On the Number of Conditions in Mining Incomplete Data Sets Using Characteristic Sets and Maximal Consistent Blocks

Patrick G. Clark and Cheng Gao
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA
Email: patrick.g.clark@gmail.com, cheng.gao@ku.edu

Jerzy W. Grzymala-Busse
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA
Department of Expert Systems and Artificial Intelligence, University of Information Technology and Management, Rzeszow, Poland
Email: jerzy@ku.edu

Teresa Mroczek
Department of Expert Systems and Artificial Intelligence, University of Information Technology and Management, Rzeszow, Poland
Email: tmroczek@wsiz.rzeszow.pl

***Abstract*—In this paper, we discuss incomplete data sets with two interpretations of missing attribute values, lost values and "do not care" conditions. For such incomplete data sets, we apply data mining based on characteristic sets and maximal consistent blocks. Our previous research shows that an error rate, evaluated by ten-fold cross validation, is sometimes smaller for characteristic sets and sometimes smaller for maximal consistent blocks. Therefore, we are taking the next step, comparing the quality of both approaches to mining incomplete data in terms of complexity of induced rule sets. We show that for data sets with lost values differences are insignificant while for data sets with "do not care" conditions rule sets are the simplest for upper approximations based on characteristic sets or maximal consistent blocks.**

Keywords–Data mining; rough set theory; probabilistic approximations; MLEM2 rule induction algorithm; lost values; "do not care" conditions.

I. INTRODUCTION

In this paper, we use two interpretations of a missing attribute value: lost values and "do not care" conditions. Lost values indicate that the original values were erased, and as a result we should use only existing, specified attribute values for rule induction. "Do not care" conditions mean that the missing attribute value may be replaced by any specified attribute value. Additionally, we use for data mining probabilistic approximations, a generalization of the idea of lower and upper approximations known in rough set theory. A probabilistic approximation is associated with a parameter (probability) α, if α = 1, a probabilistic approximation is reduced to the lower approximation; if α is small positive number, e.g., 0.001, a probabilistic approximation becomes the upper approximation. Usually probabilistic approximations are applied to completely specified data sets [1]–[9], such approximations were generalized to incomplete data sets in [10].

Characteristic sets were introduced in [11] for incomplete data sets with any interpretation of missing attribute values. On the other hand, maximal consistent blocks, introduced in [12], were restricted only to data sets with "do not care" conditions, using only lower and upper approximations. Definition of maximal consistent blocks was generalized to cover lost values and probabilistic approximations in [13]. Usefulness of characteristic sets and maximal consistent blocks to mining incomplete data in terms of an error rate was studied in [13]. It was shown that there is a small difference in quality of rule sets induced either way. Therefore, our current objective is to compare characteristic sets with maximal consistent blocks in terms of complexity of induced rule sets. In this paper, we show that for data sets with lost values differences are insignificant while for data sets with "do not care" conditions rule sets are the simplest for upper approximations based on characteristic sets or maximal consistent blocks. The Modified Learning from Examples Module, version 2 (MLEM2) [14] was used for rule induction.

This paper starts with a discussion on incomplete data in Section II where we define attribute-value blocks, characteristic sets and maximal consistent blocks. In Section III, we present probabilistic approximations based on characteristic sets and maximal consistent blocks. Section IV contains the details of our experiments. Finally, conclusions are presented in Section V.

II. INCOMPLETE DATA

We assume that the input data sets are presented in the form of a decision table. An example of a decision table is shown in Table I. 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 I, 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 will be denoted by A. In Table I, A = {*Wind*, *Humidity*, *Temperature*}. The value for a case x and an attribute a will be denoted by $a(x)$.

In this paper, we distinguish between two interpretations of missing attribute values: lost values, denoted by "?" and "do not care" conditions, denoted by "$*$". Table I presents an

 ISBN: 978-1-61208-631-6

TABLE I. A DECISION TABLE

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

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

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 $\{x \in U \mid a(x) = v\}$ [15]. 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 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 I, the blocks of attribute-value pairs are:

[(Wind, low)] = {2, 3, 4, 6, 8},
[(Wind, high)] = {1, 3, 5, 8},
[(Humidity, low)] = {1, 2, 4, 5, 8},
[(Humidity, high)] = {2, 5, 6, 7},
[(Temperature, medium)] = {1, 3, 4, 5, 6, 8}, and
[(Temperature, high)] = {2, 4, 5, 6}.

