Complexity of Rule Sets Induced from Incomplete Data Sets Using Global Probabilistic Approximations

Patrick G. Clark[1] and Jerzy W. Grzymala-Busse[1,2]

[1] Department of Electrical Engineering and Computer Science,
University of Kansas, Lawrence, KS 66045, USA
[2] Department of Expert Systems and Artificial Intelligence,
University of Information Technology and Management,
35-225 Rzeszow, Poland
patrick.g.clark@gmail.com, jerzy@ku.edu

Abstract. We consider incomplete data sets using two interpretations of missing attribute values: lost values and "do not care" conditions. Additionally, in our data mining experiments we use global probabilistic approximations (singleton, subset and concept). The results of validation of such data, using global probabilistic approximations, were published recently. A novelty of this paper is research on the complexity of corresponding rule sets, in terms of the number of rules and number of rule conditions. Our main result is that the simplest rule sets are induced from data sets in which missing attribute values are interpreted as "do not care" conditions where rule sets are induced using subset probabilistic approximations.

1 Introduction

Probabilistic approximations, for complete data sets and based on an equivalence relation, were studied for many years [14–19]. Incomplete data sets may be analyzed using global approximations such as singleton, subset and concept [5–7]. Probabilistic approximations, for incomplete data sets and based on arbitrary binary relations, were introduced in [8], while first experimental results using probabilistic approximations were published in [1].

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 *"do not care" conditions*. Lost values indicate the original value was erased or never obtained, and as a result we should use only existing, specified attribute values for rule induction. "Do not care" conditions identify data that may be replaced by any specified attribute value, typically someone refused to answer a question.

A probabilistic approximation is defined using a probability α. If α is equal to one, the probabilistic approximation is equal to the lower approximation; if α is a sufficiently small, positive number, the probabilistic approximation is equal to the upper approximation. Both lower and upper approximations are fundamental ideas of rough set theory.

A. Laurent et al. (Eds.): IPMU 2014, Part I, CCIS 442, pp. 386–395, 2014.

The main objective of this paper is research on the complexity of rule sets, in terms of the number of rules and number of rule conditions, induced from data sets with lost values and "do not care" conditions, while rule sets are induced using three global approximations: singleton, subset and concept. These approximations and their relationship to probabilistic approximations are defined in section 3. Our main result is that the simplest rule sets are induced from data sets in which missing attribute values are interpreted as "do not care" conditions where rule sets are induced using subset probabilistic approximations.

2 Attribute-Value Pair Blocks

We assume that the input data sets are presented in the form of a *decision table*. Example of decision tables are shown in Tables 1 and 2. Rows of the decision table represent *cases*, while columns are labeled by *variables*. The set of all cases will be denoted by U. In Tables 1 and 2, $U = \{1, 2, 3, 4, 5, 6, 7, 8\}$. Some variables are called *attributes* while one selected variable is called a *decision* and is denoted by d. The set of all attributes will be denoted by A. In Tables 1 and 2, $A = \{$*Wind*, *Humidity*, *Temperature*$\}$ and $d =$ *Trip*.

An important tool to analyze data sets is a *block of an attribute-value pair*. Let (a, v) be an attribute-value pair. For *complete* decision tables, i.e., decision tables in which every attribute value is specified, a block of (a, v), denoted by $[(a, v)]$, is the set of all cases x for which $a(x) = v$, where $a(x)$ denotes the value of the attribute a for the case x. For incomplete decision tables the definition of a block of an attribute-value pair is modified [5–7].

– 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,

Table 1. An incomplete decision table with lost values

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

– If for an attribute a there exists a case x such that the corresponding value is a "do not care" condition, i.e., $a(x) = *$, then the case x should be included in blocks $[(a, v)]$ for all specified values v of attribute a.

Table 2. An incomplete decision table with "do not care" conditions

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

Table 3. Blocks $[(a,\ v)]$ of attribute value pairs $(a,\ v)$

	Lost values	"Do not care" conditions
[(Wind, low)]	$\{1, 6\}$	$\{1, 2, 4, 5, 6\}$
[(Wind, high)]	$\{3, 7, 8\}$	$\{2, 3, 4, 5, 7, 8\}$
[(Humidity, low)]	$\{4\}$	$\{1, 2, 4, 6\}$
[(Humidity, high)]	$\{3, 5, 7, 8\}$	$\{1, 2, 3, 5, 6, 7, 8\}$
[(Temperature, low)]	$\{4, 5, 6\}$	$\{3, 4, 5, 6, 8\}$
[(Temperature, high)]	$\{1, 2, 7\}$	$\{1, 2, 3, 7, 8\}$

A block of a decision-value pair is called a *concept*. In Tables 1 and 2, the concepts are [(Trip, yes)] = {1, 2, 3, 4} and [(Trip, no)] = {5, 6, 7, 8}. Table 3 presents the attribute-value blocks computed for Table 1 (lost values) and Table 2 ("do not care" conditions).

