Characteristic Sets and Generalized Maximal Consistent Blocks in Mining Incomplete Data

Patrick G. Clark[1], Cheng Gao[1], Jerzy W. Grzymala-Busse[1,2(✉)], and Teresa Mroczek[2]

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

Abstract. Mining incomplete data using approximations based on characteristic sets is a well-established technique. It is applicable to incomplete data sets with a few interpretations of missing attribute values, e.g., lost values and "do not care" conditions. Typically, probabilistic approximations are used in the process. On the other hand, maximal consistent blocks were introduced for incomplete data sets with only "do not care" conditions, using only lower and upper approximations. In this paper we introduce an extension of the maximal consistent blocks to incomplete data sets with any interpretation of missing attribute values and with probabilistic approximations. Additionally, we present results of experiments on mining incomplete data using both characteristic sets and maximal consistent blocks, using lost values and "do not care" conditions. We show that there is a small difference in quality of rule sets induced either way. However, characteristic sets can be computed in polynomial time while computing maximal consistent blocks is associated with exponential time complexity.

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

1 Introduction

We report results of experiments on incomplete data sets, using two interpretations of missing attribute values: lost values and "do not care" conditions [3]. A lost value, denoted by "?", is interpreted as a value that we do not know since it was erased or not inserted into the data set. Rules are induced from existing, specified attribute values. "Do not care" conditions are interpreted as any attribute value. For example, if an attribute is the hair color, and possible values are blond, dark and red, a "do not care" condition is interpreted as any of these three colors.

L. Polkowski et al. (Eds.): IJCRS 2017, Part I, LNAI 10313, pp. 477–486, 2017.
DOI: 10.1007/978-3-319-60837-2_39

For incomplete data sets special kinds of approximations: singleton, subset and concept should be used [3]. In this paper we consider probabilistic approximations, an extension of lower and upper approximations. Such approximations are defined using a probability denoted by α. If $\alpha = 1$, the probabilistic approximation is lower, if α is a positive number, slightly greater than 0, the probabilistic approximation is upper. Such approximations were usually used for completely specified data sets [8,10–17]. Probabilistic approximations were extended to incomplete data sets in [5]. First experimental results on such approximations were reported in [1,2].

Maximal consistent blocks were introduced for incomplete data sets with only "do not care" conditions, using only lower and upper approximations [9]. The main objective of this paper is to extend the definition of maximal consistent blocks to arbitrary interpretation of missing attribute values. Additionally, the obvious question is what a better choice for data mining is: characteristic sets or maximal consistent blocks. We conducted experiments on data sets with two interpretations of missing attribute values, lost values and "do not care" conditions. As a result, we show that there is a small difference in quality of rule sets induced from approximations based on characteristic sets or on maximal consistent blocks. However, characteristic sets can be computed in polynomial time while computing maximal consistent blocks is associated with exponential time complexity.

2 Incomplete Data Sets

An example of incomplete data set is presented in Table 1. A *concept* is a set of all cases with the same decision value. In Table 1 there are two concepts, the set {1, 2, 3, 4} of all cases with flu and the other set {5, 6, 7, 8}.

Table 1. An incomplete data set

	Attributes			Decision
Case	Temperature	Headache	Cough	Flu
1	Normal	?	Yes	Yes
2	*	Yes	No	Yes
3	*	No	*	Yes
4	High	?	Yes	Yes
5	High	No	No	No
6	*	No	Yes	No
7	High	*	?	No
8	Normal	*	No	No

We use notation $a(x) = v$ if an attribute a has the value v for the case x. The set of all cases will be denoted by U. In Table 1, $U = \{1, 2, 3, 4, 5, 6, 7, 8\}$.

For complete data sets, for an attribute-value pair $(a,\ v)$, a *block* of $(a,\ v)$, denoted by $[(a,\ v)]$, is the following set

$$[(a,\ v)] = \{x | x \in U, a(x) = v\}.$$

For incomplete decision tables the definition of a block of an attribute-value pair must be modified in the following way [3,4]:

- If for an attribute a and a case x, $a(x) = ?$, 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 and a case x, $a(x) = *$, then the case x should be included in blocks $[(a, v)]$ for all specified values v of attribute a.

For the data set from Table 1, all of blocks of attribute-value pairs are

$[(Temperature, normal)] = \{1, 2, 3, 6, 8\}$,
$[(Temperature, high)] = \{2, 3, 4, 5, 6, 7\}$,
$[(Headache, no)] = \{3, 5, 6, 7, 8\}$,
$[(Headache, yes)] = \{2, 7, 8\}$,
$[(Cough, no)] = \{2, 3, 5, 8\}$,
$[(Cough, yes)] = \{1, 3, 4, 6\}$.

