A Comparison of Classification Systems for Rule Sets Induced from Incomplete Data by Probabilistic Approximations

Patrick G. Clark

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

Jerzy W. Grzymala-Busse

Department of Electrical Eng. and Computer Sci.
University of Kansas, Lawrence, KS, USA
Department of Expert Systems and Artificial Intelligence
University of Information Technology and Management,
Rzeszow, Poland
e-mail: jerzy@ku.edu

***Abstract*—In this paper, we compare four strategies used in classification systems. A classification system applies a rule set, induced from the training data set in order to classify each testing case as a member of one of the concepts. We assume that both training and testing data sets are incomplete, i.e., some attribute values are missing. In this paper, we discuss two interpretations of missing attribute values: lost values and "do not care" conditions. In our experiments rule sets were induced using probabilistic approximations. Our main results are that for lost value data sets the strength only strategy is better than conditional probability without support and that for "do not care" data sets the conditional probability with support strategy is better than strength only.**

***Index Terms*—Data mining; rough set theory; probabilistic approximations; MLEM2 rule induction algorithm; lost values and "do not care" conditions.**

I. Introduction

In this paper, we investigated the correctness of rule sets evaluated by the error rate, a result of ten-fold cross validation, with a focus on the choice of classification strategy. For a given rule set and testing data set the question is what is the best strategy for the classification system. In our experiments we used the Learning from Examples using Rough Sets (LERS) data mining system [1]–[3] with which we may use four strategies: strength of a rule combined with support, strength only, a conditional probability of the concept given the set of all training cases the rule matches combined with support, and the conditional probability, without any support.

In Sections 2 and 3, background material on incomplete data and probabilistic approximations are covered. Section 4 introduces and explains the four classification strategies used during the experiments described in Section 5. In Section 6, conclusions are discussed with the main results being that for the data sets with lost values the strategy based on strength only is better than conditional probability without support. For data sets with "do not care" conditions the strategy based on conditional probability with support is better than the strategy based on strength only.

TABLE I
TRAINING DATA SET

	Attributes			Decision
Case	Temperature	Headache	Cough	Flu
1	high	no	no	no
2	very-high	yes	*	no
3	normal	*	no	no
4	normal	no	*	no
5	?	?	yes	yes
6	very-high	yes	no	yes
7	*	yes	?	yes
8	high	yes	*	yes

II. Incomplete Data

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 will be denoted by U. In Table I, U = $\{1, 2, 3, 4, 5, 6, 7, 8\}$. Independent variables are called attributes and a dependent variable is called a decision and is denoted by d. The set of all attributes will be denoted by A. In Table I, A = $\{$*Temperature*, *Headache*, *Cough*$\}$. The value for a case x and an attribute a will be denoted by $a(x)$.

In this paper, we distinguish between two interpretations of missing attribute values: lost values and attribute-concept values. Lost values, denoted by "?", mean that the original attribute value is no longer accessible and that during rule induction we will only use existing attribute values [4][5]. "Do not care" conditions (denoted by $*$) correspond to a refusal to answer a question. With a "do not care" condition interpretation we will replace the missing attribute value by all possible attribute values. The error rate does not differ significantly for both interpretations of missing attribute values [6].

One of the most important ideas of rough set theory [7] is

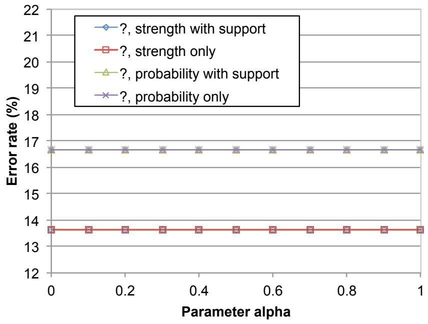

Fig. 1. Error rate for the *Bankruptcy* data set with lost values with lost values

an indiscernibility relation, defined for complete data sets. Let B be a nonempty subset of A. The indiscernibility relation $R(B)$ is a relation on U defined for $x, y \in U$ as defined in equation 1.

$$(x, y) \in R(B) \text{ if and only if } \forall a \in B \ (a(x) = a(y)) \quad (1)$$

The indiscernibility relation $R(B)$ is an equivalence relation. Equivalence classes of $R(B)$ are called *elementary sets* of B and are denoted by $[x]_B$. A subset of U is called *B-definable* if it is a union of elementary sets of B.