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

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

$$(x, y) \in R(B) \ if \ and \ only \ if \ y \in K_B(x) \qquad (1)$$

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

We quote some definitions from [13]. 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 consistent B-subset Y of U such that X is a proper subset of Y, the set X is called a *maximal B-consistent block*. The set of all B-maximal consistent blocks will be denoted by $\mathscr{C}(B)$. In our example, $\mathscr{C}(A)$ = {{1, 5, 8}, {2, 4}, {2, 6}, {3, 5}, {4, 8}, {7}}.

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

$$\{Y | Y \in \mathscr{C}(B), x \in Y\} \qquad (2)$$

will be denoted by $\mathscr{C}_x(B)$.

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 [16]. Note that in our definition, the maximal consistent blocks of B are defined for arbitrary interpretations of missing attribute values. For Table I, the maximal A-consistent blocks $\mathscr{C}_x(A)$ are

$\mathscr{C}_1(A) = \{\{1, 5, 8\}\}$,
$\mathscr{C}_2(A) = \{\{2, 4\}, \{2, 6\}\}$,
$\mathscr{C}_3(A) = \{\{3, 5\}\}$,
$\mathscr{C}_4(A) = \{\{2, 4\}, \{4, 8\}\}$,
$\mathscr{C}_5(A) = \{\{1, 5, 8\}, \{3, 5\}\}$,
$\mathscr{C}_6(A) = \{\{2, 6\}\}$,
$\mathscr{C}_7(A) = \{\{7\}\}$,
$\mathscr{C}_8(A) = \{\{1, 5, 8\}, \{4, 8\}\}$.

III. Probabilistic Approximations

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

A. Probabilistic Approximations Based on Characteristic Sets

In general, probabilistic approximations based on characteristic sets may be categorized as singleton, subset and concept [11][17]. 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

$$\cup\{K_A(x) \mid x \in X, \ Pr(X|K_A(x)) \geq \alpha\}. \qquad (3)$$

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

$$appr_{0.5}^{CS}(\{1, 2, 3, 4\}) = \{1, 2, 3, 4, 5, 6, 8\},$$

$$appr_{0.667}^{CS}(\{1, 2, 3, 4\}) = \{2, 4, 6, 8\},$$

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

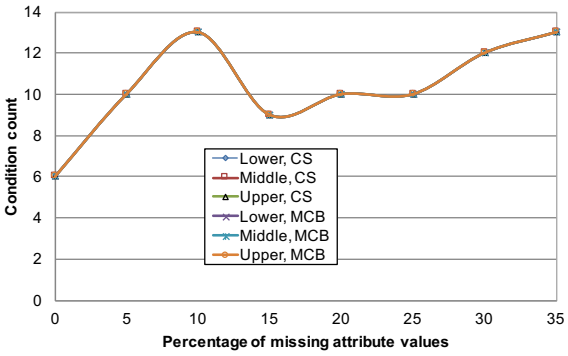

Figure 1. Number of conditions for the *Bankruptcy* data set with lost values

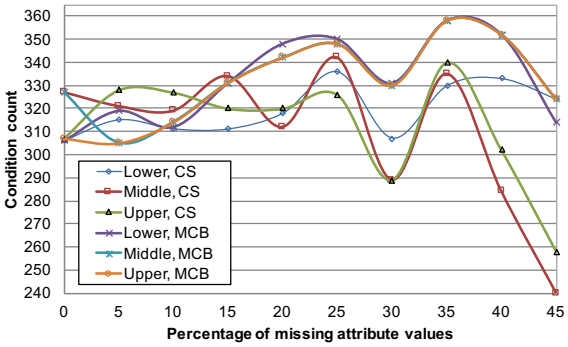

Figure 2. Number of conditions for the *Breast cancer* data set with lost values

$$appr_{0.5}^{CS}(\{5,6,7,8\}) = U,$$

$$appr_{0.75}^{CS}(\{5,6,7,8\}) = \{2,5,6,7\},$$

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

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, $appr_{0.6}^{CS}(\{1,2,3,4\}) = appr_{0.667}^{CS}(\{1,2,3,4\})$.

