threshold α, $0 < \alpha \leq 1$, denoted by $appr_{\alpha}^{CS}(X)$, is defined as follows

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

For Table 1 and both concepts $\{1, 2, 3, 4\}$ and $\{5, 6, 7, 8\}$, all distinct probabilistic approximations, based on characteristic sets, are

$appr_{0.5}^{CS}(\{1,2,3,4\}) = \{1,2,3,4,5,7,8\}$,

$appr_{1}^{CS}(\{1,2,3,4\}) = \{1\}$,

$appr_{0.667}^{CS}(\{5,6,7,8\}) = \{2,4,5,6,7,8\}$,

$appr_{1}^{CS}(\{5,6,7,8\}) = \{6\}$.

If for some β, $0 < \beta \leq 1$, a probabilistic approximation $appr_{\beta}^{CS}(X)$ is not listed above, it is equal to the probabilistic approximation $appr_{\alpha}^{CS}(X)$ with the closest α to β, $\alpha \geq \beta$. For example, $appr_{0.2}^{CS}(\{1,2,3,4\}) = appr_{0.5}^{CS}(\{1,2,3,4\})$.

3.2 Probabilistic Approximations Based on Generalized Maximal Consistent Blocks

By analogy with the definition of a probabilistic approximation based on characteristic sets, we may define a probabilistic approximation based on generalized maximal consistent blocks as follows:

A *probabilistic approximation* based on generalized maximal consistent blocks of the set X with the threshold α, $0 < \alpha \leq 1$, and denoted by $appr_{\alpha}^{MCB}(X)$, is defined as follows

$$\cup\{Y \mid Y \in \mathscr{C}_x(A),\ x \in X,\ Pr(X|Y) \geq \alpha\}.$$

All distinct probabilistic approximations based on generalized maximal consistent blocks are

$appr_{0.5}^{MCB}(\{1,2,3,4\}) = \{1,2,3,4,7,8\}$,

$appr_{0.667}^{MCB}(\{1,2,3,4\}) = \{1,2,3,4,8\}$,

$appr_{1}^{MCB}(\{1,2,3,4\}) = \{1,3\}$,

$appr_{0.333}^{MCB}(\{5,6,7,8\}) = \{2,4,5,6,7,8\}$,

$appr_{0.5}^{MCB}(\{5,6,7,8\}) = \{2,5,6,7,8\}$,

$appr_{1}^{MCB}(\{5,6,7,8\}) = \{5,6,8\}$.

4 Experiments

Our experiments were conducted on eight data sets that are available in the University of California at Irvine *Machine Learning Repository*. For any such data set a template was created by replacing (randomly) 5% of existing specified