The set X of all cases defined by the same value of the decision d is called a *concept*. For example, a concept associated with the value *yes* of the decision *Flu* is the set {5, 6, 7, 8}. The largest B-definable set contained in X is called the *B-lower approximation* of X, denoted by $\underline{appr}_B(X)$, and defined in equation 2.

$$\cup\{[x]_B \mid [x]_B \subseteq X\} \quad (2)$$

The smallest B-definable set containing X, denoted by $\overline{appr}_B(X)$ is called the *B-upper approximation* of X, and is defined in equation 3.

$$\cup\{[x]_B \mid [x]_B \cap X \neq \emptyset\} \quad (3)$$

For a variable a and its value v, (a, v) is called a variable-value pair. A *block* of (a, v), denoted by $[(a, v)]$, is the set $\{x \in U \mid a(x) = v\}$ [8].

For incomplete decision tables the definition of a block of an attribute-value pair is modified in the following way.

- If for an attribute a there exists a case x such that $a(x) = ?$, i.e., the corresponding value is lost, then the case x should not be included in any blocks $[(a, v)]$ for all values v of attribute a,
- If for an attribute a there exists a case x such that the corresponding value is 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.

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

[(Temperature, normal)] = {3, 4, 7},

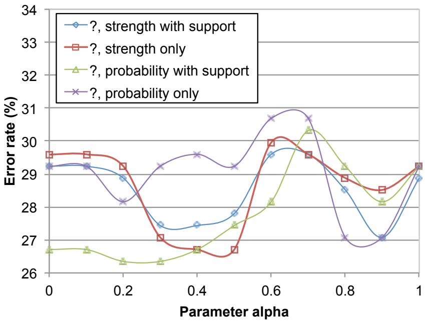

Fig. 2. Error rate for the rule set for the *Breast cancer* data set with lost values

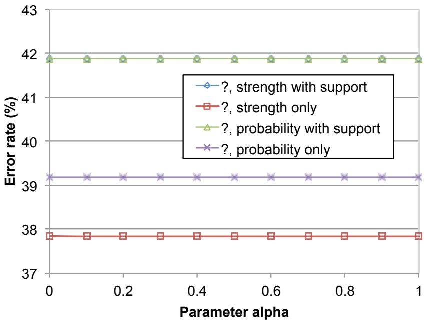

Fig. 3. Error rate for the rule set for the *Echocardiogram* data set with lost values

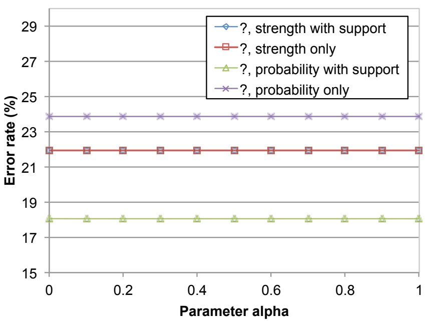

Fig. 4. Error rate for the rule set for the *Hepatitis* data set with lost values

[(Temperature, high)] = {1, 7, 8},
[(Temperature, very-high)] = {2, 6, 7},
[(Headache, no)] = {1, 3, 4},
[(Headache, yes)] = {2, 3, 6, 7, 8},
[(Cough, no)] = {1, 2, 3, 4, 6, 8}, and
[(Cough, yes)] = {2, 4, 5, 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 the set $K(x, a) = U$.

For Table I and $B = A$,

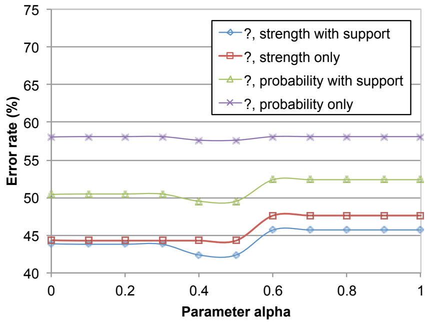

Fig. 5. Error rate for the rule set for the *Image segmentation* data set with lost values

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

Note that for incomplete data there are a few possible ways to define approximations [9], we used *concept* approximations since our previous experiments indicated that such approximations are most efficient [10]. A B-*concept lower approximation* of the concept X is defined in equation 4.