B. Probabilistic Approximations Based on Maximal Consistent Blocks

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

A *probabilistic approximation* based on 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

$$\cup\{Y \mid Y \in \mathscr{C}_x(A),\ x \in X,\ Pr(X|Y) \geq \alpha\}. \qquad (4)$$

All distinct probabilistic approximations based on maximal consistent blocks are

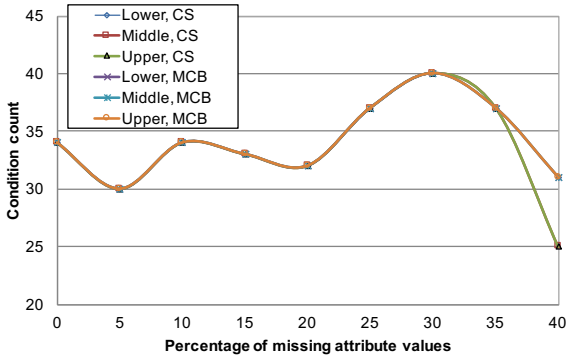

Figure 3. Number of conditions for the *Echocardiogram* data set with lost values

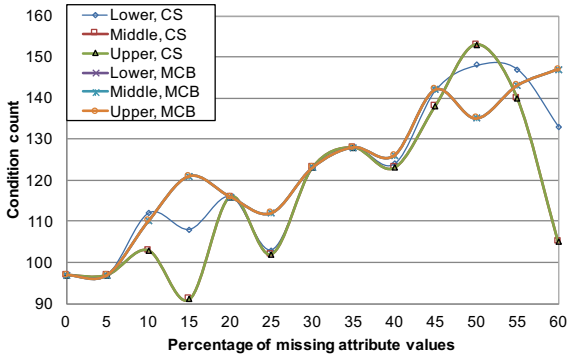

Figure 4. Number of conditions for the *Hepatitis* data set with lost values

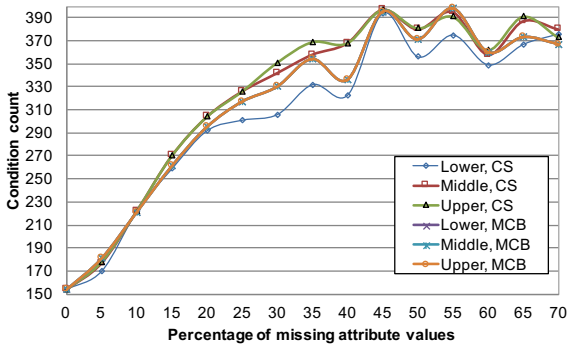

Figure 5. Number of conditions for the *Image segmentation* data set with lost values

$$appr_{0.333}^{MCB}(\{1,2,3,4\}) = \{1,2,3,4,5,6,8\},$$

$$appr_{0.5}^{MCB}(\{1,2,3,4\}) = \{2,3,4,5,6,8\},$$

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

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

$$appr_{0.667}^{MCB}(\{5,6,7,8\}) = \{1,5,7,8\},$$

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

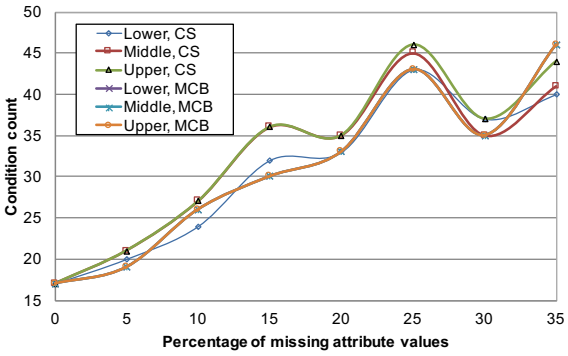

Figure 6. Number of conditions for the *Iris* data set with lost values

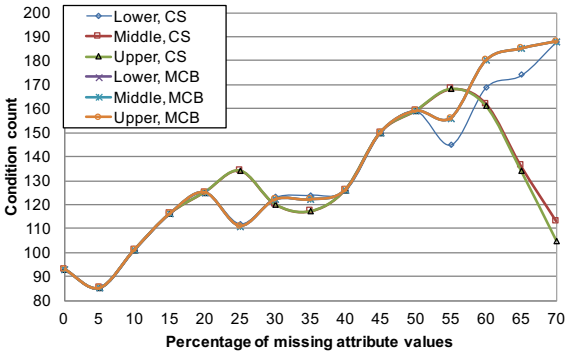

Figure 7. Number of conditions for the *Lymphography* data set with lost values

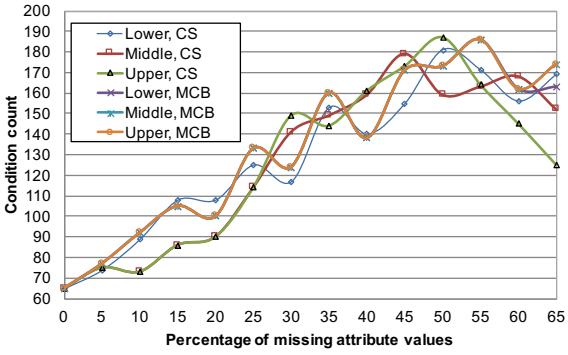

Figure 8. Number of conditions for the *Wine recognition* data set with lost values

