Mining Incomplete Data Using Global and Saturated Probabilistic Approximations Based on Characteristic Sets and Maximal Consistent Blocks

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

[1] Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS 66045, USA
jerzy@ku.edu

[2] Department of Artificial Intelligence, University of Information Technology and Management, 35-225 Rzeszow, Poland
{zhippe,tmroczek}@wsiz.rzeszow.pl

Abstract. In this paper we discuss incomplete data sets with missing attribute values interpreted as "do not care" conditions. For data mining, we use two types of probabilistic approximations, global and saturated. Such approximations are constructed from two types of granules, characteristic sets and maximal consistent blocks. We present results of experiments on mining incomplete data sets using four approaches, combining two types of probabilistic approximations, global and saturated, with two types of granules, characteristic sets and maximal consistent blocks. We compare these four approaches, using an error rate computed as the result of ten-fold cross validation. We show that there are significant differences (5% level of significance) between these four approaches to data mining. However, there is no universally best approach. Hence, for an incomplete data set, the best approach to data mining should be chosen by trying all four approaches.

Keywords: Data mining · Rough set theory · Characteristic sets · Maximal consistent blocks · Probabilistic approximations

1 Introduction

Incomplete data sets are affected by missing attribute values. In this paper, we consider an interpretation of missing attribute values called a "do not care" condition. According to this interpretation, a missing attribute value may be replaced by any specified attribute value.

For rule induction we use probabilistic approximations, a generalization of the idea of lower and upper approximations known in rough set theory. A probabilistic approximation of the concept X is associated with a probability α; if

S. Ramanna et al. (Eds.): IJCRS 2021, LNAI 12872, pp. 3–17, 2021.
https://doi.org/10.1007/978-3-030-87334-9_1

$\alpha = 1$, the probabilistic approximation becomes the lower approximation of X; if α is a small positive number, e.g., 0.001, the probabilistic approximation is reduced to the upper approximation of X. Usually, probabilistic approximations are applied to completely specified data sets [18,20–27], such approximations are generalized to incomplete data sets, using characteristic sets, in [13,14], and maximal consistent blocks in [1,2].

Missing attribute values are usually categorized into lost values and "do not care" conditions. A lost value, denoted by "?", is unavailable for the process of data mining, while a 'do not care" condition, denoted by "*", represents any value of the corresponding attribute.

Recently, two new types of approximations were introduced, global probabilistic approximations in [3] and saturated probabilistic approximations in [8]. Results of experiments on an error rate, evaluated by ten-fold cross validation, were presented for characteristic sets in [6–8] and for maximal consistent blocks in [1,2]. In these experiments, global and saturated probabilistic approximations based on characteristic sets were explored using data sets with lost values and "do not care" conditions. Results show that among these four methods there is no universally best method.

The main objective of this paper is a comparison of four approaches to mining data, using two probabilistic approximations, global and saturated, based on two granules, characteristic sets and maximal consistent blocks, in terms of an error rate evaluated by ten-fold cross validation.

Rule induction was conducted using a new version of the Modified Learning from Examples Module, version 2 (MLEM2) [5,12]. The MLEM2 algorithm is a component of the Learning from Examples using Rough Sets (LERS) data mining system [4,11,12].

2 Incomplete Data

We assume that the input data sets are presented in the form of a decision table. An example of the decision table is shown in Table 1. 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 1, $U = \{1, 2, 3, 4, 5, 6, 7, 8\}$. 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 1, $A = \{$*Temperature*, *Wind*, *Humidity*$\}$ and d is *Trip*. The value for a case x and an attribute a will be denoted by $a(x)$. For example, *Temperature*$(1) =$ *normal*.

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 $\{1, 2, 3\}$.

A *block* of the attribute-value pair (a, v), denoted by $[(a, v)]$, is the set $\{x \in U \mid a(x) = v\}$ [10]. For incomplete decision tables, the definition of a block of an attribute-value pair is modified in the following way:

– if for an attribute a and a case x we have $a(x) = ?$, the case x should not be included in any blocks $[(a, v)]$ for all values v of attribute a;

Table 1. A decision table

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

– if for an attribute a and a case x we have $a(x) = *$, the case x should be included in blocks $[(a, v)]$ for all specified values v of attribute a.

