Complexity of Rule Sets Induced from Incomplete Data with Lost Values and Attribute-Concept Values

Patrick G. Clark
Department of Electrical Eng. and Computer Sci.
University of Kansas
Lawrence, KS, USA
e-mail: patrick.g.clark@gmail.com

Jerzy W. Grzymala-Busse
Department of Electrical Eng. and Computer Sci.
University of Kansas
Lawrence, KS, USA
Institute of Computer Science
Polish Academy of Sciences
Warsaw, Poland
e-mail: jerzy@ku.edu

***Abstract*—This paper presents novel research on complexity of rule sets induced from incomplete data sets with two interpretations of missing attribute values: lost values and attribute-concept values. Experiments were conducted on 176 data sets, using three kinds of probabilistic approximations (lower, middle and upper) and the Modified Learning from Examples Module, version 2 (MLEM2) rule induction system. In our experiments, the size of the rule set was always smaller for attribute-concept values than for lost values (5% significance level). The total number of conditions was smaller for attribute-concept values than for lost values for 17 combinations of the type of data set and approximation, out of 24 combinations total. In remaining 7 cases, the difference in performance was statistically insignificant. Thus, we may claim that attribute-concept values are better than lost values in terms of rule complexity.**

Keywords–Data mining; rough set theory; probabilistic approximations; MLEM2 rule induction algorithm; lost values; attribute-concept values.

I. Introduction

Standard lower and upper approximations are fundamental concepts of rough set theory. A probabilistic approximation, associated with a probability α, is a generalization of the standard approximation. For α = 1, the probability approximation is reduced to the lower approximation; for very small α, it is reduced to the upper approximation. Research on theoretical properties of probabilistic approximations started from [1] and then was continued in many papers, see, e.g., [1]–[6].

Incomplete data sets may be analyzed using global approximations such as singleton, subset and concept [7][8]. Probabilistic approximations, for incomplete data sets and based on an arbitrary binary relation, were introduced in [9], while first experimental results using probabilistic approximations were published in [10].

In this paper, incomplete data sets are characterized by missing attribute values. We will use two interpretations of a missing attribute value: lost values and attribute-concept values.

For our experiments we used 176 incomplete data sets, with two types of missing attribute values: lost values and attribute-concept values. Additionally, in our experiments we used three types of approximations: lower, upper, and additionally the most typical probabilistic approximation, for α = 0.5, called a middle approximation.

From our previous research it follows that the correctness of the rule sets, evaluated by ten-fold cross validated error rate, do not differ significantly with different combinations of missing attribute and approximation type.

In our experiments, the size of rule set was always smaller for attribute-concept values than for lost values. The total number of conditions in rule sets was smaller for attribute-concept values for 17 combinations of the type of data set and approximation (out of 24 combinations total). In remaining seven combinations, the total number of conditions in rule sets did not differ significantly. Thus, we may claim that attribute-concept values are better than lost values in terms of rule complexity.

Our secondary objective was to check which approximation (lower, middle or upper) is the best from the point of view of rule complexity.

The smallest size of rule sets was accomplished, in five (out of 24 combinations) for lower approximations and in two combinations for upper approximations. The total number of conditions in rule sets was achieved, again, for lower approximations in five combinations and for upper approximations in other two combinations. For remaining 17 combinations the difference between all three approximations was insignificant.

This paper starts with a discussion on incomplete data in Section II where we define approximations, attribute-value blocks and characteristic sets. In Section III, we present probabilistic approximations for incomplete data. 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 = $\{$*Education*, *Skills*, *Experience*$\}$. The value for a case x and an attribute a will be denoted by $a(x)$.

 ISBN: 978-1-61208-352-0

TABLE I. A DECISION TABLE

	Attributes			Decision
Case	Education	Skills	Experience	Productivity
1	higher	high	–	high
2	?	high	low	high
3	secondary	–	high	high
4	higher	?	high	high
5	elementary	high	low	low
6	secondary	–	high	low
7	–	low	high	low
8	elementary	?	–	low

In this paper, we distinguish between two interpretations of missing attribute values: lost values and attribute-concept values. Lost values, denoted by "?", mean that the original attribute value is no longer accessible and that during rule induction we will only use existing attribute values [11][12]. Attribute-concept values, denoted by "–", mean that the original attribute value is unknown; however, because we know the concept to which a case belongs, we know all possible attribute values. Table I presents an incomplete data set affected by both lost values and attribute-concept values.

