On the Number of Rules and Conditions in Mining 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
Department of Expert Systems and Artificial Intelligence
University of Information Technology and Management
Rzeszow, Poland
e-mail: jerzy@ku.edu

***Abstract*—In mining incomplete data, we have a choice for interpretation of missing attribute values. In this paper, we consider two such interpretations: lost values and attribute-concept values. To measure the number of conditions and rules for each interpretation, we conducted experiments on eight incomplete data sets using three kinds of probabilistic approximations: singleton, subset and concept, with eleven values of probability. Using a 5% significance level, the results show that the number of rules is always smaller for attribute-concept values than for lost values. Additionally, the total number of conditions is smaller for attribute-concept values than for lost values for seven out of eight data sets.**

***Index Terms*—Data mining; rough set theory; probabilistic approximations; MLEM2 rule induction algorithm; lost values; attribute-concept values.**

I. INTRODUCTION

In this paper, we use two interpretations of a missing attribute value: lost values and attribute-concept values. Lost values indicate that the original values were erased, and as a result we should use only existing, specified attribute values for rule induction. Attribute-concept values mean that the missing attribute value may be replaced by any specified attribute value, typically for a given concept.

The idea of complexity of rule sets induced from incomplete data sets with lost values and attribute-concept values was introduced in [1]. In [1], experiments were conducted using only one type of probabilistic approximation (concept) and only three probabilities used for probabilistic approximations (0.001, 0.5 and 1.0). In this paper we use three kinds of probabilistic approximations (singleton, subset and concept) and eleven values of probability, starting from 0.001 and then from 0.1 to 1.0 with an increment of 0.1.

Our previous research [2][3] shows that the quality of rule sets, evaluated by an error rate computed by ten-fold cross validated, does not differ significantly with different combinations of missing attribute and probabilistic approximation type. As a result the main objective of this paper is research on complexity of rule sets, in terms of the number of rules and total number of rule conditions, induced from data sets with lost values and attribute-concept values. In previous work and in this work, the Modified Learning from Examples Module, version 2 (MLEM2) was used for rule induction.

The main results of this paper show that 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 22 of 24 combinations of the type of data set and probabilistic approximation. In the remaining two 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. While our previous work discussed a subset of these findings, the primary contribution of this paper is that the results are more extensive and decisive than those presented in [1].

Our secondary objective was to check which probabilistic approximation (singleton, subset or concept) is the best from the point of view of rule complexity. 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 singleton, subset and concept probabilistic approximations for incomplete data. Section IV contains the details of our experiments. Finally, conclusions are presented in Section V.

II. INCOMPLETE DATA

Fundamental concepts of rough set theory are standard lower and upper approximations. A probabilistic approximation, associated with a probability α, is an extension of the standard approximation. For α = 1, the probabilistic 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 was initiated in [4] and then was continued in, e.g., [5]–[9].

Incomplete data sets are analyzed using special approximations such as singleton, subset and concept [10][11]. Probabilistic approximations, for incomplete data sets and based on an arbitrary binary relation, were introduced in [12], while first experimental results using probabilistic approximations were published in [13]. In experiments reported in this paper, we used three kinds of probabilistic approximations: singleton, subset and concept.

We assume that the input data sets are presented in the form of a decision table. An example of a decision table

TABLE I
A DECISION TABLE

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

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 attribute-concept values, denoted by "–" [14][15]. 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 [16] 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 *yes* of the decision *Trip* 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 $\{x \in U \mid a(x) = v\}$ [17]. 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, where $V(x, a)$ 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, we have $V(2, Wind) = \{low, high\}$, $V(3, Humidity) = \{low\}$, $V(4, Wind) = \{low, high\}$, $V(6, Temperature) = \{low\}$ and $V(8, Wind) = \{low, high\}$.

For the data set from Table I the blocks of attribute-value pairs are:
[(Wind, low)] = {2, 3, 4, 6, 8},
[(Wind, high)] = {1, 2, 4, 5, 7, 8},
[(Humidity, low)] = {1, 2, 3, 4, 8},
[(Humidity, high)] = {5, 7},
[(Temperature, low)] = {4, 6, 7, 8}, and
[(Temperature, high)] = {1, 3}.

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, 3, 4, 8},
$K_A(3)$ = {3},
$K_A(4)$ = {4, 8},
$K_A(5)$ = {5, 7},
$K_A(6)$ = {4, 6, 8},
$K_A(7)$ = {7}, and
$K_A(8)$ = {4, 8}.