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

[(Temperature, normal)] = {1, 3, 4},
[(Temperature, high)] = {2, 3},
[(Temperature, very-high)] = {3, 6, 7},
[(Wind, no)] = {1, 2, 4, 6}.
[(Wind, yes)] = {1, 4, 5, 6},
[(Humidity, no)] = {1, 3, 4, 5, 7},
[(Humidity, yes)] = {4, 5, 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 $a(x) =$? or $a(x) = *$, then $K(x, a) = U$.

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

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

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

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

We quote some definitions from [1]. Let X be a subset of U. The set X is *B-consistent* if $(x, y) \in R(B)$ for any $x, y \in X$. If there does not exist a *B-*

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

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

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

will be denoted by $\mathscr{C}_B(x)$.

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 [19]. Note that in our definition, the generalized maximal consistent blocks of B are defined for arbitrary interpretations of missing attribute values. For Table 1, the generalized maximal A-consistent blocks $\mathscr{C}_A(x)$ are

$\mathscr{C}_A(1) = \{\{1, 3, 4\}\}$,
$\mathscr{C}_A(2) = \{\{2\}\}$,
$\mathscr{C}_A(3) = \{\{3, 7\}, \{1, 3, 4\}\}$,
$\mathscr{C}_A(4) = \{\{1, 3, 4\}\}$,
$\mathscr{C}_A(5) = \{\{5\}\}$,
$\mathscr{C}_A(6) = \{\{6, 7\}\}$,
$\mathscr{C}_A(7) = \{\{3, 7\}, \{6, 7\}\}$, and
$\mathscr{C}_A(8) = \{\{8\}\}$.

3 Probabilistic Approximations

In this section, we will discuss two types of probabilistic approximations: global and saturated.

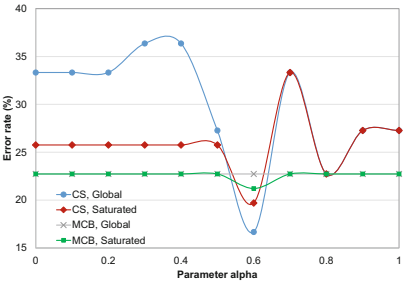

Fig. 1. The *bankruptcy* data set

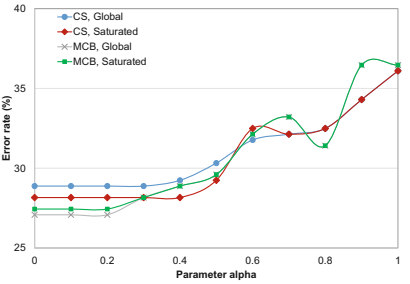

Fig. 2. The *breast cancer* data set

3.1 Global Probabilistic Approximations Based on Characteristic Sets

An idea of the global probabilistic approximation, restricted to lower and upper approximations, was introduced in [16,17], and presented in a general form in [3]. Let X be a concept, $X \subseteq U$. A *B-global probabilistic approximation* of the