3 Characteristic Sets and Maximal Consistent Blocks

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 B is a subset of the set A of all attributes and 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 the data set from Table 1, the characteristic sets are

$K_A(1) = \{1, 3, 6\}$,
$K_A(2) = \{2, 8\}$,
$K_A(3) = \{3, 5, 6, 7, 8\}$,
$K_A(4) = \{3, 4, 6\}$,
$K_A(5) = \{3, 5\}$,
$K_A(6) = \{3, 6\}$,
$K_A(7) = \{2, 3, 4, 5, 6, 7\}$,
$K_A(8) = \{2, 3, 8\}$.

The *B-characteristic relation* $R(B)$ is a relation on U defined for $x, y \in U$ as follows:

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

We say that $R(B)$ is *implied* by its B-characteristic sets $K_B(x)$, $x \in U$. The B-characteristic relation $R(B)$ is reflexive but—in general—does not need to be symmetric or transitive. For the data set from Table 1, $R(A) = \{(1, 1), (1, 3), (1, 6), (2, 2), (2, 8), (3, 3), (3, 5), (3, 6), (3, 7), (3, 8), (4, 3), (4, 4), (4, 6), (5, 3), (5, 5), (6, 3), (6, 6), (7, 2), (7, 3), (7, 4), (7, 5), (7, 6), (7, 7), (8, 2), (8, 3), (8, 8)\}$.

Let X be a subset of U. The set X is *consistent* with respect to B if $(x, y) \in R(B)$ for any $x, y \in X$. If there does not exist a consistent subset Y of U such that X is a proper subset of Y, the set X is called a *maximal consistent block* of 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 [9]. Note that in our definition the maximal consistent blocks of B are defined for arbitrary interpretations of missing attribute values. Following [9], we will denote the set of all maximal consistent blocks of B by $\mathscr{C}(\mathrm{B})$. For Table 1, the set of all maximal consistent blocks of A is $\mathscr{C}(\mathrm{A}) = \{\{1\}, \{2, 8\}, \{3, 5\}\ \{3, 6\}, \{3, 7\}, \{3, 8\}, \{4\}\}$.

4 Probabilistic Approximations

For incomplete data sets there exist a number of different definitions of approximations [3]. In this paper we will use only *concept* approximations.

4.1 Probabilistic Approximations Based on Characteristic Sets

Let B be a subset of the set A of all attributes. A B-probabilistic approximation of the set X with the threshold α, $0 < \alpha \leq 1$, based on characteristic sets and denoted by $B\text{-}appr_{\alpha}^{CS}(X)$, is defined as follows

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

where $Pr(X|K_B(x)) = \frac{|X \cap K_B(x)|}{|K_B(x)|}$ is the conditional probability of X given $K_B(x)$ [5]. A-probabilistic approximations of X with the threshold α will be denoted by $appr_{\alpha}^{CS}(X)$.

Table 2. Conditional probabilities $Pr([(Flu, yes)]|K_A(x))$

x	1	2	3	4
$K_A(x)$	{1, 3, 6}	{2, 8}	{3, 5, 6, 7, 8}	{3, 4, 6}
$P(\{1, 2, 3, 4\} \mid K_A(x))$	0.667	0.5	0.2	0.667

For Table 1 and both concepts, all conditional probabilities $P(X|K_A(x))$ are presented in Tables 2 and 3. All distinct probabilistic approximations based on characteristic sets are

Table 3. Conditional probabilities $Pr([(Flu, no)]|K_A(x))$

x	5	6	7	8
$K_A(x)$	{3, 5}	{3, 6}	{2, 3, 4, 5, 6, 7}	{2, 3, 8}
$P(\{5,6,7,8\} \mid K_A(x))$	0.5	0.5	0.5	0.333

$appr_{0.2}^{CS}(\{1,2,3,4\}) = U,$

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

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

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

$appr_{0.333}^{CS}(\{5,6,7,8\}) = \{2,3,4,5,6,7,8\},$

$appr_{0.5}^{CS}(\{5,6,7,8\}) = \{2,3,4,5,6,7\},$

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

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, $appr_{0.4}^{CS}(\{1,2,3,4\}) = appr_{0.5}^{CS}(\{1,2,3,4\})$.