$$\underline{B}X = \cup\{K_B(x) \mid x \in X, K_B(x) \subseteq X\} \tag{4}$$

The *B-concept upper approximation* of the concept X is defined by the equation 5.

$$\begin{aligned}\overline{B}X &= \cup\{K_B(x) \mid x \in X, K_B(x) \cap X \neq \emptyset\} \\ &= \cup\{K_B(x) \mid x \in X\}\end{aligned} \tag{5}$$

For Table I, A-concept lower and A-concept upper approximations of the concept $\{5, 6, 7, 8\}$ are
$\underline{A}\{5, 6, 7, 8\} = \{7, 8\}$ and
$\overline{A}\{5, 6, 7, 8\} = \{2, 3, 4, 5, 6, 7, 8\}$, respectively.

III. Probabilistic approximations

For completely specified data sets a *probabilistic approximation* is defined by equation 6, where α is a parameter, $0 < \alpha \leq 1$, see [10]–[15]. Additionally, for simplicity, the elementary sets $[x]_A$ are denoted by $[x]$. For discussion on how this definition is related to the variable precision asymmetric rough sets see [1][10].

$$appr_\alpha(X) = \cup\{[x] \mid x \in U, P(X \mid [x]) \geq \alpha\}. \tag{6}$$

For incomplete data sets, a *B-concept probabilistic approximation* is defined by equation 7 [10].

$$\cup\{K_B(x) \mid x \in X,\ Pr(X|K_B(x)) \geq \alpha\} \tag{7}$$

Where $Pr(X \mid K_B(x)) = \frac{|X\ \cap\ K_B(x)|}{|K_B(x)|}$ is the conditional probability of X given $K_B(x)$ and $|Y|$ denotes the cardinality of set Y. Note that if $\alpha = 1$, the probabilistic approximation becomes the standard lower approximation and if α is small, close to 0, in our experiments it was 0.001, the same definition describes the standard upper approximation.

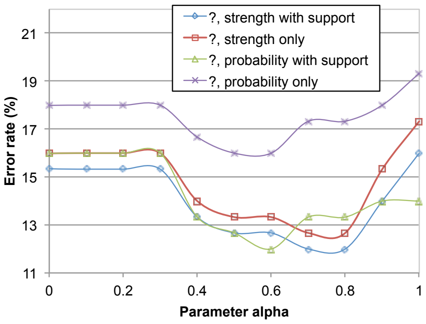

Fig. 6. Error rate for the rule set for the *Iris* data set with lost values

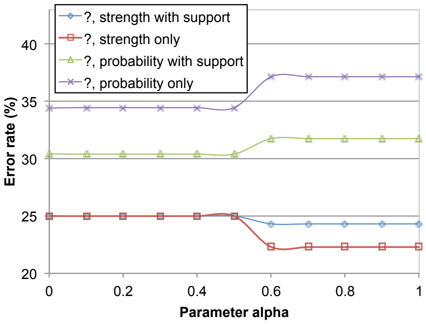

Fig. 7. Error rate for the rule set for the *Lymphography* data set with lost values

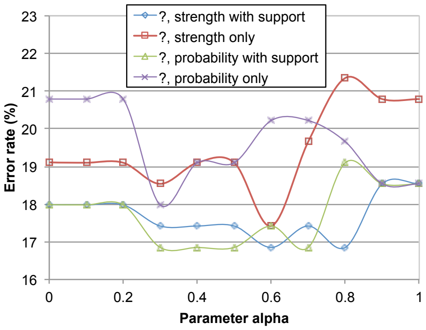

Fig. 8. Error rate for the rule set for the *Wine recognition* data set with lost values

For simplicity, we will denote $K_A(x)$ by $K(x)$ and the *A-concept probabilistic approximation* will be called a *probabilistic approximation*.

For Table I and the concept $X = \{5, 6, 7, 8\}$, there exist three distinct probabilistic approximations:
$appr_{1.0}(\{5, 6, 7, 8\}) = \{7, 8\}$
$appr_{0.6}(\{5, 6, 7, 8\}) = \{2, 3, 6, 7, 8\}$ and
$appr_{0.001}(\{5, 6, 7, 8\}) = \{2, 3, 4, 5, 6, 7, 8\}$.