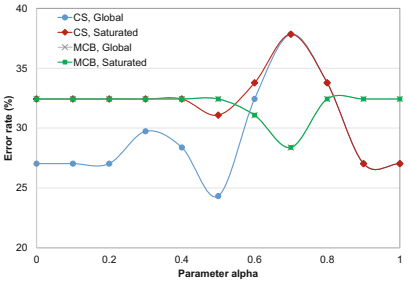

Fig. 3. The *echocardiogram* data set

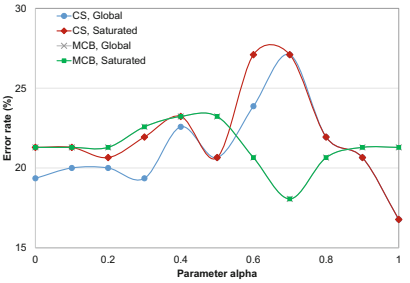

Fig. 4. The *hepatitis* data set

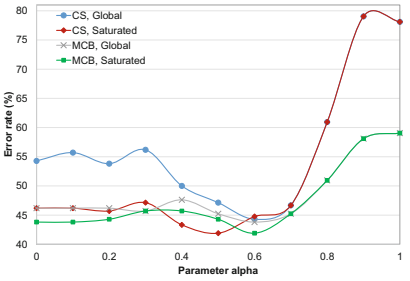

Fig. 5. The *image segmentation* data set

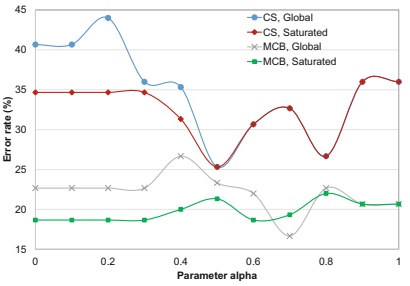

Fig. 6. The *iris* data set

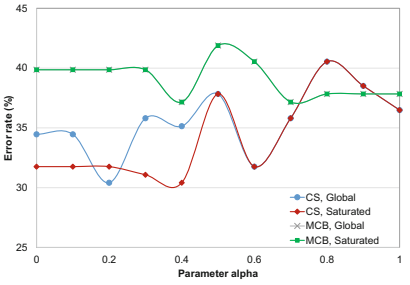

Fig. 7. The *lymphography* data set

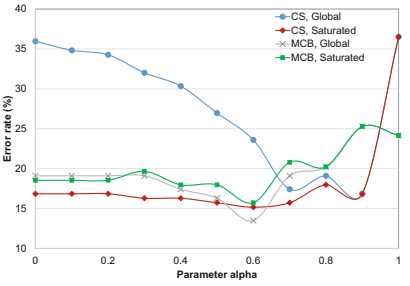

Fig. 8. The *wine recognition* data set

concept X, based on characteristic sets, with the parameter α and denoted by $appr^{global}_{\alpha,B}(X)$ is defined as the following set

$$\bigcup\{K_B(x) \mid \exists\ Y \subseteq U\ \forall x \in Y,\ Pr(X|K_B(x)) \ \geq \alpha\}. \tag{1}$$

Obviously, for some sets B and X and the parameter α, there exist many B-global probabilistic approximations of X. In addition, the algorithm for computing B-global probabilistic approximations is of exponential computational complexity. Therefore, in our experiments we used a heuristic version of the

definition of B-global probabilistic approximation, called a MLEM2 B-global probabilistic approximation of the concept X, associated with a parameter α and denoted by $appr_{\alpha,B}^{mlem2}(X)$ [3]. This definition is based on the rule induction algorithm MLEM2 [12]. The MLEM2 algorithm is used in the Learning from Examples using Rough Sets (LERS) data mining system [4,11,12]. The approximation $appr_{\alpha,B}^{mlem2}(X)$ is constructed from characteristic sets $K_B(y)$, the most relevant to the concept X, i.e., with $|X \cap K_B(y)|$ as large as possible and $Pr(X|K_B(y)) \geq \alpha$, where $y \in U$. If more than one characteristic set $K_B(y)$ satisfies both conditions, we pick the characteristic set $K_B(y)$ with the largest $Pr(X|K_B(y))$. If this criterion ends up with a tie, a characteristic set is picked up heuristically, as the first on the list [3].

In this paper, we study MLEM2 B-global probabilistic approximations based on characteristic sets, with $B = A$. Such approximations are called, for simplicity, *global probabilistic approximations* associated with the parameter α, denoted by $appr_{\alpha}^{global}(X)$. Similarly, for $B = A$, the characteristic set $K_B(X)$ is denoted by $K(x)$.

Let $E_\alpha(X)$ be the set of all eligible characteristic sets defined as follows

$$\{K(x) \mid x \in U, Pr(X|K(x)) \geq \alpha\}. \tag{2}$$

A heuristic version of the global probabilistic approximation based on characteristic sets is presented below.

