A Comparison of Global and Local Probabilistic Approximations in Mining Data with Many Missing Attribute Values

Patrick G. Clark
Department of Electrical Eng. and Computer Sci.
University of Kansas
Lawrence, KS 66045, USA
e-mail: patrick.g.clark@gmail.com

Jerzy W. Grzymala-Busse
Department of Electrical Eng. and Computer Sci.
University of Kansas
Lawrence, KS 66045, USA
Institute of Computer Science
Polish Academy of Sciences
01-237 Warsaw, Poland
e-mail: jerzy@ku.edu

***Abstract*—We present results of a novel experimental comparison of global and local probabilistic approximations. Global approximations are unions of characteristic sets while local approximations are constructed from blocks of attribute-value pairs. Two interpretations of missing attribute values are discussed: lost values and "do not care" conditions. Our main objective was to compare global and local probabilistic approximations in terms of the error rate. For our experiments we used six incomplete data sets with many missing attribute values. The best results were accomplished by global approximations (for two data sets), by local approximations (for one data set), and for the remaining three data sets the experiments ended with ties. Our next objective was to check the quality of non-standard probabilistic approximations, i.e., probabilistic approximations that were neither lower nor upper approximations. For four data sets the smallest error rate was accomplished by non-standard probabilistic approximations, for the remaining two data sets the smallest error rate was accomplished by upper approximations. Our final objective was to compare two interpretations of missing attribute values. For three data sets the best interpretation was the lost value, for one data set it was the "do not care" condition, for the remaining two cases there was a tie.**

***Keywords*-Data mining, rough set theory, probabilistic approximations, parameterized approximations, MLEM2 rule induction algorithm**

I. Introduction

Rough set theory is one of the basic approaches used in data mining. Fundamental ideas of rough set theory are approximations of a concept (a subset of the universe with the same decision value). Initially, lower and upper approximations were used for complete data sets, i.e., data sets with only specified, known attribute values.

For incomplete data sets, in which some attribute values are missing, there exist many definitions of approximations. Three frequently used are: singleton, subset and concept approximations [10], [11]. All three approximations are global, in the sense that they are constructed from characteristic sets, by analogy with approximations for complete data sets, which are created from elementary sets (equivalence classes of the indiscernibility relation), [19], [20]. Both characteristic sets and elementary sets are defined using the set of all cases.

A probabilistic approximation, associated with an additional parameter (probability) α, is a generalization of lower and upper approximations. For $\alpha = 1$, the probabilistic approximation of the set X, a subset of the universe U, is a lower approximation of X. For a quite small α it becomes an upper approximation of X. Probabilistic approximations for complete data sets were investigated for years, the idea was introduced in [23] and studied in many papers [18], [21], [22], [24], [25]. In these papers mostly theoretical properties of such approximations were discussed. The first paper on experimental validation of probabilistic approximations for complete data sets was [1].

For incomplete data sets global approximations were generalized to global probabilistic approximations in [12]. Results of experiments on global probabilistic approximation on incomplete data sets were published in [1]–[5].

Additionally, in [7] global probabilistic approximations for incomplete data were compared with probabilistic approaches to missing attribute values. The approaches included the most common value for symbolic attributes, the average value for numerical attributes, the concept most common value for symbolic attributes, and the concept average value for numerical attributes. The most common value and average value probabilistic approaches were selected since they belong to the most successful approaches to missing attribute values [15]. It turned out that global probabilistic approaches were better for five out of six data sets with many missing attribute values, the same data sets that were used for experiments described in this paper.

A new type of approximation, used for incomplete data sets, is a local approximation, introduced in [16]. This approximation is defined using blocks of attribute-value pairs, smaller granules than characteristic sets. Local approximations were extended to local probabilistic approximations in [6], [14].