One of the most important ideas of rough set theory [13] 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 defined in equation 1.

$$(x, y) \in R(B) \text{ if and only if } \forall a \in B \ (a(x) = a(y)) \quad (1)$$

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*. For example, a concept associated with the value *low* of the decision *Productivity* is the set $\{1, 2, 3, 4\}$. The largest B-definable set contained in X is called the *B-lower approximation* of X, denoted by $\underline{appr}_B(X)$, and defined in equation 2.

$$\cup\{[x]_B \mid [x]_B \subseteq X\} \quad (2)$$

The smallest B-definable set containing X, denoted by $\overline{appr}_B(X)$ is called the *B-upper approximation* of X, and is defined in equation 3.

$$\cup\{[x]_B \mid [x]_B \cap X \neq \emptyset\} \quad (3)$$

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 $\{x \in U \mid a(x) = v\}$ [14]. For incomplete decision tables the definition of a block of an attribute-value pair is modified in the following way.

- If for an attribute a there exists a case x such that $a(x) = ?$, i.e., the corresponding value is lost, then the case x should not be included in any blocks $[(a, v)]$ for all values v of attribute a,
- 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, and is defined by equation 4.

$$V(x, a) = \{a(y) \mid a(y) \textit{ is specified}, \ y \in U, \ d(y) = d(x)\} \quad (4)$$

For the data set from Table I, the attribute-concept values are defined as: $V(1, Experience) = \{low, high\}$, $V(3, Skills) = \{high\}$, $V(6, Skills) = \{low, high\}$, $V(7, Education) = \{elementary, secondary\}$ and $V(8, Experience) = \{low, high\}$.

For the data set from Table I the blocks of attribute-value pairs are: [(Education, elementary)] = $\{5, 7, 8\}$, [(Education, secondary)] = $\{3, 6, 7\}$, [(Education, higher)] = $\{1, 4\}$, [(Skills, low)] = $\{6, 7\}$, [(Skills, high)] = $\{1, 2, 3, 5, 6\}$, [(Experience, low)] = $\{1, 2, 5, 8\}$, and [(Experience, high)] = $\{1, 3, 4, 6, 7, 8\}$.

For a case $x \in U$ and $B \subseteq A$, the *characteristic set* $K_B(x)$ is defined as the intersection of the sets $K(x, a)$, for all $a \in B$, where the set $K(x, a)$ is defined in the following way:

- If $a(x)$ is specified, then $K(x, a)$ is the block $[(a, a(x))]$ of attribute a and its value $a(x)$,
- If $a(x) =?$ then the set $K(x, a) = U$, where U is the set of all cases,
- 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, $K(x, a) = U$.

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

Note that for incomplete data there are a few possible ways to define approximations [7], we used *concept* approximations [9] since our previous experiments indicated that such approximations are most efficient [9]. A B-*concept lower approximation* of the concept X is defined in equation 5.

$$\underline{B}X = \cup\{K_B(x) \mid x \in X, K_B(x) \subseteq X\} \quad (5)$$

The *B-concept upper approximation* of the concept X is defined by the equation 6.

$$\begin{aligned} \overline{B}X &= \cup\{K_B(x) \mid x \in X, K_B(x) \cap X \neq \emptyset\} \\ &= \cup\{K_B(x) \mid x \in X\} \end{aligned} \quad (6)$$

For Table I, A-concept lower and A-concept upper approximations of the concept $\{1, 2, 3, 4\}$ are $\underline{A}\{1, 2, 3, 4\} = \{1, 4\}$ and $\overline{A}\{1, 2, 3, 4\} = \{1, 2, 3, 4, 5, 6\}$, respectively.

III. Probabilistic Approximations

For completely specified data sets a *probabilistic approximation* is defined by equation 7, where α is a parameter, $0 < \alpha \leq 1$, see [1][4][9][15]–[17]. Additionally, for simplicity, the elementary sets $[x]_A$ are denoted by $[x]$. For discussion on how this definition is related to the value precision asymmetric rough sets see [9][10].