Let B be a subset of the set A of all attributes. For a case $x \in U$ 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 the set $K(x, a) = U$.

Table 4. Characteristic sets for the entire attribute set A

	Characteristic sets	
Case	Lost values	“Do not care” conditions
1	$\{1\}$	$\{1, 2\}$
2	$\{1, 2, 7\}$	$\{1, 2, 3, 7, 8\}$
3	$\{3, 7, 8\}$	$\{2, 3, 5, 7, 8\}$
4	$\{4\}$	$\{4, 6\}$
5	$\{5\}$	$\{3, 5, 6, 8\}$
6	$\{6\}$	$\{4, 5, 6\}$
7	$\{7\}$	$\{2, 3, 7, 8\}$
8	$\{3, 7, 8\}$	$\{2, 3, 5, 7, 8\}$

For example, the characteristic set for case 1 from Table 1 is

$$\begin{aligned} K_A(1) &= [(Wind, low)] \ \cap \ U \ \cap \ [(Temperature, high)] \\ &= \{1, 6\} \ \cap \ \{1, 2, 3, 4, 5, 6, 7, 8\} \ \cap \ \{1, 2, 7\} = \{1\} \end{aligned}$$

and the characteristic set for case 1 from Table 2 is

$$\begin{aligned} K_A(1) &= [(Wind, low)] \ \cap \ U \ \cap \ [(Temperature, high)] \\ &= \{1, 2, 4, 5, 6\} \ \cap \ \{1, 2, 3, 4, 5, 6, 7, 8\} \ \cap \ \{1, 2, 3, 7, 8\} = \{1, 2\}. \end{aligned}$$

All characteristic sets for Tables 1 and 2 are presented in Table 4. For a complete data set the characteristic set $K_B(x)$, where $x \in U$, is an equivalence class of the indiscernibility relation [12, 13].

3 Probabilistic Approximations

In our work we define probabilistic approximations based on the conditional probability of X given $K_B(x)$, $Pr(X \mid K_B(x)) = \frac{|X \ \cap \ K_B(x)|}{|K_B(x)|}$ with $|Y|$ denoting the cardinality of set Y. Let B be a subset of the attribute set A and X be a subset of U.

We further define three kinds of global probablistic approximations: singleton, subset and concept. A B-singleton probabilistic approximation of X with the threshold α, $0 < \alpha \leq 1$, denoted by $appr^{singleton}_{\alpha,B}(X)$, is defined as follows

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

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

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

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

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

Global probabilistic approximations for the concept [(Trip, no)] from Table 1 are presented in Table 5.

Table 5. Global approximations for [(Trip, no)], Table 1

	Probabilistic approximations		
α	singleton	subset	concept
1/3	$\{2, 3, 5, 6, 7, 8\}$	$\{1, 2, 3, 5, 6, 7, 8\}$	$\{3, 5, 6, 7, 8\}$
2/3	$\{3, 5, 6, 7, 8\}$	$\{3, 5, 6, 7, 8\}$	$\{3, 5, 6, 7, 8\}$
1	$\{5, 6, 7\}$	$\{5, 6, 7\}$	$\{5, 6, 7\}$

4 Experiments

In our experiments we used eight real-life data sets taken from the University of California at Irvine *Machine Learning Repository*. These data sets were modified by replacing 35% of existing attribute values by symbols of lost values, i.e., question marks. All data sets with lost values were edited, symbols of lost values were replaced by symbols of "do not care" conditions, i.e., by stars. Thus, for each data set, two data sets were created for experiments, one with missing attribute values interpreted as lost values and the other one as "do not care" conditions.

In our experiments we used the MLEM2 (Modified Learning from Examples Module, version 2) rule induction algorithm of the LERS (Learning from Examples using Rough Sets) data mining system [1, 3, 4].