4.2 Probabilistic Approximations Based on Maximal Consistent Blocks

By analogy with the definition of a B-probabilistic approximation based on characteristic sets, a B-probabilistic approximation of the set X with the threshold α, $0 < \alpha \leq 1$, based on maximal consistent blocks and denoted by $B\text{-}appr_{\alpha}^{MCB}(X)$, is defined as follows

$$\cup\{Y \mid Y \in \mathscr{C}(B),\ Pr(X|Y) \geq \alpha\},$$

where $Pr(X|Y) = \frac{|X \cap Y|}{|Y|}$ is the conditional probability of X given Y. A-probabilistic approximations of X, based on maximal consistent blocks, with the threshold α will be denoted by $appr_{\alpha}^{MCB}(X)$.

For Table 1 and the concept $[(Flu, yes)]$, all conditional probabilities $Pr([(Flu, yes)]|Y)$, where $Y \in \mathscr{C}(A)$, are presented in Table 4. Conditional probabilities $Pr([(Flu, no)]|Y)$, where $Y \in \mathscr{C}(A)$, may be computed in an analogous way. All distinct probabilistic approximations based on maximal consistent blocks are

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

Table 4. Conditional probabilities $Pr([(Flu, yes)]|Y)$

Y	{1}	{2, 8}	{3, 5}	{3, 6}	{3, 7}	{3, 8}	{4}
$P(\{1,2,3,4\} \mid Y)$	1	0.5	0.5	0.5	0.5	0.5	1

$$appr_1^{MCB}(\{1,2,3,4\}) = \{1,4\},$$

$$appr_{0.5}^{MCB}(\{5,6,7,8\}) = \{2,3,5,6,7,8\},$$

$$appr_1^{MCB}(\{5,6,7,8\}) = \emptyset.$$

5 Definability

Any union of characteristic sets $K_B(x)$ is called *B-globally definable* [7]. An *A*-globally definable set is called *globally definable*. Let T be a set of attribute-value pairs, where all involved attributes are distinct and are members of a set B. Such set T is called *B-complex*. A block of a B-complex T, denoted by $[T]$, is the set $\cap\{[t]|t \in T\}$. Any union of blocks of B-complexes is called B-locally definable [7]. *A*-locally definable set is called *locally definable*.

Rules are expressed by attribute-value pairs, so any set X may be described by rules if it is locally definable, as was explained in [6]. As follows from [6], maximal consistent blocks for incomplete data sets with only "do not care" conditions are locally definable, so corresponding approximations are also locally definable. However, in general, for arbitrary incomplete data sets, maximal consistent blocks are not locally definable. For example, for the data set from Table 1, sets {1} and {4}, maximal consistent blocks, are not locally definable. Indeed, case 1 occurs in only two blocks: [(Temperature, normal)] and [(Cough, yes)], and the intersection of these two sets is {1, 3, 6}. Similarly, case 4 occurs also in only two blocks: [(Temperature, high)] and [(Cough, yes)], while the intersection of these two sets is {3, 4, 6}. Thus none of the sets: {1}, {4} and {1, 4} can be expressed by rules. From the point of rule induction the set {1, 4} = $appr_1^{MCB}(\{1,2,3,4\})$ is useless.

6 Experiments

Our experiments were conducted on nine data sets obtained from the University of California at Irvine *Machine Learning Repository*. For any data set, a corresponding incomplete data set was created by a random replacement of specified values by question marks (lost values), until an entire row of a data set was full of "?"s. Such a data set was removed from experiments, we used only data sets with at least one specified value for any row. For any incomplete data set with "?"s, another incomplete data set was created by replacing all "?"s by "*"s ("do not care" conditions).