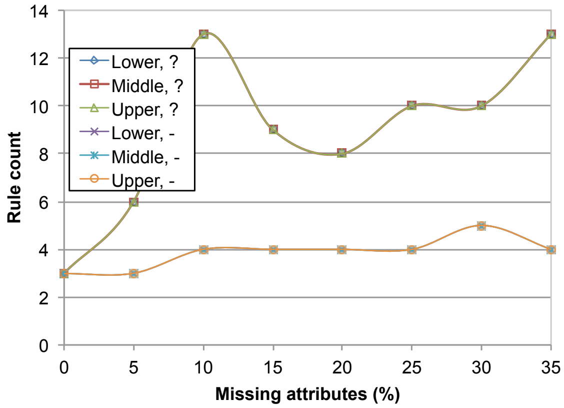

Figure 1. Size of the rule set for the *Bankruptcy* data set

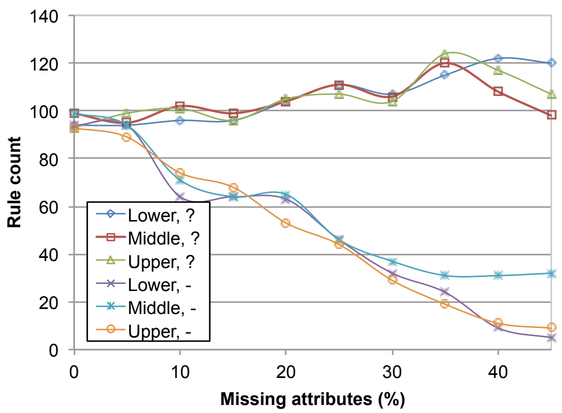

Figure 2. Size of the rule set for the *Breast cancer* data set

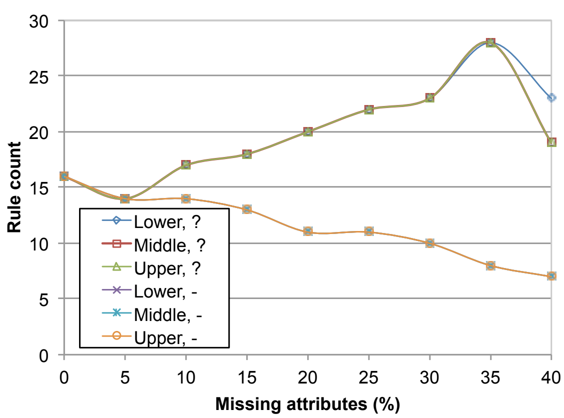

Figure 3. Size of the rule set for the *Echocardiogram* data set

$$appr_{\alpha}(X) = \cup\{[x] \mid x \in U, P(X \mid [x]) \geq \alpha\} \qquad (7)$$

Note that if $\alpha = 1$, the probabilistic approximation becomes the standard lower approximation and if α is small, close to 0, in our experiments it was 0.001, the same definition describes the standard upper approximation.

