Mining Incomplete Data with Many Missing Attribute Values

A Comparison of Probabilistic and Rough Set Approaches

Patrick G. Clark, Jerzy W. Grzymala-Busse, and Martin Kuehnhausen
Department of Electrical Engineering and Computer Science
University of Kansas
Lawrence, KS 66045, USA
E-mail: pclark@ku.edu, jerzy@ku.edu, mkuehnhausen@ku.edu

***Abstract*—In this paper, we study probabilistic and rough set approaches to missing attribute values. Probabilistic approaches are based on imputation, a missing attribute value is replaced either by the most probable known attribute value or by the most probable attribute value restricted to a concept. In this paper, in a rough set approach to missing attribute values we consider two interpretations of such value: *lost* and *"do not care"*. Additionally, we apply three definitions of approximations (*singleton*, *subset* and *concept*) and use an additional parameter called α. Our main objective was to compare probabilistic and rough set approaches to missing attribute values for incomplete data sets with many missing attribute values. We conducted experiments on six incomplete data sets with as many missing attribute values as possible. In these data sets an additional incremental replacement of known values by missing attribute values resulted with the entire records filled with only missing attribute values. Rough set approaches were better for five data sets, for one data set probabilistic approach was more successful.**

***Keywords*-Data mining; probabilistic approaches to missing attribute values; rough set theory; probabilistic approximations; parameterized approximations**

I. Introduction

In this paper, we compare two methods handling missing attribute values based on probability theory with a rough set approach to missing attribute values represented by two interpretations of missing attribute values (*lost* and *"do not care"*), three definitions of approximations (*singleton*, *subset* and *concept*) and on a parameter called α.

In probabilistic methods for missing attribute values, the most frequently used in data mining practice, in our first method called Most Common Value for Symbolic Attributes and Average Value for Numerical Attributes (MCV-AV), for symbolic attributes a missing attribute value was replaced by the most probable known attribute value (the most frequent). For numerical attributes, a missing attribute value was replaced by the average of known attribute values. In the second probabilistic method, called Concept Most Common Value for Symbolic Attributes and Concept Average Values for Numerical Attributes (CMCV-CAV), for symbolic attributes a missing attribute value was replaced by the most common value restricted to the concept that contains the missing attribute value. A *concept* is the set of all cases (records) with the same decision value (labeled the same way by an expert). Thus, for symbolic attributes a missing attribute value was replaced by a known attribute value with the largest conditional probability given the concept to which the case belongs. For numerical attributes, a missing attribute value was replaced the average of known attribute values restricted to the corresponding concept.

Using a rough set approach to missing attribute values, we may distinguish two interpretations of missing attribute values: *lost* and *"do not care"*. The former interpretation means that an attribute value was originally given, however, currently we have no access to it (e.g., the value was forgotten or erased). For data sets with *lost* values we try to induce the rule set from known data. The latter interpretation represents, e.g., a refusal to answer a question. For example, patients suspected of having flu may refuse to tell the value of the attribute *Eye color* since they may consider it irrelevant. For data mining from data sets affected by such missing attribute values we replace a *"do not care"* condition by all possible attribute values.

An idea of lower and upper approximations is a basic idea of rough set theory [1], [2]. For incomplete data sets there exist many definitions of approximations. In this paper, we use three types of approximations: singleton, subset and concept [3]. A probabilistic (or parameterized) approximation, associated with a probability (parameter) α, is a generalization of lower and upper approximations. For very small α, the probabilistic approximation becomes the upper approximation. For α = 1, the probabilistic approximation is a lower approximation [4]. Probabilistic approximations for complete data sets were studied for years, the idea was introduced in [5] and further developed in [6]–[15]. Such approximations were explored from a theoretical view point. The first paper on experimental validation of such approximations, for complete data sets, was [16]. For incomplete data sets probabilistic approximations were generalized in [4]. Results of experiments on probabilistic approximations for incomplete data sets were presented in [16]–[21]. In all of these papers, rough set approaches to mining incomplete data were not compared with any other approaches to missing attribute values.