IV. Classification

Rule sets, induced from data sets, are used most frequently to classify new, unseen cases. A *classification system* has two inputs: a rule set and a data set containing unseen cases. The classification system classifies every case as being member of some concept. A classification system used in LERS is a modification of the well-known bucket brigade algorithm [16]–[18].

The decision to which concept a case belongs is made on the basis of two factors: *strength* and *support*. *Strength* is the total number of cases correctly classified by the rule during training. The second factor, *support*, is defined as the sum of strengths for all matching rules indicating the same concept. The concept C for which the support, i.e., the following expression

$$\sum_{matching\ rules\ r\ describing\ C} Strength(r) \tag{8}$$

is the largest is the winner and the case is classified as being a member of C. This strategy is called *strength with support*. There exist three additional strategies. We may decide to which concept a case belongs on the basis of the strongest rule matching the case. This strategy will be called *strength only*. In the next strategy for every rule we compute ratios of the strength to the rule domain equal to the total number of cases matching the left-hand side of the rule. Such a ratio is a conditional probability of the concept given rule domain. A rule with the largest probability decides to which concept a case belongs. This strategy is called *probability only*. The fourth strategy, highly heuristic, in which all probabilities for rules indicating the same concept are added up is called *probability with support*.

In the classification system of LERS, if complete matching is impossible, all partially matching rules are identified. These are rules with at least one attribute-value pair matching the corresponding attribute-value pair of a case. For any partially matching rule r, the additional factor, called *Matching_factor* (r), is computed. Matching_factor (r) is defined as the ratio of the number of matched attribute-value pairs of r with a case to the total number of attribute-value pairs of r. In partial matching, the concept C for which the following expression

$$\sum_{\substack{partially\ matching\\ rules\ r\ describing\ C}} Strength(r) * Matching\_factor(r) \tag{9}$$

is the largest is the winner and the case is classified as being a member of C.

The problem is how to classify unseen cases with missing attribute values. In the LERS classification system, when an unseen case x is classified by a rule r, case x is considered to be not matched by r if for an attribute a, $a(x)$ = ? and the rule r contained a condition of the type (a, v), where v was a value of a. If for an attribute a, $a(x)$ = * and if the rule r contained a condition of the type (a, v), then case x is considered to be matched by r, does not matter what v is. In both cases interpretation of lost values and "do not care" conditions were strictly adhered to.

TABLE II
The best results for error rates (%)—experiments on data with *lost values*

Data set	Error rate (%) for			
	strength with support	strength only	probability with support	probability only
Bankruptcy	13.64	13.64	16.67	16.67
Breast cancer	27.08	26.71	26.35	27.08
Echocardiogram	41.89	37.84	41.89	39.19
Hepatitis	21.94	21.94	18.06	23.87
Image segmentation	42.38	44.29	49.52	57.62
Iris	12.00	12.67	13.33	16.00
Lymphography	24.32	22.30	30.41	34.46
Wine recognition	16.85	17.42	16.85	17.98

Using α = 0.333, the following rule set was induced by LERS from the data set from Table I

R1. (Headache, no) → (Flu, no), with strength = 3 and domain rule size = 3,

R2. (Temperature, very-high) → (Flu, no), with strength = 1 and domain rule size = 3,

R3. (Headache, yes) → (Flu, yes), with strength = 3 and domain rule size = 5, and

R4. (Cough, yes) → (Flu, yes), with strength = 2 and domain rule size = 4.

V. Experiments

Eight real-life data sets taken from the University of California at Irvine *Machine learning Repository* were used for experiments. Three of our data sets: *Bankruptcy*, *Echocardiogram* and *Iris* were numerical. All eight data sets were enhanced by replacing 35% of existing attribute values by missing attribute values, separately by lost values and by "do not care" conditions.

For all data sets there was a maximum value for the percentage of missing attribute values successfully replaced. In our experiments we chose the largest percentage common to all datasets, 35%, as it is the maximum percentage for the *bankruptcy* and *iris* data sets. As a result, 16 data sets were used, eight with 35% *lost values* and eight with 35% *"do not care" conditions*. Using the 16 data sets, experiments with 11 alpha values and four classification strategies were conducted, resulting in 704 ten-fold cross validation experiments.