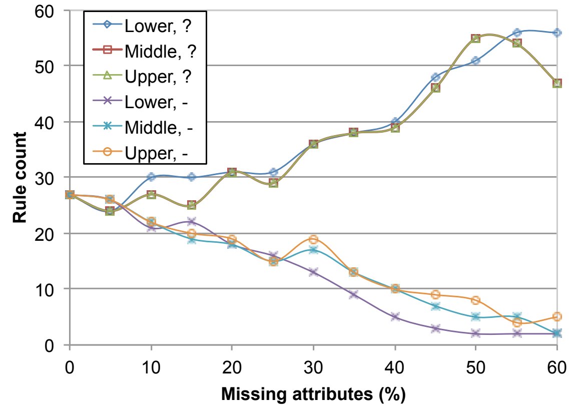

Figure 4. Size of the rule set for the *Hepatitis* data set

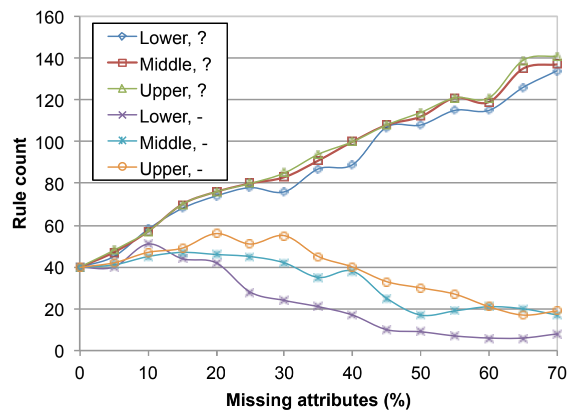

Figure 5. Size of the rule set for the *Image segmentation* data set

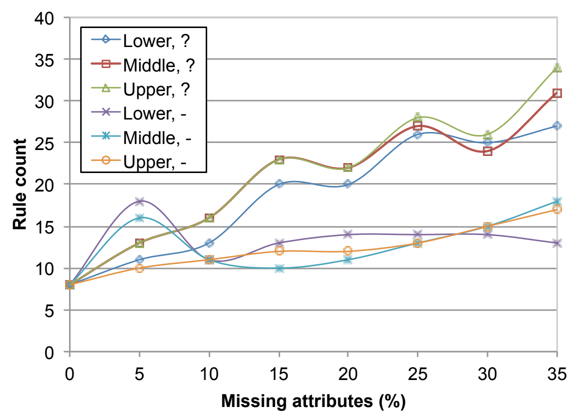

Figure 6. Size of the rule set for the *Iris* data set

For incomplete data sets, a *B-concept probabilistic approximation* is defined by equation 8 [9].

$$\cup\{K_B(x) \mid x \in X, \ Pr(X|K_B(x)) \geq \alpha\} \qquad (8)$$

For simplicity, we will denote $K_A(x)$ by $K(x)$ and the *A-concept probabilistic approximation* will be called a *probabilistic approximation*.

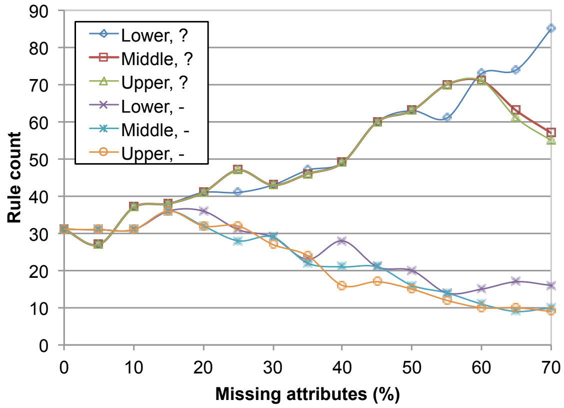

Figure 7. Size of the rule set for the *Lymphography* data set

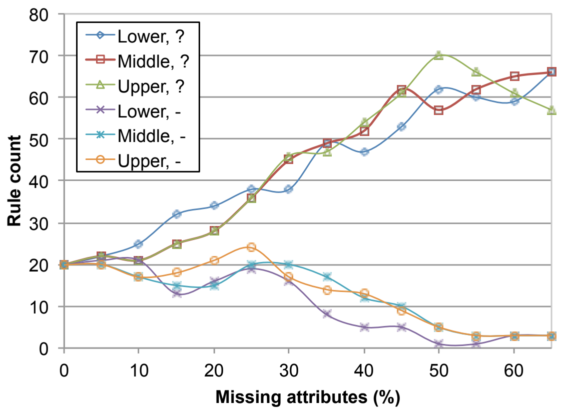

Figure 8. Size of the rule set for the *Wine recognition* data set

The special probabilistic approximations with the parameter $\alpha = 0.5$ will be called a *middle* approximation.

IV. Experiments

Our experiments are based on eight data sets that are available on the University of California at Irvine *Machine Learning Repository*.

For every data set a set of templates was created. Templates were formed by replacing incrementally (with 5% increment) existing specified attribute values by *lost values*. Thus, we started each series of experiments with no *lost values*, then we added 5% of *lost values*, then we added additional 5% of *lost values*, etc., until at least one entire row of the data sets was full of *lost values*. Then three attempts were made to change configuration of new *lost values* and either a new data set with extra 5% of *lost values* was created or the process was terminated. Additionally, the same formed templates were edited for further experiments by replacing question marks, representing *lost values* by"−"s representing *attribute-concept values*.