 ISBN: 978-1-61208-269-1

TABLE I. An incomplete decision table

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

The main objective of this paper was to compare (experimentally) probabilistic and rough set approaches to missing attribute values for incomplete data sets with many missing attribute values. Our main result is that a rough set approach to missing attribute values was successful on five out of six data sets since it provides smaller error rates, a result of ten-fold cross validation. Our data sets had as many missing attribute values as possible. With an additional incremental replacement of known values by missing attribute values, the entire records were filled with only missing attribute values.

Section II describes the formal foundation of characteristic sets that form the basis of the probabilistic approximations we explore in Section III. Furthermore, we consider the problem of definability and its relationship to probabilistic approximations in Section IV. The results of our experiments are analyzed in Section V and further examined in the conclusion of this paper.

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 (a similar table was presented in [22]). 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}. 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 = {*Temperature*, *Headache*, *Cough*} and d = *Flu*.

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, [(Flu, yes)] = {1, 2, 3, 4, 5, 6}. Additionally, for Table I

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

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. Thus, $K_A(x)$ is the set of all cases that cannot be distinguished from x using all attributes.

For Table I and $B = A$,

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

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

$(x, y) \in R(B)$ *if and only if* $y \in K_B(x)$.

The characteristic relation is reflexive but—in general—it does not need to be symmetric or transitive. In our example, $R(A)$ = {(1, 1), (1, 7), (2, 2), (2, 7), (3, 3), (4, 4), (4, 7), (5, 4), (5, 5), (5, 7), (6, 6), (6, 7), (7, 2), (7, 4), (7, 6), (7, 7)}. For Table I the relation $R(A)$ is neither symmetric nor transitive. A relation $R(A)$ that is an equivalence relation is called an *indiscernibility* relation [1], [2].

III. Probabilistic Approximations

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

$$\{x \mid x \in U,\ Pr(X|K_B(x)) \geq \alpha\}, \quad (1)$$

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

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

$$\cup\{K_B(x) \mid x \in U,\ Pr(X|K_B(x)) \geq \alpha\}. \quad (2)$$

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

$$\cup\{K_B(x) \mid x \in X,\ Pr(X|K_B(x)) \geq \alpha\}. \quad (3)$$

For Table I, all distinct probabilistic approximations (singleton, subset and concept) for [(Flu, yes)] are

$appr_{0.5}^{singleton}(\{1,2,3,4,5,6\}) = U$,

$appr_{0.667}^{singleton}(\{1,2,3,4,5,6\}) = \{3,5,7\}$,

$appr_{0.75}^{singleton}(\{1,2,3,4,5,6\}) = \{3,7\}$,

$appr_{1}^{singleton}(\{1,2,3,4,5,6\}) = \{3\}$,

$appr_{0.5}^{subset}(\{1,2,3,4,5,6\}) = U$,

$appr_{0.667}^{subset}(\{1,2,3,4,5,6\}) = \{2,3,4,5,6,7\}$,

$appr_{0.75}^{subset}(\{1,2,3,4,5,6\}) = \{2,3,4,6,7\}$,

$appr_{1}^{subset}(\{1,2,3,4,5,6\}) = \{3\}$,

$appr_{0.5}^{concept}(\{1,2,3,4,5,6\}) = U$,

$appr_{0.667}^{concept}(\{1,2,3,4,5,6\}) = \{3,4,5,7\}$,

$appr_{0.75}^{concept}(\{1,2,3,4,5,6\}) = \{3\}$,

As follows from our example, all three probabilistic approximations, in general, are distinct, even for the same value of the parameter α. If a characteristic relation $R(B)$ is an equivalence relation, all three types of probabilistic approximation: singleton, subset and concept are reduced to the same probabilistic approximation. 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* [4]. For α = 1, a probabilistic approximation is called *lower* [4].