Global probabilistic approximation
based on characteristic sets algorithm
input: a set X (a concept), a set $E_\alpha(X)$,
output: a set T ($appr_{\alpha}^{global}(X)$)
begin
    $G := X$;
    $T := \emptyset$;
    $Y := E_\alpha(X)$;
    while $G \neq \emptyset$ **and** $Y \neq \emptyset$
        begin
            select a characteristic set $K(x) \in Y$
            such that $|K(x) \cap X|$ is maximum;
            if a tie occurs, select $K(x) \in Y$
            with the smallest cardinality;
            if another tie occurs, select the first $K(x)$;
            $T := T \cup K(x)$;
            $G := G - T$;
            $Y := Y - K(x)$
        end
end

For Table 1, all distinct global probabilistic approximations based on characteristic sets are

$$appr_1^{global}(\{1,2,3\}) = \{2\},$$
$$appr_{0.667}^{global}(\{1,2,3\}) = \{1,2,3,4\},$$
$$appr_{0.4}^{global}(\{1,2,3\}) = \{1,2,3,4,5,7\},$$
$$appr_1^{global}(\{4,5,6,7,8\}) = \{4,5,7,8\},$$
$$appr_{0.75}^{global}(\{4,5,6,7,8\}) = \{1,4,5,6,7,8\},$$

3.2 Saturated Probabilistic Approximations Based on Characteristic Sets

Another heuristic version of the probabilistic approximation is based on selection of characteristic sets while giving higher priority to characteristic sets with larger conditional probability $Pr(X|K(x))$. Additionally, if the approximation covers all cases from the concept X, we stop adding characteristic sets.

Let X be a concept and let $x \in U$. Let us compute all conditional probabilities $Pr(X|K(x))$. Then, we sort the set

$$\{Pr(X|K(x)) \mid x \in U\}. \tag{3}$$

Let us denote the sorted list of such conditional probabilities by α_1, α_2,..., α_n, where α_1 is the largest. For any $i = 1, 2,..., n$, the set $E_i(x)$ is defined as follows

$$\{K(x) \mid x \in U, Pr(X|K(x)) = \alpha_i\}. \tag{4}$$

If we want to compute a saturated probabilistic approximation, denoted by $appr_\alpha^{saturated}(X)$, for some α, $0 < \alpha \leq 1$, we need to identify the index m such that

$$\alpha_m \geq \alpha > \alpha_{m+1}, \tag{5}$$

where $m \in \{1, 2, ..., n\}$ and $\alpha_{n+1} = 0$. Then, the saturated probabilistic approximation $appr_{\alpha_m}^{saturated}(X)$ is computed using the following algorithm.

Saturated probabilistic approximation
based on characteristic sets algorithm
input: a set X (a concept), a set $E_i(x)$ for
$i = 1, 2,..., n$ and $x \in U$, index m
output: a set T ($appr_{\alpha_m}^{saturated}(X)$)
begin
    $T := \emptyset$;
    $Y_i(x) := E_i(x)$ for all $i = 1, 2,..., m$ and $x \in U$;
    for $j = 1, 2,..., m$ **do**
        while $Y_j(x) \neq \emptyset$
            begin

```
                    select a characteristic set K(x) ∈ Y_j(x)
                    such that |K(x) ∩ X| is maximum;
                    if a tie occurs, select the first K(x);
                    Y_j(x) := Y_j(x) − K(x);
                    if (K(x) − T) ∩ X ≠ ∅
                      then T := T ∪ K(x);
                    if X ⊆ T then exit
                end
end
```

For Table 1, all distinct saturated probabilistic approximations based on characteristic sets are

$$appr_1^{saturated}(\{1,2,3\}) = \{2\},$$
$$appr_{0.667}^{saturated}(\{1,2,3\}) = \{1,2,3,4\},$$
$$appr_1^{saturated}(\{4,5,6,7,8\}) = \{4,5,7,8\},$$
$$appr_{0.75}^{saturated}(\{4,5,6,7,8\}) = \{1,4,5,6,7,8\},$$

3.3 Global Probabilistic Approximations Based on Maximal Consistent Blocks

A special case of the global probabilistic approximation, limited only to lower and upper approximations and to characteristic sets, was introduced in [16,17]. A general definition of the global probabilistic approximation was introduced in [9].