For any data set there was some maximum for the percentage of missing attribute values. For example, for the *bankruptcy* data set, it was 35%. Hence, for the *bankruptcy* data set, we created seven data sets with lost values and seven data sets with attribute-concept values, for the total of 15 data sets (the additional data set was complete, with no missing attribute values). By the same token, for the *breast cancer*, *echocardiogram*, *hepatitis*, *image segmentation*, *iris*, *lymphography* and *wine recognition* data sets we created 19, 17, 25, 29, 15, 29, and 27 data sets. The total number of the data sets was 176.

Results of our experiments are presented in Figures 1–16.

We compared two interpretations of missing attribute values, lost values and attribute-concept values, assuming the same type of approximations. More explicitly, we compared the complexity of rule sets, first the size of rule sets, then the total number of conditions in the rule set, separately for lower approximations, then for middle approximations, and finally, for upper approximations, using the Wilcoxon matched-pairs signed rank test, with the 5% level of significance for two-tailed test.

For all eight types of data sets and all three types of approximations, the rule set size was always smaller for attribute-concept values than for lost values. For the total number of conditions in the rule sets results were more complicated. The total number of conditions in the rule sets was smaller for attribute-concept values than for lost values for 17 combinations of the type of data set and approximation, out of 24 possible combinations. For *echocardiogram* and *iris* data sets, for all three types of approximations and for the *lymphography* data set and lower approximations, the total number of conditions in rule sets for both interpretations of missing attribute values, did not differ significantly.

We compared all three types of approximations as well, assuming the same interpretation of missing attribute values, in terms of the size of rule sets and the total number of conditions in rule sets, using the Friedman Rank Sums test, again, with 5% of significance level.

The size of the rule set was smaller for lower approximations than for upper approximations for three combinations of the type of data set and type of missing attribute values (for the *hepatitis* data set and attribute-concept values and for the *image segmentation* data set and both lost values and attribute-concept values). The size of the rule set was smaller for lower approximations than for middle approximations in two combinations of the type of data set and type of missing attribute value (for the *image segmentation* data set and both lost values and attribute-concept values). Thus, for five combinations (out of 24) lower approximations were better than other approximations. On the other hand, the size of the rule set was smaller for upper approximations than for lower approximations for one combination (for the *lymphography* data set and attribute-concept values). Additionally, the size of the rule set was smaller for upper approximations than for middle approximations also for one combination (for the *breast cancer* data set and the attribute-concept values). Thus, for two combinations (out of 24) upper approximations were better than other approximations. For remaining 17 combinations the difference between all three approximations was insignificant.

The total number of conditions in rule sets was smaller for lower approximations than for upper approximations in four combinations of the type of data set and type of missing attribute value (for the *hepatitis* data set and attribute-concept