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

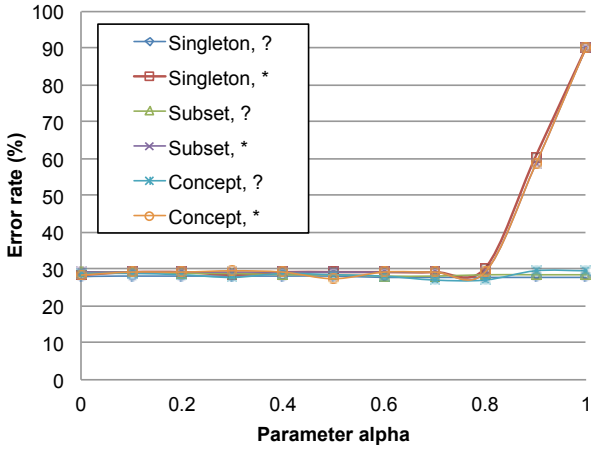

Figure 1. Results of experiments with *Breast cancer* data set

IV. Definability

Let B be a subset of the set A of all attributes. For incomplete decision tables, a union of some intersections of attribute-value pair blocks, where such attributes are members of B and are distinct, will be called B-*locally definable* sets. A union of characteristic sets $K_B(x)$, where $x \in X \subseteq U$ will be called a B-*globally definable* set. Any set X that is *B*-globally definable is *B*-locally definable, the converse is not true [4].

Singleton probabilistic approximations—in general—are not even locally definable. For example, the singleton probabilistic approximation of [(Flu, yes)] with the threshold α = 0.667, i.e., the set {3, 5, 7}, is not A-locally definable. Indeed, in all attribute blocks which contain the case 5 (there are two such blocks, [(Temperature, high)] and [(Cough, yes)], 5 is in the same block with cases 4 and 7. Thus, any definable case containing 5 must contain cases 4 and 7. The set {3, 5, 7} does not contain case 4.

On the other hand, both subset and concept probabilistic approximations are A-globally definable. Obviously, if a set is not B-locally definable then it cannot be expressed by rule sets using attributes from B.