A *B-global probabilistic approximation based on Maximal Consistent Blocks* of the concept X, with the parameter α and denoted by $appr_{\alpha,B}^{global}(X)$ is defined as follows

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

Obviously, for given sets B and X and the parameter α, there exist many B-global probabilistic approximations of X. Additionally, an algorithm for computing B-global probabilistic approximations is of exponential computational complexity. So, we decided to use a heuristic version of the definition of B-global probabilistic approximation, called the MLEM2 B-global probabilistic approximation of the concept X, associated with a parameter α and denoted by $appr_{\alpha,B}^{mlem2}(X)$ [3]. This definition is based on the rule induction algorithm MLEM2. The approximation $appr_{\alpha,B}^{mlem2}(X)$ is a union of the generalized maximal consistent blocks $Y \in \mathscr{C}(B)$, the most relevant to the concept X, i.e., with $|X \cap Y|$ as large as possible and with $Pr(X|Y) \geq \alpha$. If more than one generalized maximal consistent block Y satisfies both conditions, the generalized maximal consistent block Y with the largest $Pr(X|Y) \geq \alpha$ is selected. If this

criterion ends up with a tie, a generalized maximal consistent block Y is picked up heuristically, as the first on the list [3].

Special MLEM2 B-global probabilistic approximations, with $B = A$, are called *global probabilistic approximations* associated with the parameter α, and are denoted by $appr_{\alpha}^{mlem2}(X)$.

Let $E_{\alpha}(X)$ be the set of all eligible generalized maximal consistent blocks defined as follows

$$\{Y \mid Y \subseteq \mathscr{C}(A), Pr(X|Y) \geq \alpha\}.$$

A heuristic version of the global probabilistic approximation is computed using the following algorithm

Global probabilistic approximation
based on maximal consistent blocks algorithm
input: a set X (a concept), a set $E_{\alpha}(X)$,
output: a set T (a global probabilistic approximation $appr_{\alpha}^{mlem2}(X)$) of X
begin
$G := X$;
$T := \emptyset$;
$\mathscr{Y} := E_{\alpha}(X)$;
while $G \neq \emptyset$ **and** $\mathscr{Y} \neq \emptyset$
begin
select a generalized maximal consistent block $Y \in \mathscr{Y}$
such that $|X \cap Y|$ is maximum;
if a tie occurs, select $Y \in \mathscr{Y}$
with the smallest cardinality;
if another tie occurs, select the first $Y \in \mathscr{Y}$;
$T := T \cup Y$;
$G := G - T$;
$\mathscr{Y} := \mathscr{Y} - Y$
end
end

For Table 1, all distinct global probabilistic approximations based on maximal consistent blocks are

$$appr_{1}^{global}(\{1,2,3\}) = \{2\},$$
$$appr_{0.667}^{global}(\{1,2,3\}) = \{1,2,3,4\},$$
$$appr_{1}^{global}(\{4,5,6,7,8\}) = \{5,6,7,8\},$$
$$appr_{0.5}^{global}(\{4,5,6,7,8\}) = \{3,5,6,7,8\},$$
$$appr_{0.333}^{global}(\{4,5,6,7,8\}) = \{1,3,4,5,6,7,8\},$$

3.4 Saturated Probabilistic Approximations Based on Maximal Consistent Blocks

Saturated probabilistic approximations are unions of generalized maximal consistent blocks while giving higher priority to generalized maximal consistent blocks with larger conditional probability $Pr(X|Y)$. Additionally, if the approximation covers all cases from the concept X, we stop adding generalized maximal consistent blocks.