First we will quote some definitions from [18]. Let X be a subset of U. The B-*singleton lower approximation* of X, denoted by $\underline{appr}_B^{singleton}(X)$, is defined by equation 5

$$\{x \mid x \in U, K_B(x) \subseteq X\} \quad (5)$$

The B-*singleton upper approximation* of X, denoted by $\overline{appr}_B^{singleton}(X)$, is defined by equation 6.

$$\{x \mid x \in U, K_B(x) \cap X \neq \emptyset\} \quad (6)$$

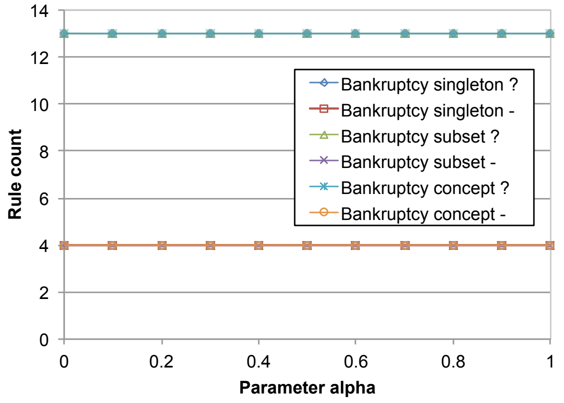

Fig. 1. Size of the rule set for the *Bankruptcy* data set

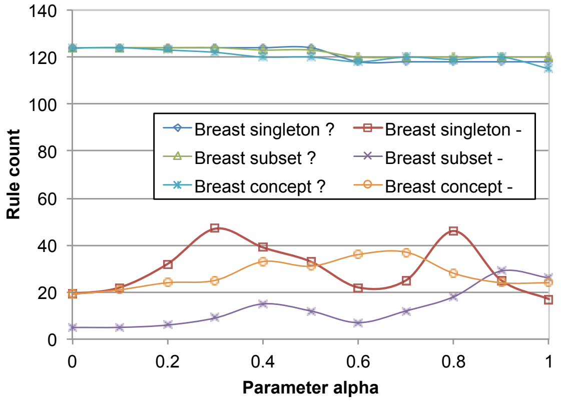

Fig. 2. Size of the rule set for the *Breast cancer* data set

The B-*subset lower approximation* of X, denoted by $\underline{appr}_B^{subset}(X)$, is defined by equation 7.

$$\cup \{K_B(x) \mid x \in U, K_B(x) \subseteq X\} \quad (7)$$

The B-*subset upper approximation* of X, denoted by $\overline{appr}_B^{subset}(X)$, is defined by equation 8.

$$\cup \{K_B(x) \mid x \in U, K_B(x) \cap X \neq \emptyset\} \quad (8)$$

The B-*concept lower approximation* of X, denoted by $\underline{appr}_B^{concept}(X)$, is defined by equation 9.

$$\cup \{K_B(x) \mid x \in X, K_B(x) \subseteq X\} \quad (9)$$

The B-*concept upper approximation* of X, denoted by $\overline{appr}_B^{concept}(X)$, is defined by equation 10.

$$\cup \{K_B(x) \mid x \in X, K_B(x) \cap X \neq \emptyset\} = \\ \cup \{K_B(x) \mid x \in X\} \quad (10)$$

For Table I and X = {5, 6, 7, 8}, all A-singleton, A-subset and A-concept lower and upper approximations are:

$\underline{appr}_A^{singleton}(X) = \{5, 7\}$,
$\overline{appr}_A^{singleton}(X) = \{2, 4, 5, 6, 7, 8\}$,
$\underline{appr}_A^{subset}(X) = \{5, 7\}$,
$\overline{appr}_A^{subset}(X) = U$,
$\underline{appr}_A^{concept}(X) = \{5, 7\}$,
$\overline{appr}_A^{concept}(X) = \{4, 5, 6, 7, 8\}$.