Table I
AN INCOMPLETE DECISION TABLE

	Attributes			Decision
Case	Age	Education	Gender	Hobby
1	under-21	elementary	male	fishing
2	?	secondary	female	fishing
3	21–40	*	?	fishing
4	41-and-over	higher	male	hunting
5	21–40	?	female	hunting
6	?	*	female	jogging
7	21–40	secondary	female	jogging
8	under-21	elementary	*	jogging

In the experiments reported in this paper we distinguish between two interpretations of missing attribute values: lost values and "do not care" conditions. We assume that lost values were erased and are not accessible since an operator forgot to enter the value, etc. We induce rules only from existing, known attribute values. "Do not care" conditions are interpreted as a result of a refusal to answer a question since the question is considered to be irrational, embarrassing, etc. In this case we assume that such a missing attribute values may be replaced by any existing, known attribute value.

The main objective of this paper was to compare global and local approximations in terms of the error rate, computed as a result of ten-fold cross validation. For our experiments, we used six incomplete data sets with as many missing attribute values as possible, i.e., any additional incremental replacement of known values by missing attribute values resulted in the entire records filled with only missing attribute values. Our next objective was to check the quality of non-standard probabilistic approximations, i.e., probabilistic approximations that were neither lower nor upper approximations. Our final objective was to compare two interpretations of missing attribute values, lost values and "do not care" conditions.

II. CHARACTERISTIC SETS

We assume that the input data sets are presented in the form of a *decision table*. An example of a decision table is shown in Table I. Rows of the decision table represent *cases*, while columns are labeled by *variables*. The set of all cases is denoted by U. In Table I, 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 Table I, A = {*Age*, *Education*, *Gender*} and d = *Hobby*.

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.

- 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 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.

A special block of a decision-value pair is called a *concept*. In Table I, [(*Hobby*, *fishing*)] = {1, 2, 3}. Additionally, for Table I

[(*Age*, *under*-21)] = {1, 8},
[(*Age*, 21–40)] = {3, 5, 7},
[(*Age*, 41-*and-over*)] = {4},
[(*Education*, *elementary*)] = {1, 3, 6, 8},
[(*Education*, *secondary*)] = {2, 3, 6, 7},
[(*Education*, *higher*)] = {3, 4, 6},
[(*Gender*, *female*)] = {2, 5, 6, 7, 8},
[(*Gender*, *male*)] = {1, 4, 8}.

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 $K(x, a) = U$.

Characteristic set $K_B(x)$ may be interpreted as the set of cases that are indistinguishable from x using all attributes from B and using a given interpretation of missing attribute values. For Table I and $B = A$,

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

III. GLOBAL PROBABILISTIC APPROXIMATIONS

There are three different global probabilistic approximations, called *singleton*, *subset* and *concept*, as discussed, e.g., in [12]. For our experiments we used only concept probabilistic approximations since it has been proved that the are the best approaches among global probabilistic approximations to handle missing attribute values [7].

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

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

where $Pr(X|K_A(x)) = \frac{|X \cap K_A(x)|}{|K_A(x)|}$ is the conditional probability of X given $K_A(x)$.

For Table I, all distinct concept probabilistic approximations for [(*Hobby*, *fishing*)] = {1, 2, 3} are

$$appr_{0.333}^{concept}(\{1,2,3\}) = \{1,2,3,5,6,7,8\},$$

$$appr_{0.5}^{concept}(\{1,2,3\}) = \{1,8\} \text{ and}$$

$$appr_{1}^{concept}(\{1,2,3\}) = \emptyset.$$

Note that if, for a given set X, a global probabilistic approximation of X, associated with β, $0 < \beta \leq 1$, not listed above, it is equal to the closest global probabilistic approximation of the same type with α larger than or equal to β. For example, $appr_{0.4}^{concept}(\{1,2,3\})$ = $appr_{0.5}^{concept}(\{1,2,3\})$ = {1, 8}. Additionally, if α is small but greater than 0 (in our experiments such α was equal to 0.001), a probabilistic approximation is called *upper* [12]. For α = 1, a probabilistic approximation is called *lower* [12].

IV. Local Probabilistic Approximations

A set T of attribute-value pairs, where all attributes belong to the set A of all attributes and are distinct, will be called a *complex*. We will discuss only *nontrivial complexes*, i.e., such complexes that the intersection of all attribute-value blocks from a given complex is not the empty set. A block of the complex T, denoted by $[T]$, is a subset of the set U, defined as the set $\cap\{[t] \mid t \in T\}$.