Let X be a concept and let $x \in U$. Let us compute all conditional probabilities $Pr(X|Z)$, where $Z \in \{Y \mid Y \subseteq \mathscr{C}(A), Pr(X|Y) \geq \alpha\}$. Then we sort the set

$$\{Pr(X|Y) \mid Y \subseteq \mathscr{C}(A)\}$$

in descending order. Let us denote the sorted list of such conditional probabilities by α_1, α_2,..., α_n. For any $i = 1, 2, ..., n$, the set $E_i(X)$ is defined as follows

$$\{Y \mid Y \subseteq \mathscr{C}(A), Pr(X|Y) = \alpha_i\}.$$

If we want to compute a saturated probabilistic approximation, denoted by $appr_{\alpha}^{saturated}(X)$, for some α, $0 < \alpha \leq 1$, we need to identify the index m such that

$$\alpha_m \geq \alpha > \alpha_{m+1},$$

where $m \in \{1, 2, ..., n\}$ and $\alpha_{n+1} = 0$. The saturated probabilistic approximation $appr_{\alpha_m}^{saturated}(X)$ is computed using the following algorithm

Saturated probabilistic approximation
based on maximal consistent blocks algorithm
input: a set X (a concept), a set $E_i(X)$ for $i = 1, 2, ..., n$, index m
output: a set T (a saturated probabilistic approximation $appr_{\alpha_m}^{saturated}(X)$) of X
begin
    $T := \emptyset$;
    $\mathscr{Y}_i(X) := E_i(X)$ for all $i = 1, 2, ..., m$;
    for $j = 1, 2, ..., m$ **do**
        while $\mathscr{Y}_j(X) \neq \emptyset$
            begin
                select a generalized maximal consistent block $Y \in \mathscr{Y}_j(X)$
                such that $|X \cap Y|$ is maximum;
                if a tie occurs, select the first Y;
                $\mathscr{Y}_j(X) := \mathscr{Y}_j(X) - Y$;
                if $(Y - T) \cap X \neq \emptyset$
                  then $T := T \cup Y$;
                if $X \subseteq T$ **then exit**
            end
end

For Table 1, any saturated probabilistic approximation based on maximal consistent blocks for is the same as corresponding global probabilistic approximation based on maximal consistent blocks for the same concept.

3.5 Rule Induction

Once the global and saturated probabilistic approximations associated with a parameter α are constructed, rule sets are induced using the rule induction algorithm based on another parameter, also interpreted as a probability, and denoted by β. This algorithm also uses the MLEM2 principles [15], and was presented, e.g., in [3].

MLEM2 rule induction algorithm
input: a set Y (an approximation of X) and a parameter β,
output: a set $\mathcal{T}$ (a rule set),
begin
$G := Y$;
$D := Y$;
$\mathcal{T} := \emptyset$;
$\mathcal{J} := \emptyset$;
while $G \neq \emptyset$
begin
$T := \emptyset$;
$T_s := \emptyset$;
$T_n := \emptyset$;
$T(G) := \{t \mid [t] \cap G \neq \emptyset\}$;
while $(T = \emptyset$ **or** $[T] \not\subseteq D)$ **and** $T(G) \neq \emptyset$
begin
select a pair $t = (a_t, v_t) \in T(G)$ with
maximum of $|[t] \cap G|$; if a tie occurs,
select a pair $t \in T(G)$ with the smallest
cardinality of $[t]$; if another tie occurs,
select the first pair;
$T := T \cup \{t\}$;
$G := [t] \cap G$;
$T(G) := \{t \mid [t] \cap G \neq \emptyset\}$;
if a_t is symbolic {let V_{a_t} be the domain of a_t}
then
$T_s := T_s \cup \{(a_t, v) | v \in V_{a_t}\}$
else {a_t is numerical, let $t = (a_t, u..v)$}
and $T_n := T_n \cup \{(a_t, x..y) \mid$ disjoint $x..y$
and $u..v\} \cup \{(a_t, x..y \mid x..y \supseteq u..v\}$;
$T(G) := T(G) - (T_s \cup T_n)$;
end {while};
if $Pr(X \mid [T]) \geq \beta$
then

```
                                        begin
                                            D := D ∪ [T];
                                            T := T ∪ {T};
                                        end {then}
                                    else J := J ∪ {T};
                                G := D − ∪_{S∈T∪J}[S];
                    end {while};
            for each T ∈ T do
                    for each numerical attribute a_t with
                    (a_t, u..v) ∈ T do
                        while (T contains at least two different
                        pairs (a_t, u..v) and (a_t, x..y) with
                        the same numerical attribute a_t)
                            replace these two pairs with a new pair
                            (a_t, common part of (u..v) and (x..y));
                    for each t ∈ T do
                        if [T − {t}] ⊆ D then T := T − {t};
            for each T ∈ T do
                if ∪_{S∈(T−{T})}[S] = ∪_{S∈T}[S] then T := T − {T};
end {procedure}.
```