III. Probabilistic approximations

In this section we will extend definitions of singleton, subset and concept approximations to corresponding probabilistic approximations. A B-singleton probabilistic approximation of X with the threshold α, $0 < \alpha \leq 1$, denoted by $appr_{\alpha,B}^{singleton}(X)$, is defined by equation 11.

$$\{x \mid x \in U,\ Pr(X \mid K_B(x)) \geq \alpha\} \quad (11)$$

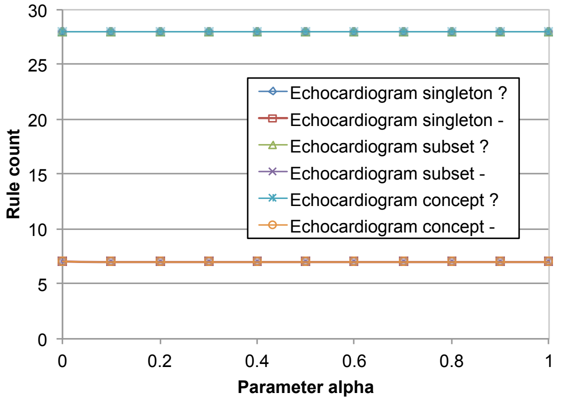

Fig. 3. Size of the rule set for the *Echocardiogram* data set

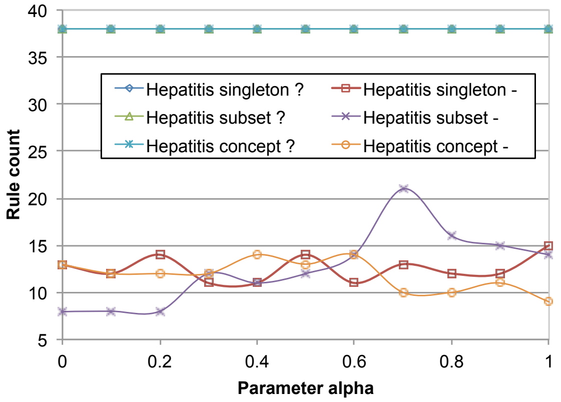

Fig. 4. Size of the rule set for the *Hepatitis* data set

Here $Pr(X \mid K_B(x)) = \frac{|X \cap K_B(x)|}{|K_B(x)|}$ is the conditional probability of X given $K_B(x)$ and $|Y|$ denotes the cardinality of set Y. 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 by equation 12.