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

$$\cup\{[T] \mid \exists \text{ a family } \mathcal{T} \text{ of complexes } T \text{ of } X, T \in \mathcal{T}, Pr(X|[T]) \geq \alpha\}.$$

For a given set X and parameter α, in general, there exist many local probabilistic approximations. Additionally, searching for all local probabilistic approximations is a process of exponential complexity [14]. Therefore, in our experiments we used a heuristic approach to computing another local probabilistic approximation, denoted by $appr_{\alpha}^{mlem2}(X)$, since it uses ideas of the MLEM2 rule induction algorithm [17]. Using this approach, $appr_{\alpha}^{mlem2}(X)$ is constructed from the complexes T that are the most relevant to X, i.e., with $|X \cap [T]|$ as large as possible, if there is more than one complex that satisfies this criterion, the largest conditional probability of X given $[T]$ is the next criterion to select a complex. Note that if two complexes are equally relevant, then the second criterion selects a complex with the smaller cardinality of $[T]$. This criterion selects the complex with the larger conditional probability of X given T.

To be more specific, a local probabilistic approximation $appr_{\alpha}^{mlem2}(X)$ is defined by the following algorithm that computes not only the approximation $appr_{\alpha}^{local}(X)$ but also the corresponding rule set [14]. The rule set is represented by the family $\mathcal{T}$ of complexes T, where every T corresponds to a rule. The local probabilistic approximation $appr_{\alpha}^{mlem2}(X)$ is defined as $\cup\{[T] \mid T \in \mathcal{T}\}$.

Algorithm for determining
a single local probabilistic approximation
input: a set X (a subset of U) and a parameter α,
output: a set $\mathcal{T}$ of the set X,
begin
  $G := X$;
  $D := X$;
  $\mathcal{T} := \emptyset$;
  $\mathcal{J} := \emptyset$;
  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 $t = (a_t, v_t) \in T(G)$ such that $|[t] \cap G|$ is maximum; if a tie occurs, select a pair $t \in T(G)$ with the smallest cardinality of $[t]$; if another tie occurs, select 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
              $T_s := T_s \cup \{(a_t, v) \mid v \in V_{a_t}\}$
            else {a_t is numerical, let $t = (a_t, u..v)$}
              $T_n := T_n \cup \{(a_t, x..y) \mid$ 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 $Pr(X \mid [T]) \geq \alpha$
        then
          begin
            $D := D \cup [T]$;
            $\mathcal{T} := \mathcal{T} \cup \{T\}$;
          end {then}
        else $\mathcal{J} := \mathcal{J} \cup \{T\}$;
      $G := D - \cup_{S \in \mathcal{T} \cup \mathcal{J}}[S]$;
  end {while};
for each $T \in \mathcal{T}$ **do**
  for each numerical attribute a_t with $(a_t, u..v) \in T$ **do**

Table II
DATA SETS USED FOR EXPERIMENTS

Data set	Number of cases	attributes	concepts	Percentage of missing attribute values
Breast cancer	277	9	2	44.81
Echocardiogram	74	7	2	40.15
Hepatitis	155	19	2	60.27
Image segmentation	210	19	7	69.85
Lymphography	148	18	4	69.89
Wine recognition	178	13	3	64.65

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 $t \in T$ **do**
if $[T - \{t\}] \subseteq D$ **then** $T := T - \{t\}$;
for each $T \in \mathcal{T}$ **do**
if $\cup_{S \in (\mathcal{T} - \{T\})}[S] = \cup_{S \in \mathcal{T}}[S]$ **then** $\mathcal{T} := \mathcal{T} - \{T\}$;
end {procedure}.