For example, for Table 1 and $\alpha = \beta = 0.5$, using the global probabilistic approximations, the MLEM2 rule induction algorithm induces the following rules:

(Temperature, very-high) & (Headache, yes) → (Flu, yes)

(Temperature, high) & (Cough, yes) → (Flu, yes)

(Headache, no) & (Cough, no) → (Flu, no)

and

(Temperature, normal) → (Flu, no)

4 Experiments

For our experiments, we used eight data sets taken from the *Machine Learning Repository* at the University of California at Irvine. For every data set, a new record was created by randomly replacing 35% of existing specified attribute values by *"do not care" conditions*.

In our experiments, the parameter α varied between 0.001 and 1 while the parameter β was equal to 0.5. For any data set, ten-fold cross validation was conducted. Results of our experiments are presented in Figs. 1, 2, 3, 4, 5 6 and 7, where "CS" denotes a characteristic set, "MCB" denotes a generalized maximal consistent block, "Global" denotes a MLEM2 global probabilistic approximation and "Saturated" denotes a saturated probabilistic approximation. In our

experiments, four methods for mining incomplete data sets were used, since we combined two types of granules from which approximations are constructed: characteristic sets and generalized maximal consistent blocks with two versions of probabilistic approximations: global and saturated.

These four methods were compared by applying the distribution free Friedman rank sum test and then by the post-hoc test (distribution-free multiple comparisons based on the Friedman rank sums), with a 5% level of significance.

For three data sets: *bankruptcy*, *image segmentation* and *iris*, two methods: global and saturated probabilistic approximations based on maximal consistent blocks are significantly better (error rates evaluated by ten-fold cross validation are smaller) than global probabilistic approximations based on characteristic sets. Additionally, for the *iris* data set saturated probabilistic approximations based on maximal consistent blocks are significantly better than saturated probabilistic approximations based on characteristic sets.

On the other hand, for the data set *lymphography*, saturated global approximations based on characteristic sets are better than both global and saturated probabilistic approximations based on maximal consistent blocks. For the data set *wine recognition*, saturated probabilistic approximations based on characteristic sets are better than both global probabilistic approximations based on characteristic sets and saturated probabilistic approximations based on maximal consistent blocks.

For three data sets, *breast cancer*, *echocardiogram* and *hepatitis*, pairwise differences in an error rate, evaluated by ten-fold cross validation between these four approaches to data mining, are statistically insignificant.

5 Conclusions

We compared four methods for mining incomplete data sets, combining two granules, characteristic sets and generalized maximal consistent blocks with two types of probabilistic approximations, global and saturated. Our criterion of quality was an error rate evaluated by ten-fold cross validation. As follows from our experiments, there are no significant differences between the four methods. The main conclusion is that for data mining all four methods should be applied.