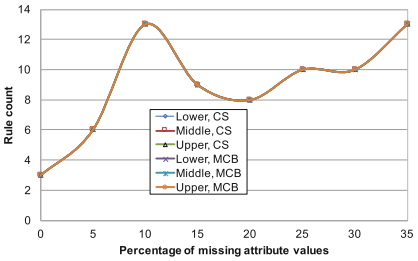

Fig. 1. Error rate for the *bankruptcy* data set with lost values

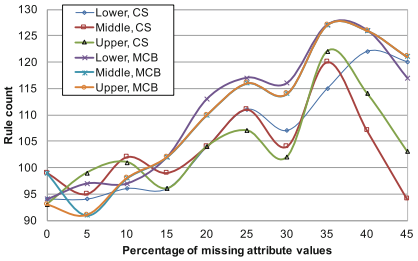

Fig. 2. Error rate for the *breast cancer* data set with lost values

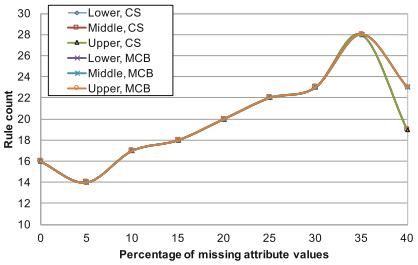

Fig. 3. Error rate for the *echocardiogram* data set with lost values

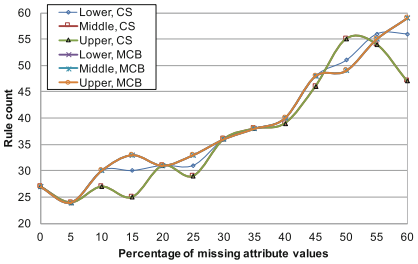

Fig. 4. Error rate for the *hepatitis* data set with lost values

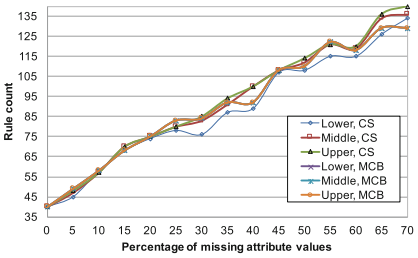

Fig. 5. Error rate for the *image segmentation* data set with lost values

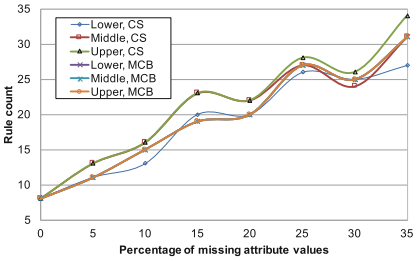

Fig. 6. Error rate for the *iris* data set with lost values

attribute values by *lost values*, then adding another 5% of lost values, and so on, until an entire row was full of lost values. The same templates were used for constructing data sets with "do not care" conditions, by replacing "?"s with "*"s, so we created 16 families of incomplete data sets.

In our experiments we used the MLEM2 rule induction algorithm of the LERS (Learning from Examples using Rough Sets) data mining system [2,5, 6]. We used characteristic sets and generalized maximal consistent blocks for mining incomplete datasets. Additionally, we used three different probabilistic

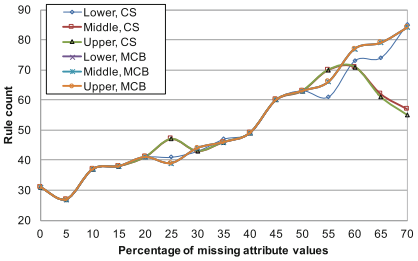

Fig. 7. Error rate for the *lymphography* data set with lost values

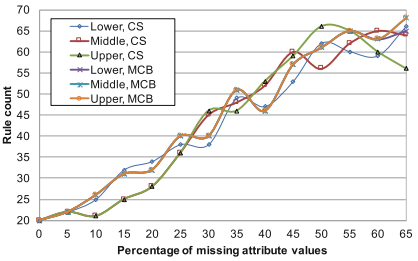

Fig. 8. Error rate for the *wine recognition* data set with lost values

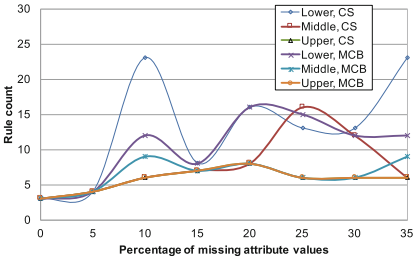

Fig. 9. Number of rules for the *bankruptcy* data set with "do not care" conditions

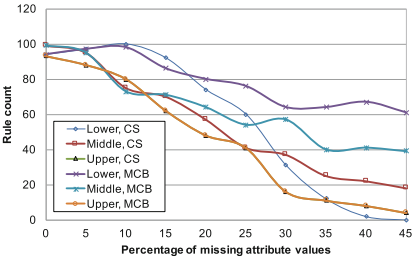

Fig. 10. Error rate for the *breast cancer* data set with "do not care" conditions

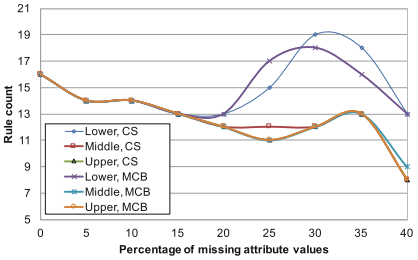

Fig. 11. Error rate for the *echocardiogram* data set with "do not care" conditions