Results of our experiments are presented as Figures 1 - 16 and Tables II and III. Results of experiments are presented in

TABLE III
THE BEST RESULTS FOR ERROR RATES (%)–EXPERIMENTS ON DATA WITH *"do not care" conditions*

Data set	Error rate (%) for			
	strength with support	strength only	probability with support	probability only
Bankruptcy	16.67	22.73	15.15	19.70
Breast cancer	28.16	28.88	27.08	27.80
Echocardiogram	24.32	27.03	27.03	28.38
Hepatitis	19.35	18.71	18.71	19.35
Image segmentation	47.14	51.43	46.19	49.52
Iris	36.00	38.67	28.67	25.33
Lymphography	24.32	26.35	25.00	31.76
Wine recognition	14.04	17.42	14.04	15.73

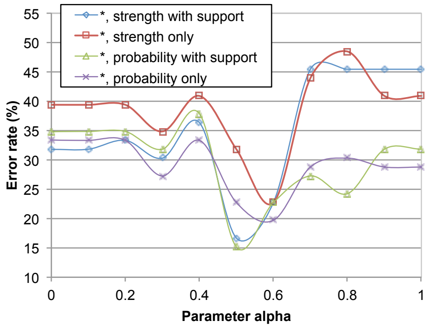

Fig. 9. Error rate for the *Bankruptcy* data set with "do not care" conditions

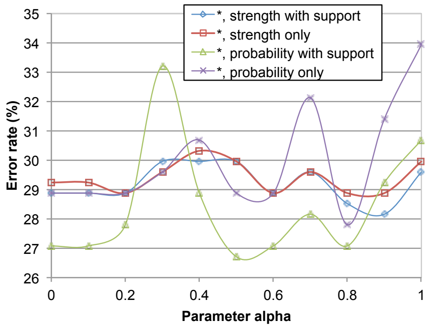

Fig. 10. Error rate for the rule set for the *Breast cancer* data set with "do not care" conditions

terms of error rate, a percentage of incorrectly classified cases when run in a 10-fold cross validation system.

In Tables II and III, the best results for all four strategies are shown. For each data set, strategy and interpretation of missing attribute values, we selected the smallest error rate from Figures 1 - 16. It is justified by practice of data mining, we always pick the value of the parameter α that corresponds to the smallest error rate. For example, for the *bankruptcy* data set, for two strategies, *strength with support* and *strength only*, for lost values, the error rate is 13.64%, so the corresponding entries in Table II are 13.64 (in this specific situation, the error rate does not depend on α).

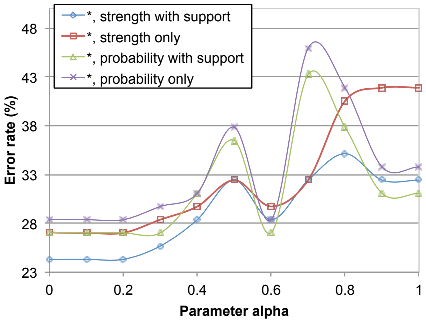

Fig. 11. Error rate for the rule set for the *Echocardiogram* data set with "do not care" conditions

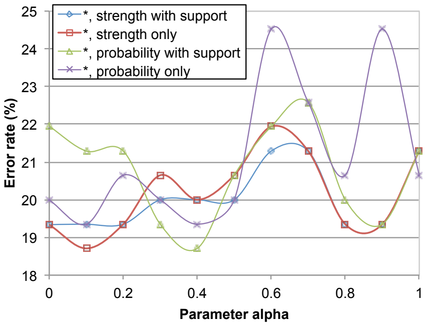

Fig. 12. Error rate for the rule set for the *Hepatitis* data set with "do not care" conditions

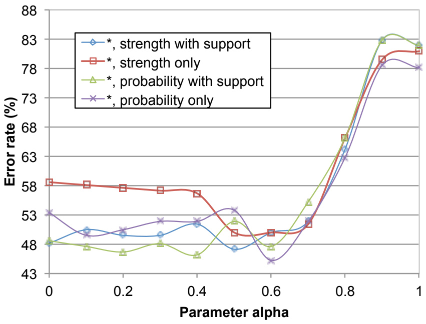

Fig. 13. Error rate for the rule set for the *Image segmentation* data set with "do not care" conditions

Surprisingly, the strategy *strength only* seems to be the best strategy for data with lost values while the same strategy looks like the worst strategy for data with "do not care" conditions.