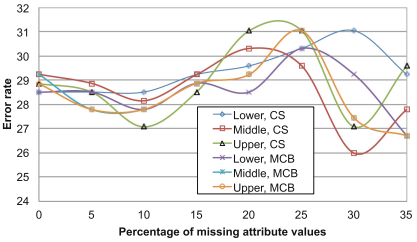

Fig. 1. Number of rules for the *breast cancer* data set

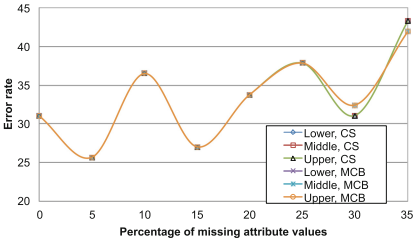

Fig. 2. Error rate for the *echocardiogram* data set with lost values

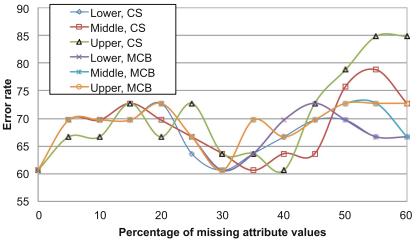

Fig. 3. Error rate for the *global climate* data set with lost values

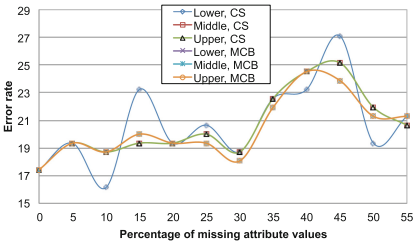

Fig. 4. Error rate for the *hepatitis* data set with lost values

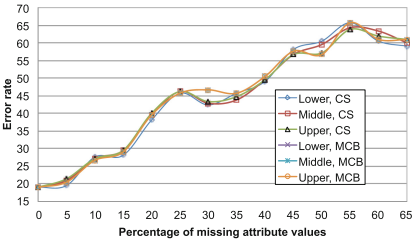

Fig. 5. Error rate for the *image segmentation* data set with lost values

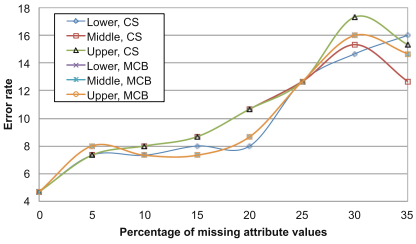

Fig. 6. Error rate for the *iris* data set with lost values

Our main objective was to compare the quality of two approaches to rule induction, based on characteristic sets and maximal consistent blocks, respectively, in terms of an error rate. Note that due to computational complexity, our experiments were restricted to only some percentage of missing attribute values and to some type of incomplete data sets. Results of our experiments, presented in Figs. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 and 12, are restricted to only three data sets with "do not care" conditions, due to excessive computational complexity. In Figs. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 and 12, "Lower" means a lower approximation ($\alpha = 1$), "Middle" means a middle probabilistic approximation ($\alpha = 0.5$), and "Upper" means an upper approximation ($\alpha = 0.001$). Additionally, "CS" means a characteristic set and "MCB" means a maximal consistent block.

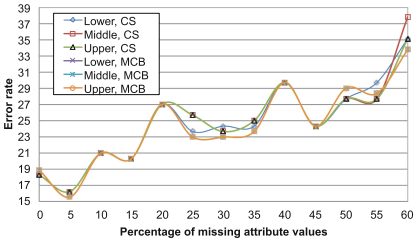

Fig. 7. Error rate for the *lymphography* data set with lost values

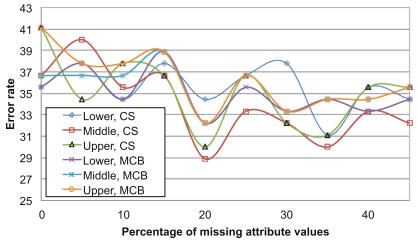

Fig. 8. Error rate for the *postoperative patient* data set with lost values

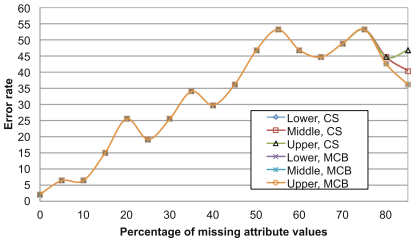

Fig. 9. Error rate for the *small soybean* data set with lost values

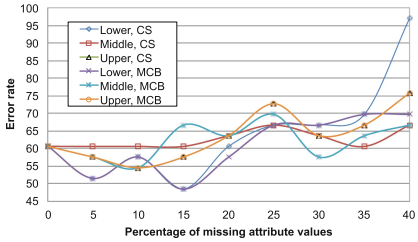

Fig. 10. Error rate for the *global climate* data set with "do not care" conditions

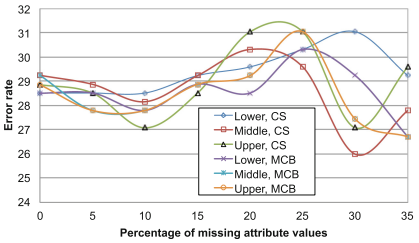

Fig. 11. Error rate for the *echocardiogram* data set with "do not care" conditions