 ISBN: 978-1-61208-269-1

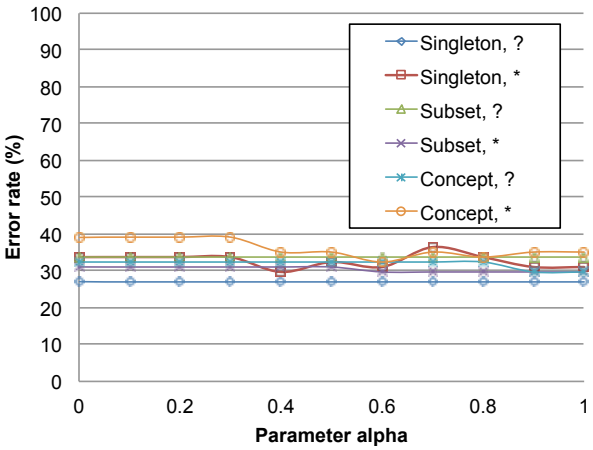

Figure 2. Results of experiments with *Echocardiogram* data set

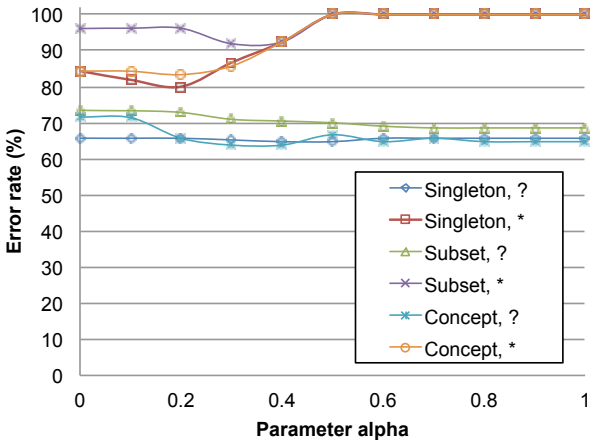

Figure 4. Results of experiments with *Image segmentation* data set

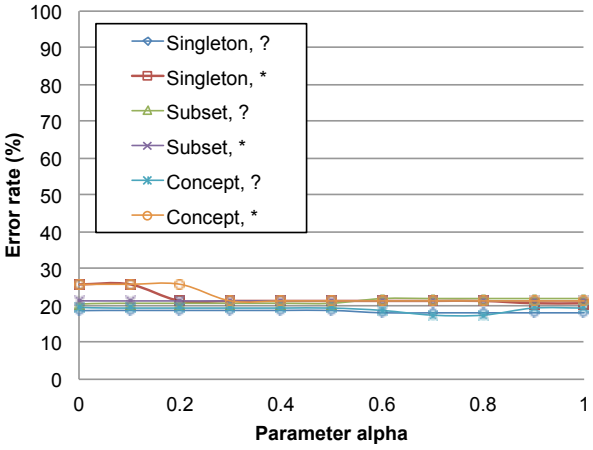

Figure 3. Results of experiments with *Hepatitis* data set

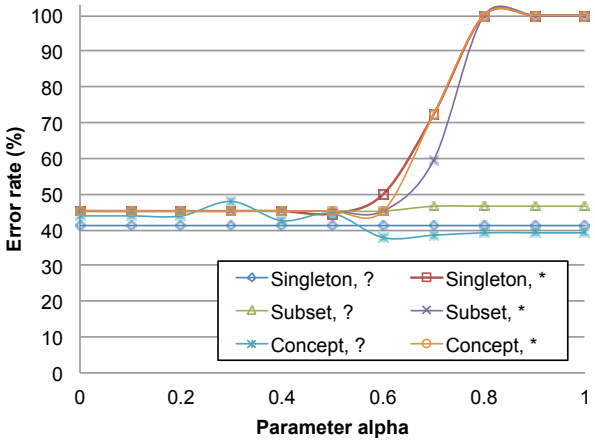

Figure 5. Results of experiments with *Lymphography* data set

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%.

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 [23], [24].

VI. Conclusions

A comparison of two probabilistic approaches to missing attribute values is presented on Table III. This table shows that for incomplete data with many missing attribute values, the MCV-AV method is better for four data sets (*Echocardiogram*, *Hepatitis*, *Lymphography* and *Wine recognition*) while the CMCV-CAV method is better for two data sets (*Breast cancer* and *Image segmentation*).

 ISBN: 978-1-61208-269-1

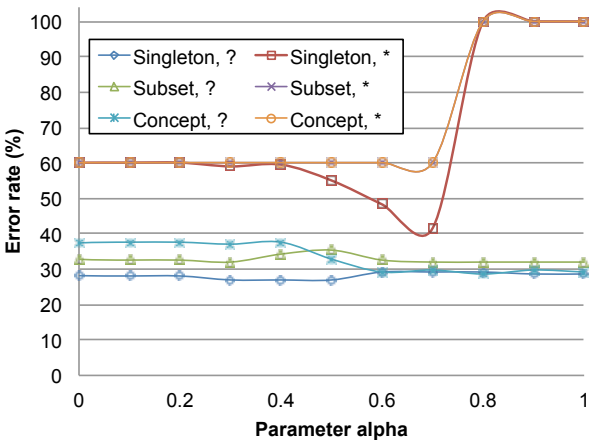

Figure 6. Results of experiments with *Wine recognition* data set

TABLE III. Error rate for MCV-AV and CMCV-CAV

Data set	Error rate in percent	
	MCV-AV	CMCV-CAV
Breast cancer	30.69	29.96
Echocardiogram	22.97	27.03
Hepatitis	19.35	21.29
Image segmentation	72.86	64.76
Lymphography	41.22	47.97
Wine recognition	31.46	44.38