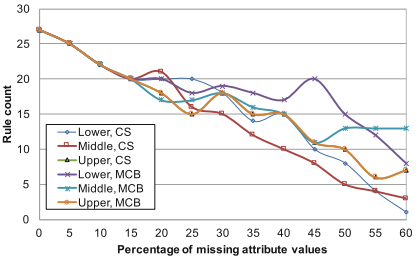

Fig. 12. Error rate for the *hepatitis* data set with "do not care" conditions

approximations, lower ($\alpha = 1$), middle ($\alpha = 0.5$) and upper ($\alpha = 0.001$). Thus our experiments were conducted on six different approaches to mining incomplete data sets. These six approaches were compared by applying the Friedman rank sum test combined with multiple comparisons, with a 5% level of significance. We applied this test to all 16 families of data sets, eight with lost values and eight with "do not care" conditions.

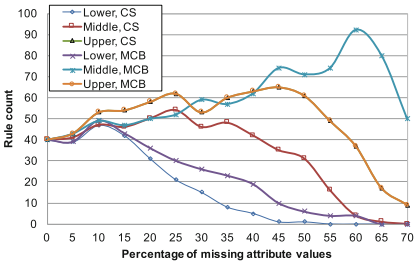

Fig. 13. Error rate for the *image segmentation* data set with "do not care" conditions

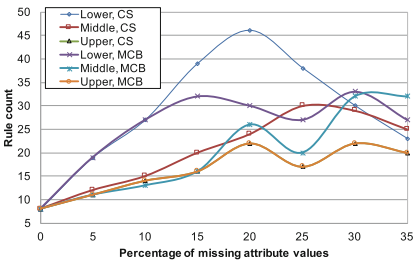

Fig. 14. Error rate for the *iris* data set with "do not care" conditions

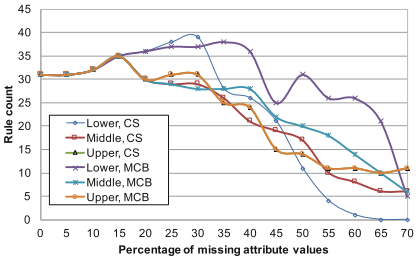

Fig. 15. Error rate for the *lymphography* data set with "do not care" conditions

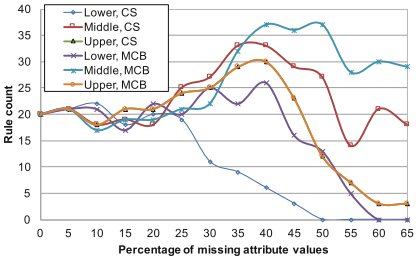

Fig. 16. Error rate for the *wine recognition* data set with "do not care" conditions

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, where "CS" denotes a characteristic set and "MCB" denotes a maximal consistent block. For eight data sets with lost values, the null hypothesis H_0 of the Friedman test saying that differences between these approaches are insignificant was rejected for four families of data sets (*breast cancer*, *hepatitis*, *image recognition* and *iris*). However, the post-hoc test (distribution-free multiple comparisons based on the Friedman rank sums) indicated that the differences between all six approaches were statistically insignificant for *breast cancer* and *hepatitis*. Results for *image recognition* and *iris* are listed in Table 2.

For eight data sets with "do not care" conditions, the null hypothesis H_0 of the Friedman test was rejected for all eight families of data sets. Additionally, for three families of data sets (*bankruptcy*, *echocardiogram* and *hepatitis* families of data sets the post-hoc test shown that the differences between all six approaches were insignificant. Results for the remaining five data sets are presented in Table 2. Obviously, for data sets with "do not care" conditions, *concept* upper approximations based on characteristic sets are identical with upper approximations based on maximal consistent blocks [11].