 ISBN: 978-1-61208-352-0

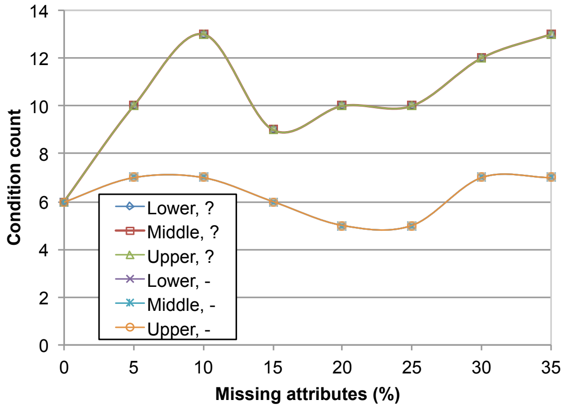

Figure 9. Number of conditions for the *Bankruptcy* data set

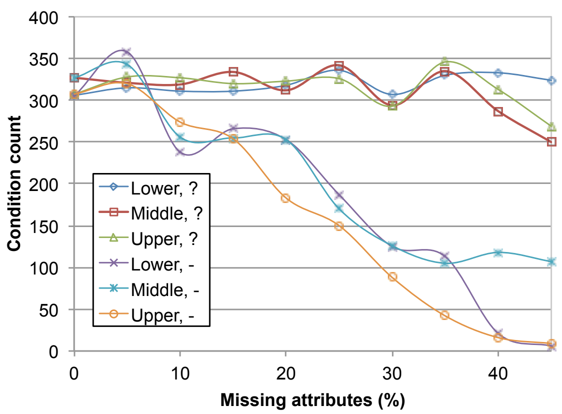

Figure 10. Number of conditions for the *Breast cancer* data set

values and for the *image segmentation* data set and both lost values and attribute-concept values and for the *iris* data set and lost values). The total number of conditions in rule sets was smaller for lower approximations than for middle approximations in one combination (for the *image segmentation* data set and lost values). Thus, for five combinations (out of 24) lower approximations were better than other approximations. The total number of conditions in rule sets was smaller for middle approximations than for lower approximations for one combination (for the *lymphography* data set and the attribute-concept values). Additionally, the total number of conditions in rules sets was smaller for upper approximations than for lower approximations also for one combination (for the *lymphography* data set and the attribute-concept values). Thus, for two combinations (out of 24) other approximations were better than lower approximations. For remaining 17 combinations the difference between all three approximations was insignificant.

In our experiments, we used the MLEM2 rule induction algorithm of the Learning from Examples using Rough Sets (LERS) data mining system [10][18][19].

V. Conclusions

As follows from our experiments, the size of rule set was always smaller for attribute-concept values than for lost values. The total number of conditions in rule sets was smaller for

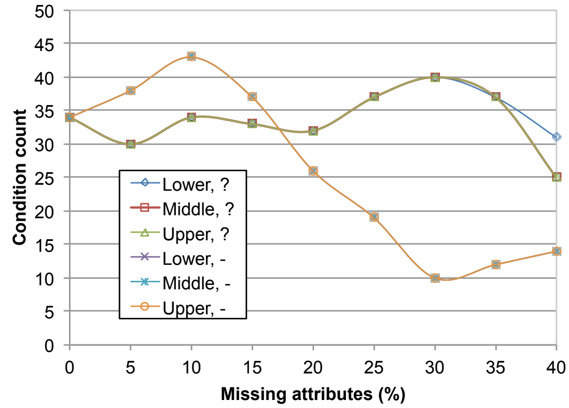

Figure 11. Number of conditions for the *Echocardiogram* data set

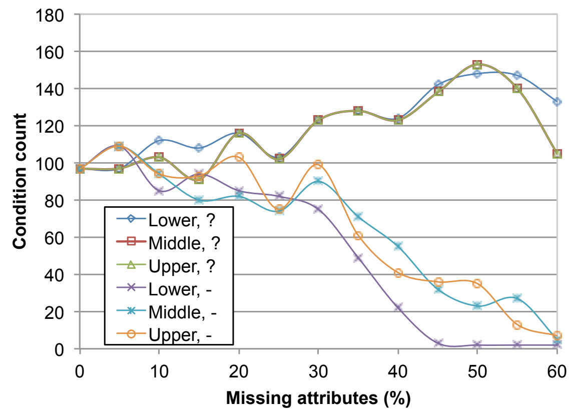

Figure 12. Number of conditions for the *Hepatitis* data set

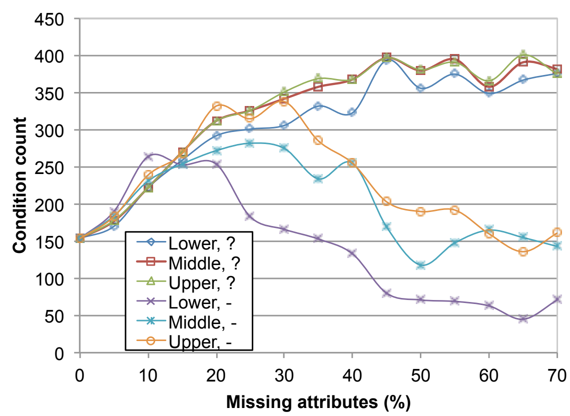

Figure 13. Number of conditions for the *Image segmentation* data set

attribute-concept values for 17 combinations of the type of data set and approximation (out of 24 combinations total). In remaining seven combinations, the total number of conditions in rule sets did not differ significantly. Thus, we may claim attribute-concept values are better than lost values in terms of rule complexity.