Probabilistic rules were induced from modified data sets. For each concept X and the set Y equal to a probabilistic approximation of X of a given type (singleton, subset or concept) a modified data set was created, see [9–11]. In this data set all cases from Y had the same decision values as in the original data set, all remaining cases were labeled with a special, additional value. The LERS system, using the MLEM2 algorithm, was used to induce a rule set. Blocks of attribute-value pairs in the MLEM2 algorithm were modified, taking into account

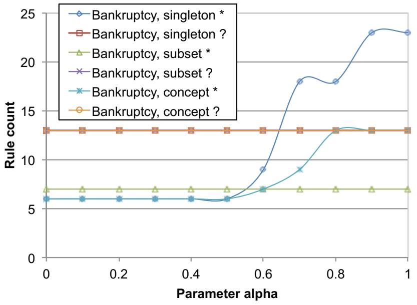


Fig. 1. Rule set size for the *bankruptcy* data set

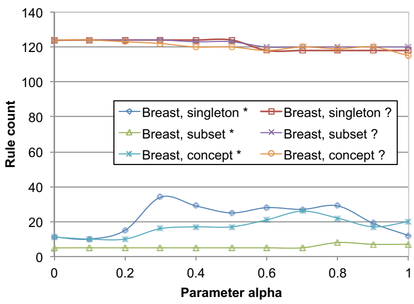


Fig. 2. Rule set size for the *breast cancer* data set

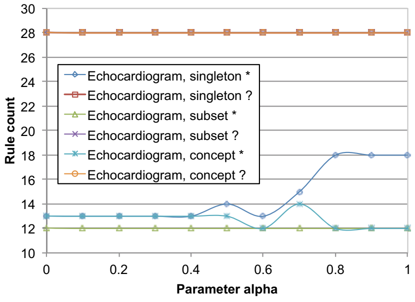


Fig. 3. Rule set size for the *echocardiogram* data set

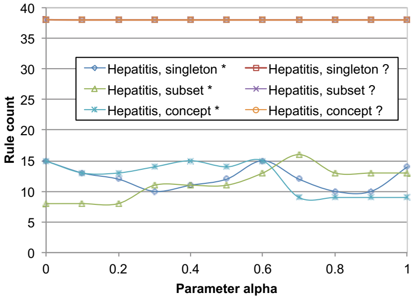


Fig. 4. Rule set size for the *hepatitis* data set

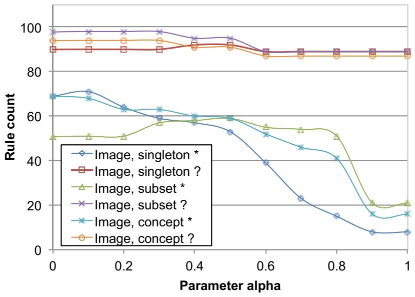


Fig. 5. Rule set size for the *image segmentation* data set

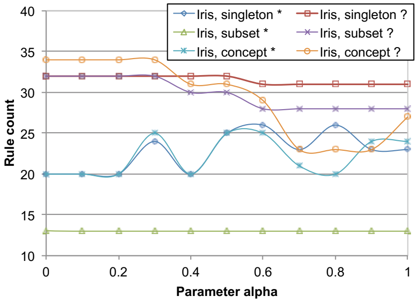


Fig. 6. Rule set size for the *iris* data set

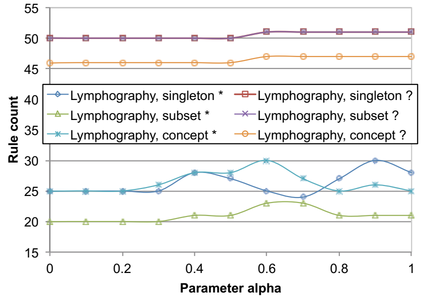


Fig. 7. Rule set size for the *lymphography* data set

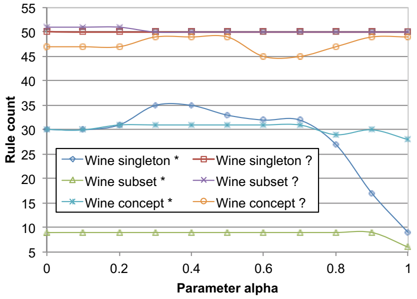


Fig. 8. Rule set size for the *wine recognition* data set

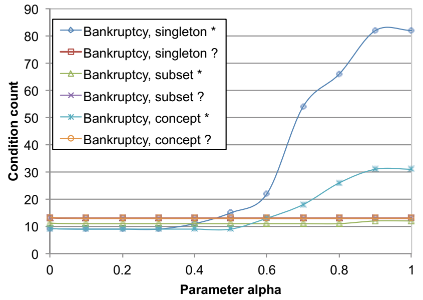


Fig. 9. Condition counts for the *bankruptcy* data set

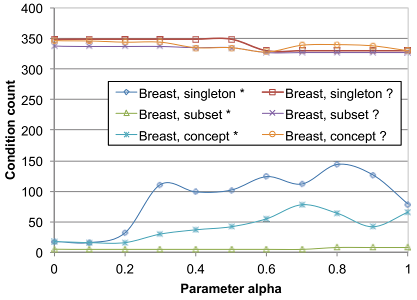


Fig. 10. Condition counts for the *breast cancer* data set

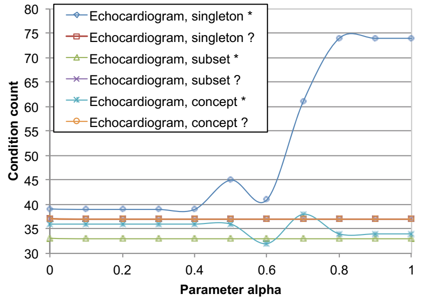


Fig. 11. Condition counts for the *echocardiogram* data set

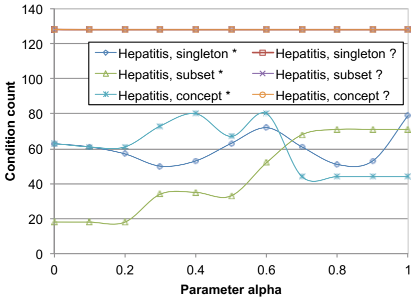
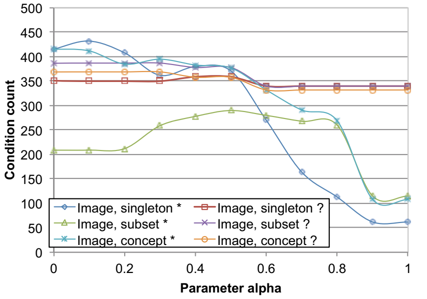


Fig. 12. Condition counts for the *hepatitis* data set



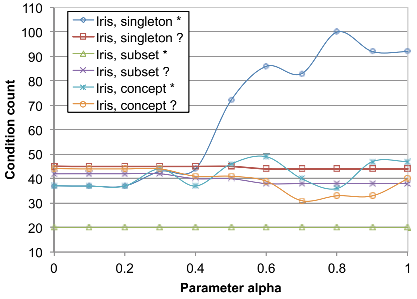


Fig. 13. Condition counts for the *image* data set

Fig. 14. Condition counts for the *iris* data set

missing attribute values. For the modified data set, only rules describing the concept X survived, remaining rules were deleted. The aggregate rule set was combined from rule sets induced from all modified data sets.

For any data set we tested six methods of handling missing attribute values:

- singleton probabilistic approximation combined with lost values, denoted as Singleton, ?,
- singleton probabilistic approximation combined with "do not care" conditions, denoted as Singleton, $*$,
- subset probabilistic approximation combined with lost values, denoted as Subset, ?,
- subset probabilistic approximation combined with "do not care" conditions, denoted as Subset, $*$,
- concept probabilistic approximation combined with lost values, denoted as Concept, ?, and
- concept probabilistic approximation combined with "do not care" conditions, denoted as Concept, $*$.

As follows from [2], all six methods do not differ significantly (Friedman's test (5% significance level) in terms of the error rate.

Our main objective was to compare all six methods in terms of the complexity of rule sets. It is clear that for our data sets the method (Subset, $*$) provides smaller size of rule sets than all three methods associated with lost values: (Singleton, ?), (Subset, ?) and (Concept, ?). Additionally, the same method produces rule sets with smaller total number of conditions than all three methods associated with lost values.

Results of our experiments on the size of rule sets are presented in Figures 1–8. Six selected results on the total number of conditions (because of the space limit) are presented in Figures 9–14.

The method (Subset, $*$) provides smaller size of rule sets than (Singleton, $*$) and (Concept, $*$) for five out of eight data sets: Breast cancer, Echocardiogram, Iris, Lymphography and Wine recognition and smaller total number of conditions for the same data sets (Wilcoxon test, 5% significance level was used for *Echocardiogram*).

Note that on some occasions the difference in performance is quite spectacular, for example, for the Breast cancer data set, (Subset $*$) method provides 5–7 rules (with α between 0.001 and 1) and with 5–8 conditions, while (Singleton, ?), (Subset, ?) and (Concept, ?) methods provide rule sets with 118–124 rules and 330–349 conditions. The error rate for (Subset, $*$) is between 28.52% and 29.90%, for all three methods associated with lost values, the error rate is between 27.44% and 29.90%.