For Table I, all distinct local probabilistic approximations of the type $appr_{\alpha}^{mlem2}(X)$, for [(*Hobby*, *fishing*)] = {1, 2, 3}, are

$$appr_{1}^{mlem2}(\{1,2,3\}) = \emptyset,$$

$$appr_{0.5}^{mlem2}(\{1,2,3\}) = \{1,3,7,8\},$$

$$appr_{0.333}^{mlem2}(\{1,2,3\}) = \{1,2,3,6,7,8\},$$

V. EXPERIMENTS

For our experiments we used six real-life data sets that are available on the University of California at Irvine *Machine learning Repository*, see Table II. For every data set a set of templates was created. Templates were formed by replacing incrementally (with 5% increment) existing specified attribute values by *lost* values. Thus, we started each series of experiments with no *lost* values, then we added 5% of *lost* values, then we added additional 5% of *lost* values, etc., until at least one entire row of the data sets was full of *lost* values. Then three attempts were made to change configuration of new *lost* values and either a new data set with extra 5% of *lost* values were created or the process was terminated. Additionally, the same formed templates were edited for further experiments by replacing question marks, representing *lost* values by "*"s, representing *"do not care"* conditions.

For any data set there was some maximum for the percentage of missing attribute values. For example, for the *Breast cancer* data set, it was 44.81%.

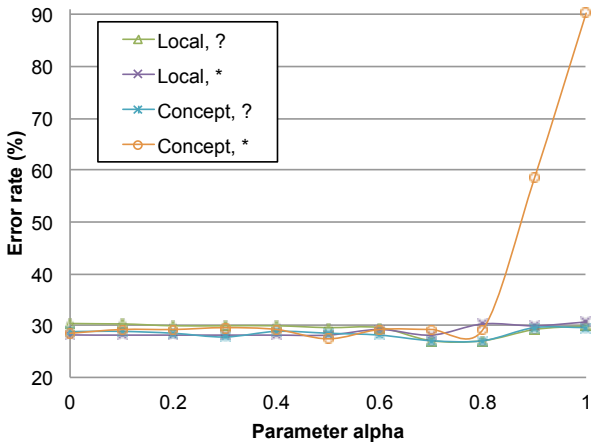

Figure 1. Results of experiments with *Breast cancer*data set

Rule sets were induced from data sets with many missing attribute values using two different algorithms, based on global and local probabilistic approximations. For rule induction using global probabilistic approximations we used the methodology explained in [12], [13]. Using this approach, rule sets were induced from modified data sets. For a given data set, for any concept, a modified data set was constructed in which for all cases from a global concept approximation of the concept the decision values were the same as in the original data set. For all remaining cases the decision values were special (non-existing in the original data set). Only rules indicating concept were saved, all remaining rules, with special values on the right-hand side, were deleted. A union of such rule sets was the final rule set. For rule induction we used the Modified Learning from Examples Module version 2 (MLEM2) rule induction algorithm, a component of the Learning from Examples based on Rough Sets (LERS) data mining system [8], [9].

An algorithm for rule induction using local probabilistic approximations was explained in Section IV.

Results of ten-fold cross validation are presented in Figures 1–6. In these figures the global concept probabilistic approximation is denoted by "Concept". In the same figures, local probabilistic approximations are denoted by "Local". Additionally, both methods may be applied to data sets with lost values, denoted by "?" and by "do not care" conditions denoted by "*".

VI. CONCLUSIONS

Our main objective was to compare global and local probabilistic approximations in mining data with many missing attribute values. In two cases (*hepatitis* and *lymphography* data sets) the best results were accomplished by global approach, in one case the winner was a local approach (*echocardiogram* data set), in remaining three cases there was a tie.