Similarly, Figures 1–6 show that for four data sets (*Breast cancer*, *Hepatitis*, *Image segmentation* and *Lymphography* the best rough set approach is based on concept approximations. For remaining two data sets (*Echocardiogram* and *Wine recognition* the best rough set approach is based on singleton approximations. For all six data sets the smallest error rate is always associated with missing attribute values interpreted as *lost*. Note that with the exception of the *Echocardiogram* data set (where the best value is associated with any alpha, between 0.001 and 1), for remaining five data sets the smallest error rate occurs always for some alpha larger than 0.001 and smaller than 1. Moreover, for the parameter α close to one, error rate associated with *"do not care"* conditions is close to 100% due to small probabilistic approximations, as exemplified by *Breast cancer*, *Image segmentation* and *Wine recognition* data sets.

TABLE IV. Error rate for the better of MCV-AV and CMCV-CAV and the best rough set approach

Data set	Error rate in percent	
	better of MCV-AV and CMCV-CAV	best rough set approach
Breast cancer	29.96	27.08
Echocardiogram	22.97	27.03
Hepatitis	19.35	17.42
Image segmentation	64.76	63.81
Lymphography	41.22	37.84
Wine recognition	31.46	26.97

Finally, a comparison of the best of the two probabilistic approaches to missing attribute values (MCV-AV and CMCV-CAV) is presented in Table IV. This table shows the best results accomplished using rough set theory with three parameters: kind of approximation (singleton, subset and concept), two interpretations of missing attribute values (*lost* and *"do not care"* condition) and α (0.001, 0.1, 0.2,..., 1). The better results for five of the data sets is the rough set approach, while for only one data set (*Echocardiogram*) the better approach is achieved by the MCV-AV method. However, statistically, the superiority of a rough set approach is insignificant. Clearly, more experiments are needed to compare probabilistic and rough set approaches to missing attribute values.

Theoretically, we should not use singleton probabilistic approximations for data mining, see Section IV. Nevertheless, our experiments show that—in practice—singleton approximations are worth considering, since rule sets induced from such approximations classify unknown cases using the sophisticated LERS classification system [23], [24].

REFERENCES

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

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

[3] J. W. Grzymala-Busse, "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.

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

[5] 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.

[6] 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.

 ISBN: 978-1-61208-269-1

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

[8] 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.

[9] D. Ślęzak and W. Ziarko, "The investigation of the bayesian rough set model," *International Journal of Approximate Reasoning*, vol. 40, pp. 81–91, 2005.

[10] S. Tsumoto and H. Tanaka, "PRIMEROSE: probabilistic rule induction method based on rough sets and resampling methods," *Computational Intelligence*, vol. 11, pp. 389–405, 1995.

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

[12] Y. Y. Yao and S. K. M. Wong, "A decision theoretic framework for approximate concepts," *International Journal of Man-Machine Studies*, vol. 37, pp. 793–809, 1992.

[13] Y. Y. Yao, S. K. M. Wong, and P. Lingras, "A decision-theoretic rough set model," in *Proceedings of the 5th International Symposium on Methodologies for Intelligent Systems*, 1990, pp. 388–395.

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

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

[16] 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.

[17] P. G. Clark and J. W. Grzymala-Busse, "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.

[18] P. G. Clark and J. W. Grzymala-Busse, "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.

[19] P. G. Clark and J. W. Grzymala-Busse, "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.

[20] 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.

[21] 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.

[22] J. W. Grzymala-Busse and W. Rzasa, "Local and global approximations for incomplete data," *Transactions on Rough Sets*, vol. 8, pp. 21–34, 2008.

[23] 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.

[24] J. W. Grzymala-Busse, "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.

 ISBN: 978-1-61208-269-1