Rule sets induced from data sets with "do not care" conditions are simpler, in general, than rule sets induced from data sets with lost values since for any data set, an attribute-value block for the data set with "do not care" conditions is a superset of the corresponding block (the same attribute-value pair) for the data set with lost values. The MLEM2 rule induction algorithm induces rules using these attribute-value blocks, so a rule induced from the data set with "do not care" conditions covers more cases than a rule induced from the data set with lost values.

5 Conclusions

For a given data set, all six methods of handling missing attribute values (using three kinds of global probabilistic approximations and two interpretations of missing attribute values) do not differ significantly with respect to the error rate [2]. However, as follows from our research presented in this paper, these methods differ significantly with respect to the complexity of rule sets; the simplest rule sets are induced using subset probabilistic approximations and missing attribute values interpreted as "do not care" conditions. Therefore, if we have a choice how to interpret missing attribute values, the best rule set would be induced by subset probabilistic approximations with missing attribute values interpreted as "do not care" conditions.

The focus of this work was a study of rule set complexity using different missing attribute interpretations and approximation methods while applying the same rule induction algorithm, MLEM2. Further investigation with other rule induction algorithms would be need in order to determine if the results are algorithm dependent.

References

1. Clark, P.G., Grzymala-Busse, J.W.: Experiments on probabilistic approximations. In: Proceedings of the 2011 IEEE International Conference on Granular Computing, pp. 144–149 (2011)
2. Clark, P.G., Grzymala-Busse, J.W., Rzasa, W.: Mining incomplete data with singleton, subset and concept approximations (2013) (submitted for publication)
3. Grzymala-Busse, J.W.: A new version of the rule induction system LERS. Fundamenta Informaticae 31, 27–39 (1997)

4. 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)
5. Grzymala-Busse, J.W.: Rough set strategies to data with missing attribute values. In: Workshop Notes, Foundations and New Directions of Data Mining, in Conjunction with the 3rd International Conference on Data Mining, pp. 56–63 (2003)
6. Grzymala-Busse, J.W.: Characteristic relations for incomplete data: A generalization of the indiscernibility relation. In: Tsumoto, S., Słowiński, R., Komorowski, J., Grzymała-Busse, J.W. (eds.) RSCTC 2004. LNCS (LNAI), vol. 3066, pp. 244–253. Springer, Heidelberg (2004)
7. Grzymala-Busse, J.W.: Data with missing attribute values: Generalization of indiscernibility relation and rule induction. Transactions on Rough Sets 1, 78–95 (2004)
8. Grzymala-Busse, J.W.: Generalized parameterized approximations. In: Yao, J., Ramanna, S., Wang, G., Suraj, Z. (eds.) RSKT 2011. LNCS, vol. 6954, pp. 136–145. Springer, Heidelberg (2011)
9. Grzymala-Busse, J.W.: Generalized probabilistic approximations. Transactions on Rough Sets 16, 1–16 (2013)
10. Grzymala-Busse, J.W., Marepally, S.R., Yao, Y.: A comparison of positive, boundary, and possible rules using the MLEM2 rule induction algorithm. In: Proceedings of the 10th International Conference on Hybrid Intelligent Systems, pp. 7–12 (2010)
11. Grzymala-Busse, J.W., Yao, Y.: Probabilistic rule induction with the LERS data mining system. International Journal of Intelligent Systems 26, 518–539 (2011)
12. Pawlak, Z.: Rough Sets. Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Dordrecht (1991)
13. Pawlak, Z., Grzymala-Busse, J.W., Slowinski, R., Ziarko, W.: Rough sets. Communications of the ACM 38, 89–95 (1995)
14. Pawlak, Z., Skowron, A.: Rough sets: Some extensions. Information Sciences 177, 28–40 (2007)
15. Pawlak, Z., Wong, S.K.M., Ziarko, W.: Rough sets: probabilistic versus deterministic approach. International Journal of Man-Machine Studies 29, 81–95 (1988)
16. Ślęzak, D., Ziarko, W.: The investigation of the bayesian rough set model. International Journal of Approximate Reasoning 40, 81–91 (2005)
17. Yao, Y.Y.: Probabilistic rough set approximations. International Journal of Approximate Reasoning 49, 255–271 (2008)
18. Yao, Y.Y., Wong, S.K.M.: A decision theoretic framework for approximate concepts. International Journal of Man-Machine Studies 37, 793–809 (1992)
19. Ziarko, W.: Probabilistic approach to rough sets. International Journal of Approximate Reasoning 49, 272–284 (2008)