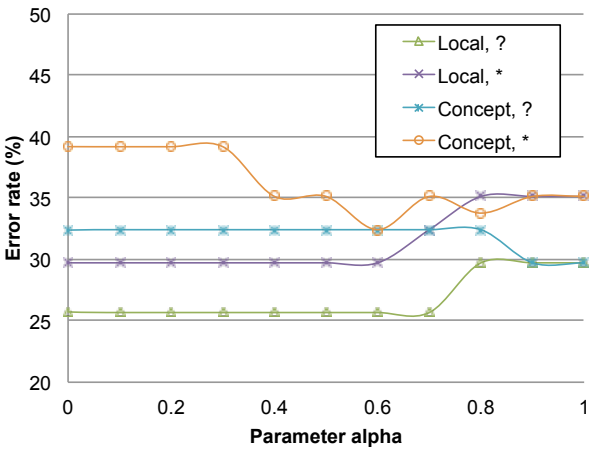

Figure 2. Results of experiments with *Echocardiogram* data set

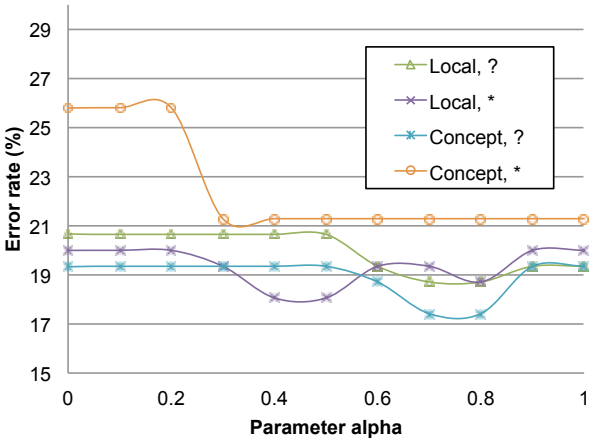

Figure 3. Results of experiments with *Hepatitis* data set

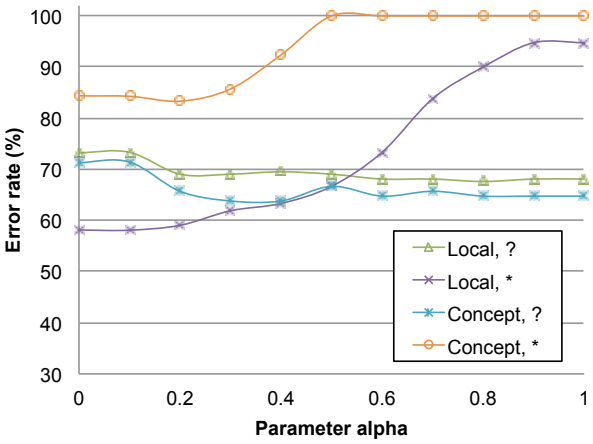

Figure 4. Results of experiments with *Image Segmentation* data set

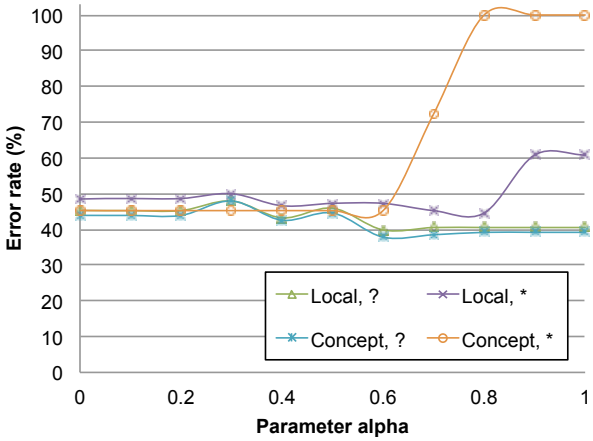

Figure 5. Results of experiments with *Lymphography* data set

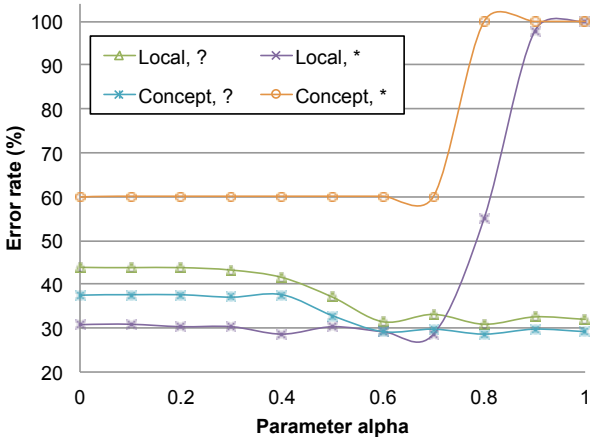

Figure 6. Results of experiments with *Wine recognition* data set

In our experiments, for four data sets the best results (the smallest error rate) were accomplished by non-standard probabilistic approximations, i.e., probabilistic approximations that were neither lower nor upper approximations. In two cases (*echocardiogram* and *image segmentation* data sets) the same results could be accomplished by upper approximations. An error rate may increase for some probabilistic approximations different from lower and upper approximations. For example, for *lymphography* data set with lost values, for the global approach the error rate was 39.19% for lower approximation, 43.92% for upper approximation, and 47.97% for the probabilistic approximation with α = 0.3.