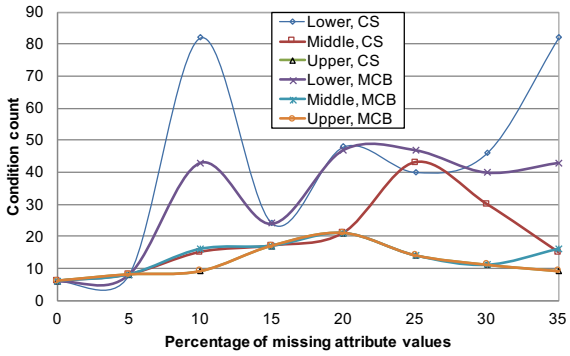

Figure 9. Number of conditions for the *Bankruptcy* data set with "do not care" conditions

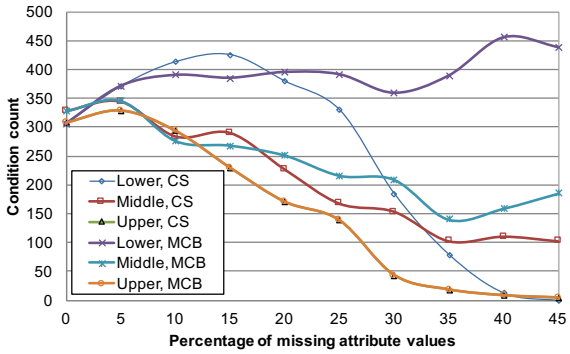

Figure 10. Number of conditions for the *Breast cancer* data set with "do not care" conditions

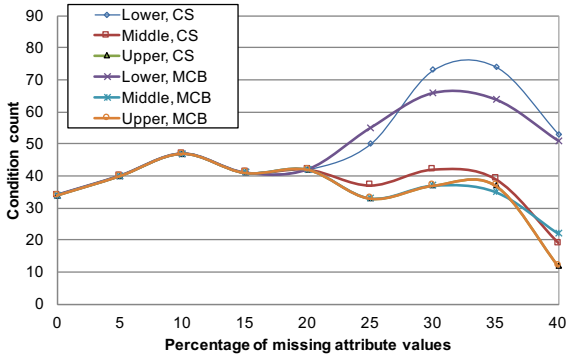

Figure 11. Number of conditions for the *Echocardiogram* data set with "do not care" conditions

IV. Experiments

For our experiments, we used eight data sets that are available in the University of California at Irvine *Machine Learning Repository*.

For every data set, a template was created. Such a template was formed by replacing randomly 5% of existing specified attribute values by *lost values*, then adding another 5% of specified 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.

In our experiments, we used an MLEM2 rule induction algorithm of the Learning from Examples using Rough Sets (LERS) data mining system [18]–[20]. Results of our experiments are presented in Figures 1–16, where "CS" denotes a characteristic set and "MCB" denotes a maximal consistent block. In our experiments, six approaches for mining incomplete data sets were used, since we combined two options: characteristic sets and maximal consistent blocks with three

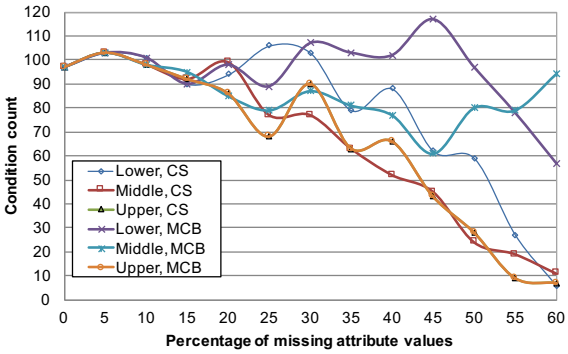

Figure 12. Number of conditions for the *Hepatitis* data set with "do not care" conditions