$$\cup\{K_B(x) \mid x \in U,\ Pr(X \mid K_B(x)) \geq \alpha\} \quad (12)$$

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

ISBN: 978-1-61208-408-4

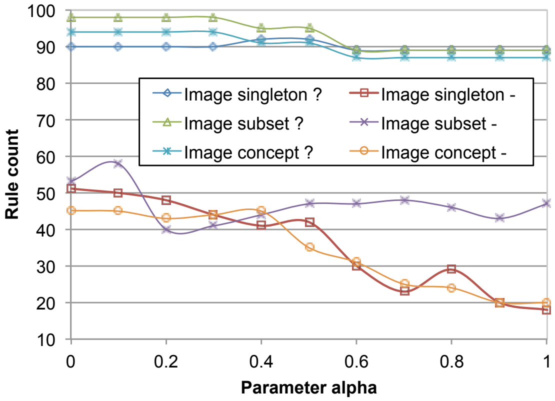

Fig. 5. Size of the rule set for the *Image segmentation* data set

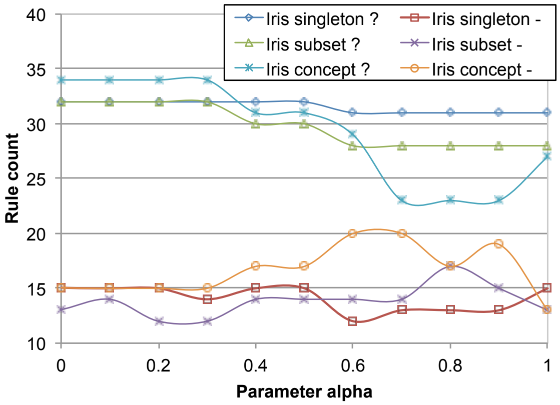

Fig. 6. Size of the rule set for the *Iris* data set

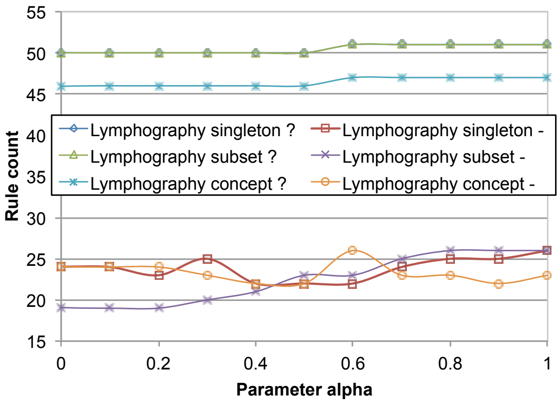

Fig. 7. Size of the rule set for the *Lymphography* data set

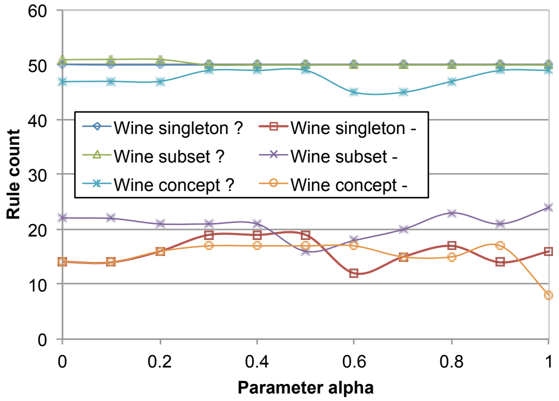

Fig. 8. Size of the rule set for the *Wine recognition* data set

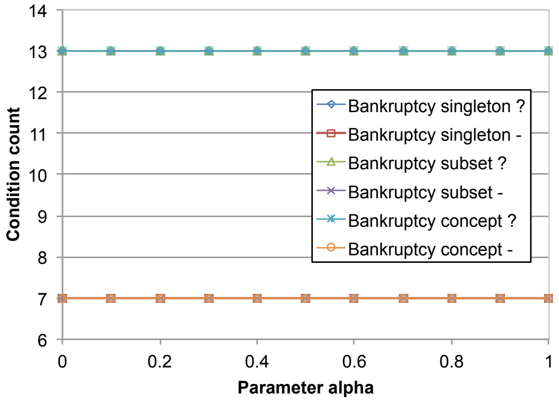

Fig. 9. Number of conditions for the *Bankruptcy* data set

defined by equation 13.

$$\cup\{K_B(x) \mid x \in X,\ Pr(X \mid K_B(x)) \geq \alpha\} \qquad (13)$$

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.

For Table I and the concept X = [(*Trip*, *yes*)] = {1, 2, 3, 4}, there exist the following distinct probabilistic approximations:

$appr^{singleton}_{1.0,A}(\{1,2,3,4\}) = \{1,3\}$,
$appr^{singleton}_{0.8,A}(\{1,2,3,4\}) = \{1,2,3\}$,
$appr^{singleton}_{0.5,A}(\{1,2,3,4\}) = \{1,2,3,4,8\}$,
$appr^{singleton}_{0.333,A}(\{1,2,3,4\}) = \{1,2,3,4,6,8\}$,
$appr^{subset}_{1.0,A}(\{1,2,3,4\}) = \{1,3\}$,
$appr^{subset}_{0.5,A}(\{1,2,3,4\}) = \{1,2,3,4,8\}$,
$appr^{subset}_{0.333,A}(\{1,2,3,4\}) = \{1,2,3,4,6,8\}$,
$appr^{concept}_{1.0,A}(\{1,2,3,4\}) = \{1,3\}$,
$appr^{concept}_{0.333,A}(\{1,2,3,4\}) = \{1,2,3,4,8\}$,

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 template was created. Such a template was formed by replacing randomly 35% of existing specified attribute values by *lost values*. The same template was used for constructing a corresponding data set with *attribute-concept values*, by replacing "?"s with "−"s.