The next problem was what the better interpretation of missing attribute values is: lost values or "do not care" conditions. In three cases (*echocardiogram*, *hepatitis* and *lymphography* data sets) the best approach was a lost value interpretation, in one case (*wine recognition* data set) it was

a "do not care" condition, in remaining two cases there was a tie.

Data sets with many "do not care" conditions and α close to 1 may cause the error rate for lower approximation to increase up to 100% due to so large characteristic sets that the corresponding probabilistic approximations are empty.

For a given data set, choice for the interpretation of the missing attribute value type and the approximation type should be selected individually.

REFERENCES

[1] 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.

[2] ——, "Experiments on rule induction from incomplete data using three probabilistic approximations," in *Proceedings of the 2012 IEEE International Conference on Granular Computing*, 2012, pp. 90–95.

[3] ——, "Experiments using three probabilistic approximations for rule induction from incomplete data sets," in *Proceeedings of the MCCSIS 2012, IADIS European Conference on Data Mining ECDM 2012*, 2012, pp. 72–78.

[4] ——, "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*, 2012, pp. 235–242.

[5] P. G. Clark, J. W. Grzymala-Busse, and Z. S. Hippe, "How good are probabilistic approximations for rule induction from data with missing attribute values?" in *Proceedings of the RSCTC 2012, the 8-th International Conference on Rough Sets and Current Trends in Computing*, 2012, pp. 46–55.

[6] P. G. Clark, J. W. Grzymala-Busse, and M. Kuehnhausen, "Local probabilistic approximations for incomplete data," in *Proceedings of the ISMIS 2012, the 20-th International Symposium on Methodologies for Intelligent Systems*, 2012, pp. 93–98.

[7] ——, "Mining incomplete data with many missing attribute values. a comparison of probabilistic and rough set approaches," in *Proceedings of the INTELLI 2013, the Second International Conference on Intelligent Systems and Applications*, 2013, pp. 12–17.

[8] 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.

[9] ——, "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.

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

[11] ——, "Characteristic relations for incomplete data: A generalization of the indiscernibility relation," in *Proceedings of the Fourth International Conference on Rough Sets and Current Trends in Computing*, 2004, pp. 244–253.

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

[13] ——, "Generalized probabilistic approximations," *Transactions on Rough Sets*, vol. 16, pp. 1–16, 2013.

[14] J. W. Grzymala-Busse, P. G. Clark, and M. Kuehnhausen, "Generalized probabilistic approximations of incomplete data," *International Journal of Approximate Reasoning*, vol. in press, 2013.

[15] J. W. Grzymala-Busse and M. Hu, "A comparison of several approaches to missing attribute values in data mining," in *Proceedings of the Second International Conference on Rough Sets and Current Trends in Computing*, 2000, pp. 340–347.

[16] J. W. Grzymala-Busse and W. Rzasa, "Local and global approximations for incomplete data," in *Proceedings of the RSCTC 2006, the Fifth International Conference on Rough Sets and Current Trends in Computing*, 2006, pp. 244–253.

[17] ——, "A local version of the MLEM2 algorithm for rule induction," *Fundamenta Informaticae*, vol. 100, pp. 99–116, 2010.

[18] J. W. Grzymala-Busse and W. Ziarko, "Data mining based on rough sets," in *Data Mining: Opportunities and Challenges*, J. Wang, Ed. Hershey, PA: Idea Group Publ., 2003, pp. 142–173.

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

[20] ——, *Rough Sets. Theoretical Aspects of Reasoning about Data*. Dordrecht, Boston, London: Kluwer Academic Publishers, 1991.

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

[22] 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.

[23] S. K. M. Wong and W. Ziarko, "INFER—an adaptive decision support system based on the probabilistic approximate classification," in *Proceedings of the 6-th International Workshop on Expert Systems and their Applications*, 1986, pp. 713–726.

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

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