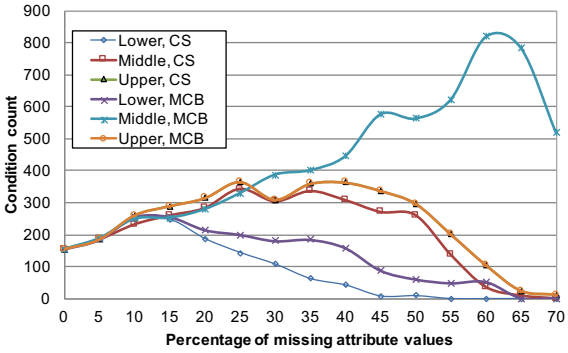

Figure 13. Number of conditions for the *Image segmentation* data set with "do not care" conditions

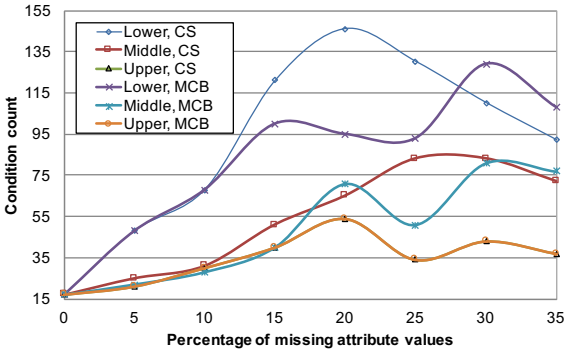

Figure 14. Number of conditions for the *Iris* data set with "do not care" conditions

options of probabilistic approximations: lower (α = 1), middle (α = 0.5) and upper (α = 0.001).

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 data sets, eight with lost values and eight with "do not care" conditions.

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 *image recognition*

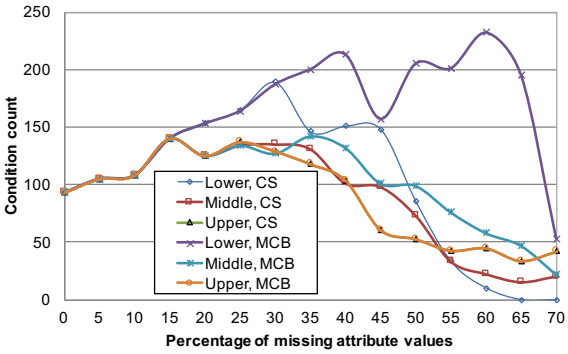

Figure 15. Number of conditions for the *Lymphography* data set with "do not care" conditions

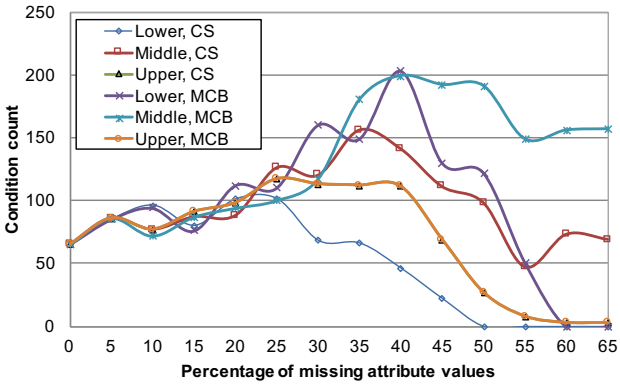

Figure 16. Number of conditions for the *Wine recognition* data set with "do not care" conditions

as the only data set. 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 eight data sets with "do not care" conditions, the null hypothesis H_0 of the Friedman test was rejected for all data sets except *wine recognition*. Additionally, for *echocardiogram* data set the post-hoc test shown that the differences between all six approaches were insignificant. Results for the remaining six data sets are presented in Table II. *Image segmentation* data set needs an additional explanation. For all three best approaches (lower approximation based on characteristic sets, lower approximation based on maximal consistent blocks and middle approximation based on characteristic sets) and for large percentages of missing attribute values, lower approximations are reduced to empty sets. This is due to the fact that both characteristic sets and maximal consistent blocks are large, so they cannot be subsets of corresponding concepts. Thus we may as well exclude this data set from further analysis. For remaining five data sets, clean winners are upper approximation based on characteristic sets and maximal consistent blocks. Obviously, for data sets with "do not care" conditions, *concept* upper approximations are identical with upper approximations based on maximal consistent blocks [12].