The smallest size of rule sets was accomplished, in five

 ISBN: 978-1-61208-352-0

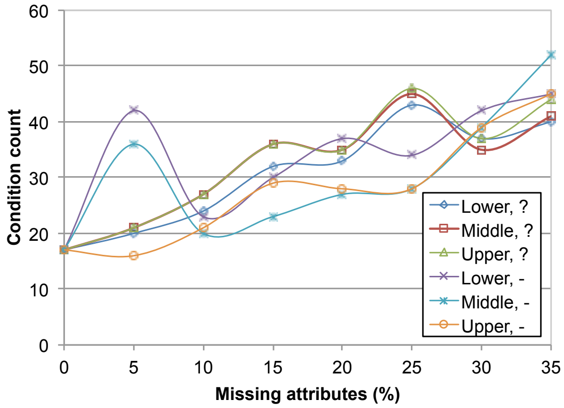

Figure 14. Number of conditions for the *Iris* data set

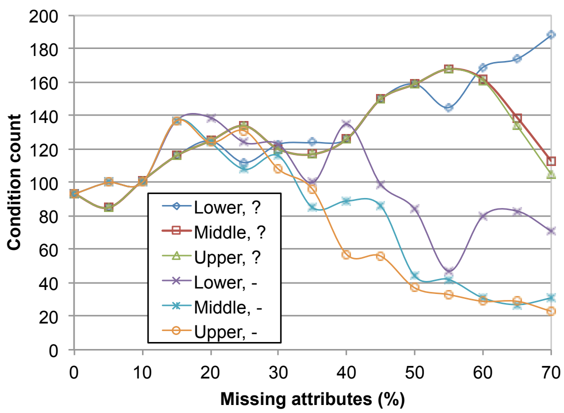

Figure 15. Number of conditions for the *Lymphography* data set

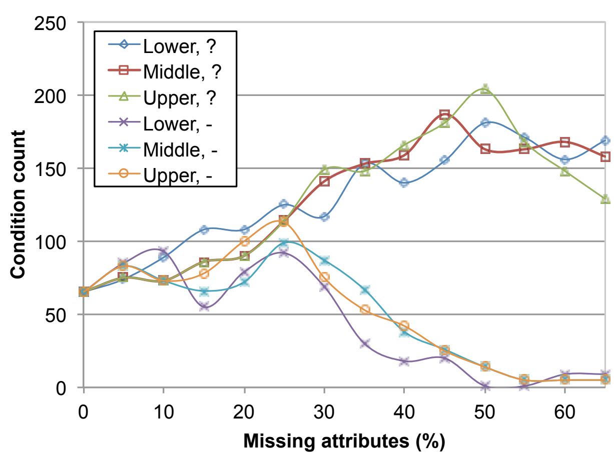

Figure 16. Number of conditions for the *Wine recognition* data set

(out of 24 combinations for lower approximations and in two combinations for upper approximations. The total number of conditions in rule sets was achieved, again, for lower approximations in five combinations and for upper or middle approximations in other two combinations. For remaining 17 combinations the difference between all three approximations was insignificant.

REFERENCES

[1] 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.

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

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

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

[5] 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.

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

[7] J. W. Grzymala-Busse, "Rough set strategies to data with missing attribute values," in Workshop Notes, Foundations and New Directions of Data Mining, in conjunction with the 3-rd International Conference on Data Mining, 2003, pp. 56–63.

[8] ——, "Data with missing attribute values: Generalization of indiscernibility relation and rule induction," Transactions on Rough Sets, vol. 1, 2004, pp. 78–95.

[9] ——, "Generalized parameterized approximations," in Proceedings of the RSKT 2011, the 6-th International Conference on Rough Sets and Knowledge Technology, 2011, pp. 136–145.

[10] P. G. Clark and J. W. Grzymala-Busse, "Experiments on probabilistic approximations," in Proceedings of the 2011 IEEE International Conference on Granular Computing, 2011, pp. 144–149.

[11] J. W. Grzymala-Busse and A. Y. Wang, "Modified algorithms LEM1 and LEM2 for rule induction from data with missing attribute values," in Proceedings of the Fifth International Workshop on Rough Sets and Soft Computing (RSSC'97) at the Third Joint Conference on Information Sciences (JCIS'97), 1997, pp. 69–72.

[12] J. Stefanowski and A. Tsoukias, "Incomplete information tables and rough classification," Computational Intelligence, vol. 17, no. 3, 2001, pp. 545–566.

[13] Z. Pawlak, "Rough sets," International Journal of Computer and Information Sciences, vol. 11, 1982, pp. 341–356.

[14] 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.

[15] 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.

[16] 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.

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

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

[19] ——, "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.