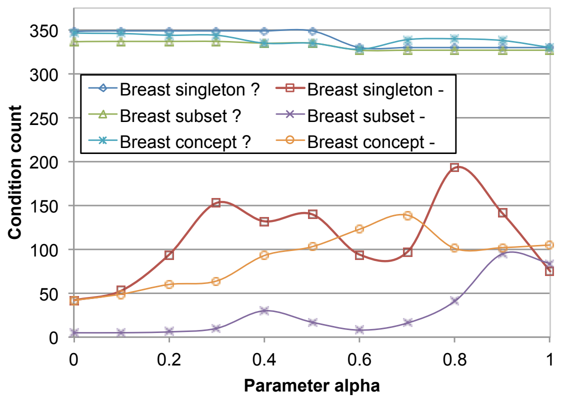

Fig. 10. Number of conditions for the *Breast cancer* data set

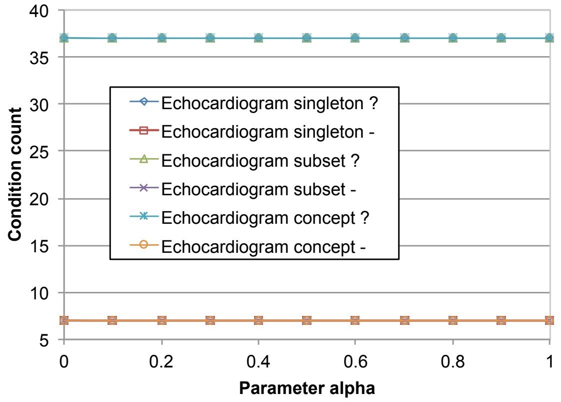

Fig. 11. Number of conditions for the *Echocardiogram* data set

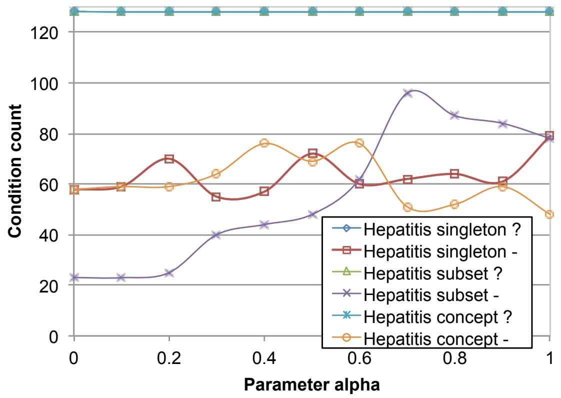

Fig. 12. Number of conditions for the *Hepatitis* data set

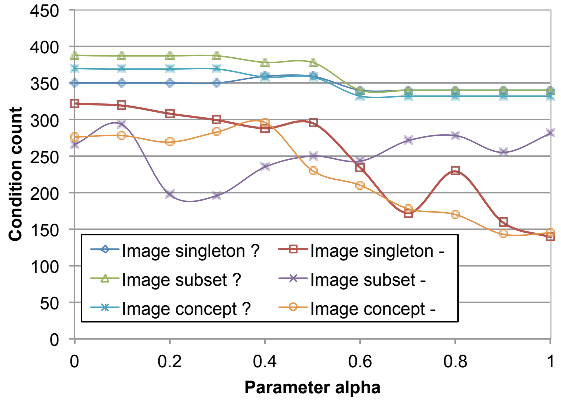

Fig. 13. Number of conditions for the *Image segmentation* data set

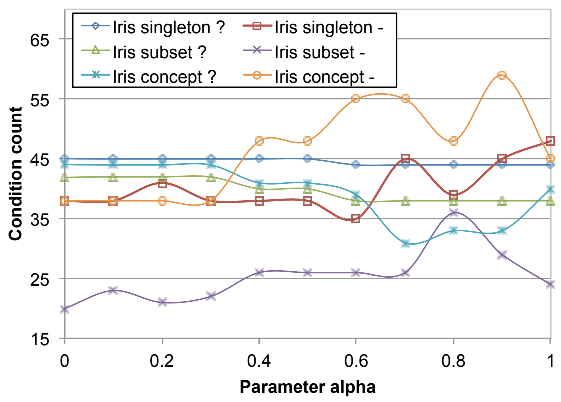

Fig. 14. Number of conditions for the *Iris* data set

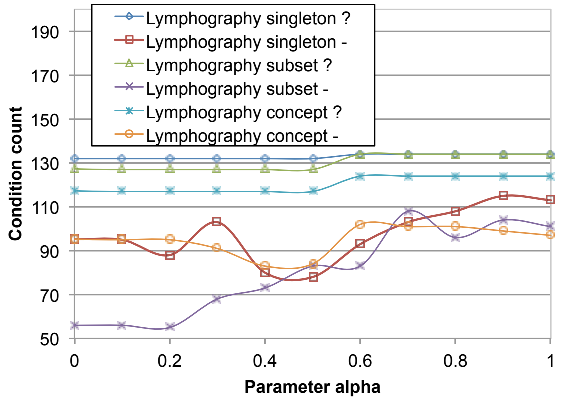

Fig. 15. Number of conditions for the *Lymphography* data set

For any data set we compared the size of rule set and the total number of conditions in the rule set for two interpretations of missing attribute values assuming the same type of probabilistic approximation. We used the Wilcoxon matched-pairs signed-ranked test with 5% significance level and with Bonferroni correction for multiple comparisons. In our experiments, we used the MLEM2 rule induction algorithm of the Learning from Examples using Rough Sets (LERS) data mining system [13][19][20]. Results of our experiments are presented in Figures 1–16.