The Friedman test (5% level of significance), ties were taken into account shows that for both Tables II and III the null hypothesis that all four strategies do not differ significantly with respect to error rate must be rejected. For post hoc analysis we used the distribution-free pairwise comparisons based on Friedman rank sums (5% level of significance). The only results are: for data sets with lost values, the strategy based on *strength only* is better than the strategy

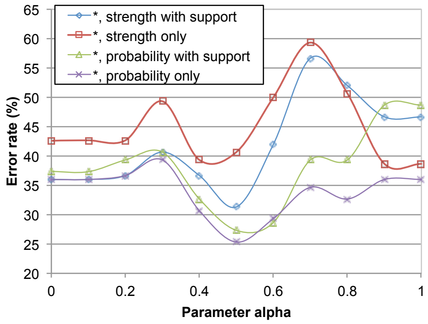

Fig. 14. Error rate for the rule set for the *Iris* data set with "do not care" conditions

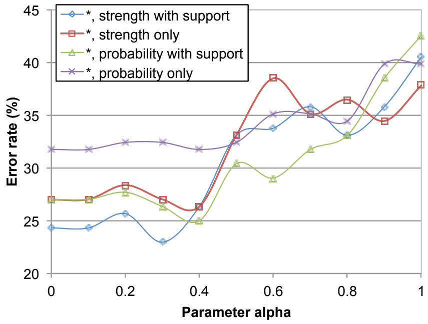

Fig. 15. Error rate for the rule set for the *Lymphography* data set with "do not care" conditions

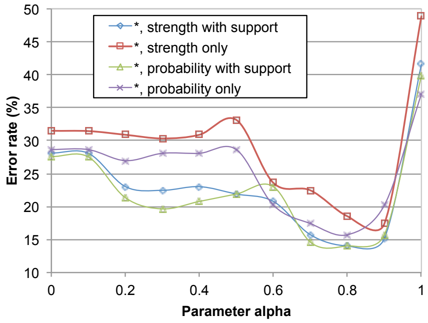

Fig. 16. Error rate for the rule set for the *Wine recognition* data set with "do not care" conditions

based on *probability only*, for data sets with "do not care" conditions, the strategy based on *probability and support* is significantly better than the strategy based on *strength only*. For other strategies differences are not statistically significant. For example, as follows from Table II, for data with lost values, the strategy based on *strength with support* is in most cases better than the strategy based on *probability only*, but that difference is not statistically significant.

VI. Conclusions

In this paper we report results of experiments on four different strategies of classification: *strength with support*, *strength only*, *probability with support* and *probability only* used for classification incomplete data by rule sets induced from incomplete data using probabilistic approximations.

Our main result is that for the data sets with lost values the strategy based on strength only is better than conditional probability without support. For data sets with "do not care" conditions the strategy based on conditional probability with support is better than the strategy based on strength only.

Additionally, results of our experiments show that for any given incomplete data set all four strategies should be applied and the best strategy should be selected as a result of ten-fold cross validation.

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

[3] ——, "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.

[4] J. W. Grzymala-Busse and A. Y. Wang, "Modified algorithms LEM1 and LEM2 for rule induction from data with missing attribute values," in *Proceedings of the Fifth International Workshop on Rough Sets and Soft Computing (RSSC'97) at the Third Joint Conference on Information Sciences (JCIS'97)*, 1997, pp. 69–72.

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

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

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

[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] ——, "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.

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

[11] 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.

[12] 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.

[13] 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.

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

[15] W. Ziarko, "Variable precision rough set model," *Journal of Computer and System Sciences*, vol. 46, no. 1, pp. 39–59, 1993.

[16] L. B. Booker, D. E. Goldberg, and J. F. Holland, "Classifier systems and genetic algorithms," in *Machine Learning. Paradigms and Methods*, J. G. Carbonell, Ed. Boston: MIT Press, 1990, pp. 235–282.

[17] J. H. Holland, K. J. Holyoak, and R. E. Nisbett, *Induction. Processes of Inference, Learning, and Discovery*. Boston: MIT Press, 1986.

[18] J. Stefanowski, *Algorithms of Decision Rule Induction in Data Mining*. Poznan, Poland: Poznan University of Technology Press, 2001.