 ISBN: 978-1-61208-631-6

TABLE II. Results of statistical analysis

Data set	The best approaches	The worst approaches
Bankruptcy	Upper, CS; Upper, MCB	Lower, CS
Breast cancer	Upper, CS; Upper, MCB	Lower, MCB
Hepatitis	Upper, CS; Upper, MCB	Lower, MCB
Image recognition	Lower, CS; Lower, MCB; Middle, CS	Middle, MCB; Upper, CS; Upper, MCB
Iris	Upper, CS; Upper, MCB	Lower, CS; Lower, MCB
Lymphography	Middle, CS; Upper, CS; Upper, MCB	Lower, MCB

V. CONCLUSIONS

In this paper, we compare six approaches for mining incomplete data in terms of complexity of the rule sets. As follows from our experiments, for data sets with lost values, there is not significant difference between all six approaches. For data sets with "do not care" conditions, rule sets induced from upper approximations, based on characteristic sets or maximal consistent blocks, are the simplest in terms of the total number of conditions, in terms of complexity of rule sets.

REFERENCES

[1] J. W. Grzymala-Busse and W. Ziarko, "Data mining based on rough sets," in Data Mining: Opportunities and Challenges, J. Wang, Ed. Hershey, PA: Idea Group Publ., 2003, pp. 142–173.

[2] Z. Pawlak and A. Skowron, "Rough sets: Some extensions," Information Sciences, vol. 177, 2007, pp. 28–40.

[3] Z. Pawlak, S. K. M. Wong, and W. Ziarko, "Rough sets: probabilistic versus deterministic approach," International Journal of Man-Machine Studies, vol. 29, 1988, pp. 81–95.

[4] D. Ślęzak and W. Ziarko, "The investigation of the bayesian rough set model," International Journal of Approximate Reasoning, vol. 40, 2005, pp. 81–91.

[5] S. K. M. Wong and W. Ziarko, "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, 1986, pp. 713–726.

[6] Y. Y. Yao, "Probabilistic rough set approximations," International Journal of Approximate Reasoning, vol. 49, 2008, pp. 255–271.

[7] Y. Y. Yao and S. K. M. Wong, "A decision theoretic framework for approximate concepts," International Journal of Man-Machine Studies, vol. 37, 1992, pp. 793–809.

[8] W. Ziarko, "Variable precision rough set model," Journal of Computer and System Sciences, vol. 46, no. 1, 1993, pp. 39–59.

[9] ——, "Probabilistic approach to rough sets," International Journal of Approximate Reasoning, vol. 49, 2008, pp. 272–284.

[10] J. W. Grzymala-Busse, "Generalized parameterized approximations," in Proceedings of the 6-th International Conference on Rough Sets and Knowledge Technology, 2011, pp. 136–145.

[11] ——, "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, 2003, pp. 56–63.

[12] Y. Leung, W. Wu, and W. Zhang, "Knowledge acquisition in incomplete information systems: A rough set approach," European Journal of Operational Research, vol. 168, 2006, pp. 164–180.

[13] P. G. Clark, C. Gao, J. W. Grzymala-Busse, and T. Mroczek, "Characteristic sets and generalized maximal consistent blocks in mining incomplete data, part i," in Proceedings of the International Joint Conference on Rough Sets, 2017, pp. 477–486.

[14] P. G. Clark and J. W. Grzymala-Busse, "Experiments on rule induction from incomplete data using three probabilistic approximations," in Proceedings of the 2012 IEEE International Conference on Granular Computing, 2012, pp. 90–95.

[15] J. W. Grzymala-Busse, "LERS—a system for learning from examples based on rough sets," in Intelligent Decision Support. Handbook of Applications and Advances of the Rough Set Theory, R. Slowinski, Ed. Dordrecht, Boston, London: Kluwer Academic Publishers, 1992, pp. 3–18.

[16] Y. Leung and D. Li, "Maximal consistent block technique for rule acquisition in incomplete information systems," Information Sciences, vol. 153, 2003, pp. 85–106.

[17] P. G. Clark and J. W. Grzymala-Busse, "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, 2012, pp. 72–78.

[18] ——, "Experiments on probabilistic approximations," in Proceedings of the 2011 IEEE International Conference on Granular Computing, 2011, pp. 144–149.

[19] J. W. Grzymala-Busse, "A new version of the rule induction system LERS," Fundamenta Informaticae, vol. 31, 1997, pp. 27–39.

[20] ——, "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, 2002, pp. 243–250.

 ISBN: 978-1-61208-631-6