For rule sets, in all 24 combinations of the type of probabilistic approximation and data set, the size of the rule set

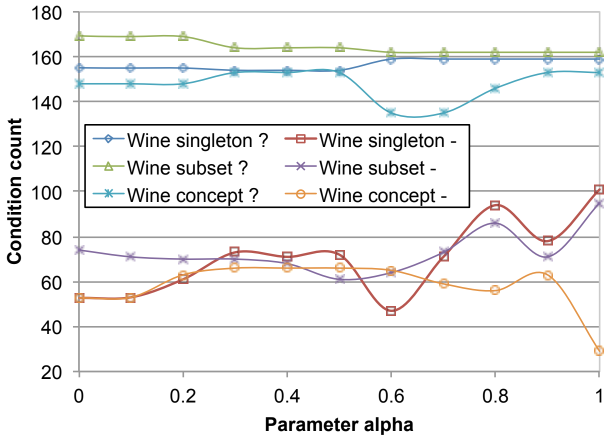

Fig. 16. Number of conditions for the *Wine recognition* data set

was always smaller for attribute-concept values than for lost values. For the total number of conditions, for the *iris* data set and for singleton and concept probabilistic approximations (Figure 6), the results were statistically inconclusive. For remaining 22 combinations, the total number of conditions was always smaller for attribute-concept values than for lost values.

In six out of the eight data sets, results of our experiments show some level of variability in the number of rules and conditions as the parameter alpha changes. However, for the *bankruptcy* and *echocardiogram* data sets, within given interpretation of missing attribute values, these numbers are constant for all values of the parameter α. The results for the *bankruptcy* and *echocardiogram* are different from the results for the remaining six data sets since for the former all attributes have numeric values while for the latter attributes are symbolic. For numeric attributes, during rule induction, the same numeric intervals are created for all possible values of the parameter α.

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 attribute-concept values for 22 combinations of the type of data set and probabilistic approximation (out of 24 combinations total). In remaining two combinations, the total number of conditions in rule sets did not differ significantly. Thus, we claim that attribute-concept values are better than lost values in terms of rule complexity.

Additionally, results of our experiments show that in induction of least complex rule sets the difference between all three probabilistic approximations (singleton, subset and concept) was not statistically significant.

References

[1] P. G. Clark and J. W. Grzymala-Busse, “Complexity of rule sets induced from incomplete data with lost values and attribute-concept values,” in *Proceedings of the Third International Conference on Intelligent Systems and Applications*, 2014, pp. 91–96.

[2] P. G. Clark, J. W. Grzymala-Busse, and W. Rzasa, “Mining incomplete data with singleton, subset and concept approximations,” *Information Sciences*, vol. 280, pp. 368–384, 2014.

[3] P. G. Clark and J. W. Grzymala-Busse, “Mining incomplete data with lost values and attribute-concept values,” in *Proceedings of the 2014 IEEE International Conference on Granular Computing*, 2014, pp. 49–54.

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

[5] Z. Pawlak and A. Skowron, “Rough sets: Some extensions,” *Information Sciences*, vol. 177, pp. 28–40, 2007.

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

[7] Y. Y. Yao, “Probabilistic rough set approximations,” *International Journal of Approximate Reasoning*, vol. 49, pp. 255–271, 2008.

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

[9] W. Ziarko, “Probabilistic approach to rough sets,” *International Journal of Approximate Reasoning*, vol. 49, pp. 272–284, 2008.

[10] J. W. Grzymala-Busse, “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.

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

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

[13] 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.

[14] 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 5-th International Workshop on Rough Sets and Soft Computing in conjunction with the Third Joint Conference on Information Sciences*, 1997, pp. 69–72.

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

[16] Z. Pawlak, “Rough sets,” *International Journal of Computer and Information Sciences*, vol. 11, pp. 341–356, 1982.

[17] 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.

[18] J. W. Grzymala-Busse and W. Rzasa, “Definability and other properties of approximations for generalized indiscernibility relations,” *Transactions on Rough Sets*, vol. 11, pp. 14–39, 2010.

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

[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: 978-1-61208-408-4