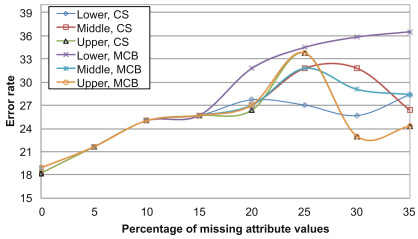

Fig. 12. Error rate for the *lymphography* data set with "do not care" conditions

For a comparison of the two approaches to rule induction, based on characteristic sets and maximal consistent blocks, we used the Friedman rank sum test combined with multiple comparisons, with a 5% level of significance. For all twelve possibilities, presented in Figs. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 and 12, for only one possibility (presented in Fig. 8 for *postoperative patient* with lost values) the null hypothesis of nonsignificant differences between the six methods is rejected: methods based on characteristic sets combined with middle approxi-

mations are better than methods based on maximal consistent blocks with upper approximations.

Thus, we may conclude that there is a small difference in quality of rule sets induced by characteristic sets and maximal consistent blocks. Taking into account computational complexity, it is better to apply data mining to characteristic sets than to maximal consistent blocks.

Originally, maximal consistent blocks were introduced for incomplete data sets with "do not care" conditions [9]. For such data sets rule induction is much more time consuming than for data sets with lost values.

7 Conclusions

In experiments reported in this paper, we compared quality of rule sets induced from characteristic sets and maximal consistent blocks. Results of our experiments show that there is a small difference in quality of rule sets induced using both approaches. Taking into account computational complexity, it is better to apply data mining to characteristic sets than to maximal consistent blocks.

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.: Rule induction using probabilistic approximations and data with missing attribute values. In: Proceedings of the 15-th IASTED International Conference on Artificial Intelligence and Soft Computing ASC 2012, pp. 235–242 (2012)
3. Grzymala-Busse, J.W.: Rough set strategies to data with missing attribute values. In: Notes of the Workshop on Foundations and New Directions of Data Mining, in Conjunction with the Third International Conference on Data Mining, pp. 56–63 (2003)
4. Grzymala-Busse, J.W.: Three approaches to missing attribute values—a rough set perspective. In: Proceedings of the Workshop on Foundation of Data Mining, in Conjunction with the Fourth IEEE International Conference on Data Mining, pp. 55–62 (2004)
5. Grzymala-Busse, J.W.: Generalized parameterized approximations. In: Proceedings of the 6-th International Conference on Rough Sets and Knowledge Technology, pp. 136–145 (2011)
6. Grzymala-Busse, J.W., Mroczek, T.: Definability in mining incomplete data. In: Proceedings of the 20-th International Conference on Knowledge Based and Intelligent Information and Engineering Systems, pp. 179–186 (2016)
7. Grzymala-Busse, J.W., Rzasa, W.: Local and global approximations for incomplete data. In: Proceedings of the Fifth International Conference on Rough Sets and Current Trends in Computing, pp. 244–253 (2006)
8. Grzymala-Busse, J.W., Ziarko, W.: Data mining based on rough sets. In: Wang, J. (ed.) Data Mining: Opportunities and Challenges, pp. 142–173. Idea Group Publishing, Hershey (2003)

9. Leung, Y., Li, D.: Maximal consistent block technique for rule acquisition in incomplete information systems. Inf. Sci. **153**, 85–106 (2003)
10. Pawlak, Z., Skowron, A.: Rough sets: some extensions. Inf. Sci. **177**, 28–40 (2007)
11. Pawlak, Z., Wong, S.K.M., Ziarko, W.: Rough sets: probabilistic versus deterministic approach. Int. J. Man-Mach. Stud. **29**, 81–95 (1988)
12. Ślęzak, D., Ziarko, W.: The investigation of the bayesian rough set model. Int. J. Approx. Reason. **40**, 81–91 (2005)
13. Wong, S.K.M., Ziarko, W.: INFER—an adaptive decision support system based on the probabilistic approximate classification. In: Proceedings of the 6-th International Workshop on Expert Systems and their Applications, pp. 713–726 (1986)
14. Yao, Y.Y.: Probabilistic rough set approximations. Int. J. Approx. Reason. **49**, 255–271 (2008)
15. Yao, Y.Y., Wong, S.K.M.: A decision theoretic framework for approximate concepts. Int. J. Man-Mach. Stud. **37**, 793–809 (1992)
16. Ziarko, W.: Variable precision rough set model. J. Comput. Syst. Sci. **46**(1), 39–59 (1993)
17. Ziarko, W.: Probabilistic approach to rough sets. Int. J. Approx. Reason. **49**, 272–284 (2008)