References

1. Clark, P.G., Gao, C., Grzymala-Busse, J.W., Mroczek, T.: Characteristic sets and generalized maximal consistent blocks in mining incomplete data. In: Proceedings of the International Joint Conference on Rough Sets, Part 1, pp. 477–486 (2017)
2. Clark, P.G., Gao, C., Grzymala-Busse, J.W., Mroczek, T.: Characteristic sets and generalized maximal consistent blocks in mining incomplete data. Inf. Sci. **453**, 66–79 (2018)
3. Clark, P.G., Gao, C., Grzymala-Busse, J.W., Mroczek, T., Niemiec, R.: A comparison of concept and global probabilistic approximations based on mining incomplete data. In: Vasiljevienė, G. (ed.) Information and Software Technologies. ICIST 2018. Communications in Computer and Information Science, **920**, pp. 324–335 (2018). https://doi.org/10.1007/978-3-319-99972-2_26

4. 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)
5. Clark, P.G., Grzymala-Busse, J.W.: Experiments on rule induction from incomplete data using three probabilistic approximations. In: Proceedings of the 2012 IEEE International Conference on Granular Computing, pp. 90–95 (2012)
6. Clark, P.G., Grzymala-Busse, J.W., Hippe, Z.S., Mroczek, T., Niemiec, R.: Global and saturated probabilistic approximations based on generalized maximal consistent blocks. In: Proceedings of the 15-th International Conference on Hybrid Artificial Intelligence Systems, pp. 387–396 (2020)
7. Clark, P.G., Grzymala-Busse, J.W., Hippe, Z.S., Mroczek, T., Niemiec, R.: Mining data with many missing attribute values using global and saturated probabilistic approximations. In: Proceedings of the 26-th International Conference on Information and Software Technologies, pp. 72–83 (2020)
8. Clark, P.G., Grzymala-Busse, J.W., Mroczek, T., Niemiec, R.: A comparison of global and saturated probabilistic approximations using characteristic sets in mining incomplete data. In: Proceedings of the Eight International Conference on Intelligent Systems and Applications, pp. 10–15 (2019)
9. Clark, P.G., Grzymala-Busse, J.W., Mroczek, T., Niemiec, R.: Rule set complexity in mining incomplete data using global and saturated probabilistic approximations. In: Proceedings of the 25-th International Conference on Information and Software Technologies, pp. 451–462 (2019)
10. Grzymala-Busse, J.W.: LERS–a system for learning from examples based on rough sets. In: Slowinski, R. (ed.) Intelligent Decision Support. Handbook of Applications and Advances of the Rough Set Theory, pp. 3–18. Kluwer Academic Publishers, Dordrecht, Boston, London (1992)
11. Grzymala-Busse, J.W.: A new version of the rule induction system LERS. Fund. Inf. **31**, 27–39 (1997)
12. 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)
13. 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)
14. 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)
15. Grzymala-Busse, J.W., Clark, P.G., Kuehnhausen, M.: Generalized probabilistic approximations of incomplete data. Int. J. Approx. Reason. **132**, 180–196 (2014)
16. 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)
17. Grzymala-Busse, J.W., Rzasa, W.: Local and global approximations for incomplete data. Trans. Rough Sets **8**, 21–34 (2008)
18. 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 Publ., Hershey, PA (2003)
19. Leung, Y., Li, D.: Maximal consistent block technique for rule acquisition in incomplete information systems. Inf. Sci. **153**, 85–106 (2003)

20. Pawlak, Z., Skowron, A.: Rough sets: some extensions. Inf. Sci. **177**, 28–40 (2007)
21. Pawlak, Z., Wong, S.K.M., Ziarko, W.: Rough sets: probabilistic versus deterministic approach. Int. J. Man-Mach. Stud. **29**, 81–95 (1988)
22. Ślęzak, D., Ziarko, W.: The investigation of the Bayesian rough set model. Int. J. Approx. Reason. **40**, 81–91 (2005)
23. 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)
24. Yao, Y.Y.: Probabilistic rough set approximations. Int. J. Approx. Reason. **49**, 255–271 (2008)
25. Yao, Y.Y., Wong, S.K.M.: A decision theoretic framework for approximate concepts. Int. J. Man-Mach. Stud. **37**, 793–809 (1992)
26. Ziarko, W.: Variable precision rough set model. J. Comput. Syst. Sci. **46**(1), 39–59 (1993)
27. Ziarko, W.: Probabilistic approach to rough sets. Int. J. Approx. Reason. **49**, 272–284 (2008)