Table 2. Results of statistical analysis

Data set	Friedman test results (5% significance level)
Image recognition, ?	Lower, CS is better than Middle, CS and Upper, CS
	Lower, CS is better than all three approaches with MCB
Iris, ?	Lower, CS is better than Upper, CS
Breast cancer, *	Upper, CS is better than Lower, MCB
	Upper, MCB is better than Lower, MCB
Image recognition, *	Lower, CS is better than Upper, CS; Middle, MCB and Upper MCB
	Lower, MCB is better than Upper, CS; Middle, MCB and Upper MCB
Iris, *	Upper, CS is better than Lower, CS and Lower, MCB
	Upper, MCB is better than Lower, CS and Lower, MCB
Lymphography, *	Middle, CS is better than Lower, MCB
Wine recognition, *	Lower, CS is better than Middle, CS and Middle, MCB

5 Conclusions

Our objective was to compare six approaches to mining incomplete data sets (combining characteristic sets and generalized maximal consistent blocks with three types of probabilistic approximations). Our conclusion is that the choice between characteristic sets and generalized maximal consistent blocks and between types of probabilistic approximation is important, since there are statistically significant differences in complexity of induced rule sets. However, for every data set all six approaches should be tested and the best one should be selected. There is no universally best approach.

References

1. Clark, P.G., Gao, C., Grzymala-Busse, J.W., Mroczek, T.: Characteristic sets and generalized maximal consistent blocks in mining incomplete data. In: Polkowski, L., Yao, Y., Artiemjew, P., Ciucci, D., Liu, D., Ślęzak, D., Zielosko, B. (eds.) IJCRS 2017. LNCS (LNAI), vol. 10313, pp. 477–486. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-60837-2_39
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. Clark, P.G., Grzymala-Busse, J.W.: Experiments using three probabilistic approximations for rule induction from incomplete data sets. In: Proceeedings of the MCCSIS 2012, IADIS European Conference on Data Mining ECDM 2012, pp. 72–78 (2012)

4. 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)
5. Grzymala-Busse, J.W.: A new version of the rule induction system LERS. Fundam. Inform. **31**, 27–39 (1997)
6. 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)
7. 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)
8. Grzymała-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). https://doi.org/10.1007/978-3-642-24425-4_20
9. Grzymala-Busse, J.W., Ziarko, W.: Data mining based on rough sets. In: Wang, J. (ed.) Data Mining: Opportunities and Challenges, pp. 142–173. Idea Group Publishing, Hershey (2003)
10. Leung, Y., Li, D.: Maximal consistent block technique for rule acquisition in incomplete information systems. Inf. Sci. **153**, 85–106 (2003)
11. Leung, Y., Wu, W., Zhang, W.: Knowledge acquisition in incomplete information systems: a rough set approach. Eur. J. Oper. Res. **168**, 164–180 (2006)
12. Pawlak, Z., Skowron, A.: Rough sets: some extensions. Inf. Sci. **177**, 28–40 (2007)
13. Pawlak, Z., Wong, S.K.M., Ziarko, W.: Rough sets: probabilistic versus deterministic approach. Int. J. Man Mach. Stud. **29**, 81–95 (1988)
14. Ślęzak, D., Ziarko, W.: The investigation of the bayesian rough set model. Int. J. Approx. Reason. **40**, 81–91 (2005)
15. Wong, S.K.M., Ziarko, W.: INFER–an adaptive decision support system based on the probabilistic approximate classification. In: Proceedings of the 6-th International Workshop on Expert Systems and their Applications, pp. 713–726 (1986)
16. Yao, Y.Y.: Probabilistic rough set approximations. Int. J. Approx. Reason. **49**, 255–271 (2008)
17. Yao, Y.Y., Wong, S.K.M.: A decision theoretic framework for approximate concepts. Int. J. Man Mach. Stud. **37**, 793–809 (1992)
18. Ziarko, W.: Variable precision rough set model. J. Comput. Syst. Sci. **46**(1), 39–59 (1993)
19. Ziarko, W.: Probabilistic approach to rough sets. Int. J. Approx. Reason. **49**, 272–284 (2008)