Experiments on Probabilistic Approximations

Patrick G. Clark *, Jerzy W. Grzymala-Busse *†
* *Department of Electrical Engineering and Computer Science*
University of Kansas
Lawrence, KS 66045, USA
† *Institute of Computer Science*
Polish Academy of Sciences
01-237 Warsaw, Poland
E-mail: pclark@ku.edu, jerzy@ku.edu

***Abstract*—Recently much attention has been paid to probabilistic (parameterized) approximations that are generalizations of ordinary lower and upper approximations known from rough set theory. The first objective of this paper is to compare the quality of such approximations and ordinary, lower and upper approximations. The second objective is to show that the number of distinct probabilistic approximations is quite limited. In our experiments we used six real-life data sets. Obviously, inconsistent data sets are required for such experiments, so the level of consistency in all data sets used for our experiments was decreased to enhance our experiments. Our main result is rather pessimistic: probabilistic approximations, different from ordinary lower or upper approximations, were better than ordinary approximations for only two out of these six data sets.**

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

I. Introduction

Recently much attention has been paid to probabilistic approximations, also known as parameterized approximations. Probabilistic approximations are generalizations of ordinary lower and upper approximations, which are fundamental concepts of rough set theory. Usually, two definitions of probabilistic approximations are considered, lower and upper, depending on two parameters, α and β, respectively. Since we are interested in all distinct probabilistic approximations, it is sufficient to use one parameter, denoted by α. If the parameter α is a positive number close to zero, the probabilistic approximation becomes the ordinary possible approximation, known from the standard rough set theory. On the other hand, if the parameter α is equal to one, the probabilistic approximation is identical with the ordinary lower approximation.

A basic question is whether probabilistic approximations are more valuable than ordinary approximations. More precisely, what is the estimate of the error rate for rule sets induced using probabilistic approximations from real-life data sets? Thus, the main objective of our research was to test whether probabilistic approximations, different from lower and upper approximations are truly better than standard lower and upper approximations.

For a given data set the number of distinct probabilistic approximations is quite limited. In this paper we report the exact number of distinct probabilistic approximations for six real-life data sets.

Table I
A decision table

	Attributes			Decision
Case	Temperature	Headache	Cough	Flu
1	high	yes	no	no
2	high	no	yes	no
3	normal	no	no	no
4	normal	no	no	no
5	high	yes	no	yes
6	high	yes	no	yes
7	high	no	yes	yes
8	high	no	no	maybe
9	high	no	no	maybe

II. Indiscernibility relation

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, 9}. 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 $a(x)$.

One of the most important ideas of rough set theory [1], [2] is an indiscernibility relation. 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 follows:

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

978-1-4577-0371-3/11/$26.00 ©2011 IEEE

The indiscernibility relation $R(B)$ is an equivalence relation. Equivalence classes of $R(B)$ are called *elementary sets* of B and are denoted by $[x]_B$. For Table I, all elementary sets of A are {1, 5, 6}, {2, 7}, {3, 4} and {8, 9}. A subset of U is called *A-definable* if it is a union of elementary sets.

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 *no* of the decision *Flu* is the set {1, 2, 3, 4}. This concept is not A-definable. The largest B-definable set contained in X is called the *B-lower approximation* of X, denoted by $\underline{appr}_B(X)$, and defined as follows

$$\cup\{[x]_B \mid [x]_B \subseteq X\}$$

while the smallest B-definable set containing X, denoted by $\overline{appr}_B(X)$ is called the *B-upper approximation* of X, and is defined as follows

$$\cup\{[x]_B \mid [x]_B \cap X \neq \emptyset\}.$$

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 $\{x \in U \mid a(x) = v\}$ [3]. For Table I, there are three concepts: the blocks $[(Flu, no)] = \{1, 2, 3, 4\}$, $[(Flu, yes)] = \{5, 6, 7\}$, and $[(Flu, maybe)] = \{8, 9\}$. A-approximations of the concept {1, 2, 3, 4} are:

$$\begin{aligned} \underline{appr}_A([(Flu, no)]) &= \{3, 4\}, \\ \overline{appr}_A([(Flu, no)]) &= \{1, 2, 3, 4, 5, 6, 7\}. \end{aligned}$$

III. Probabilistic approximations

In this paper we will assume that a data set is described by an indiscernibility relation $R(A)$ which is an equivalence relation. Additionally, we will denote the A-elementary set $[x]_A$ by $[x]$.

A generalization of ordinary lower and upper approximations, based on probability theory was introduced in [4] and then studied in many papers [5]–[14]. Such approximations are called *probabilistic* or *parameterized*.

In the *variable precision asymmetric* rough set model, see, e.g., [8], [14], probabilistic lower and upper approximations of the set $X \subseteq U$ are defined using the prior probability $P(X)$, a conditional probability $P(X \mid [x])$, and two parameters, denoted by α and β, where $1 \geq \alpha > P(X) > \beta \geq 0$. The *lower probabilistic approximation* of X (also called a *positive region* of X) is defined by

$$\underline{appr}_\alpha(X) = \cup\{[x] \mid x \in U, P(X \mid [x]) \geq \alpha\}$$

and the *boundary region* of X (the difference between the upper and lower probabilistic approximations of X) is defined by

$$BND_{\alpha,\beta}(X) = \cup\{[x] \mid x \in U, \beta < P(X \mid [x]) < \alpha\}.$$

Table II
Conditional probabilities

$[x]$	{1, 5, 6}	{2, 7}	{3, 4}	{8, 9}
$P(\{1, 2, 3, 4\} \mid [x])$	0.333	0.5	1.0	0

Hence the upper approximation of X, defined as a union of the lower approximation of X and the boundary region of X, is, in turn, defined by

$$\overline{appr}_\beta = \cup\{[x] \mid x \in U, P(X \mid [x]) > \beta\}.$$

Similar definitions of probabilistic approximations were studied in [11].

In this paper we are exploring all probabilistic approximations that can be defined for a given concept X. Our *probabilistic approximation* is defined as follows

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

where $1 \geq \alpha > 0$. We excluded the case of $\alpha = 0$ since then $appr_\alpha(X) = U$ for any X. Since we consider all possible values of α, our definition of $appr_\alpha(X)$ covers both lower and upper probabilistic approximations.

Thus we need only one parameter α (for a similar approach to probabilistic approximations see [15]). Note that if $\alpha = 1$, the probabilistic approximation becomes the standard lower approximation and if α is small, close to 0, the same definition describes the standard upper approximation.

For Table I and the concept X = [(*Flu*, *no*)] = {1, 2, 3, 4}, for any elementary set [x], $x \in U$, conditional probabilities $P(X|[x])$ are presented in Table II.

Thus, for the concept {1, 2, 3, 4} we may define only three distinct probabilistic approximations:
$appr_{0.333}(\{1, 2, 3, 4\}) = \{1, 2, 3, 4, 5, 6, 7\}$,
$appr_{0.5}(\{1, 2, 3, 4\}) = \{2, 3, 4, 7\}$,
$appr_{1.0}(\{1, 2, 3, 4\}) = \{3, 4\}$.

Note that there are only three distinct probabilistic approximations for the concept [(*Flu*, *yes*)] as well (the third one is a lower approximation of the set [(*Flu*, *yes*)], equal to the empty set). For the remaining concept, [(*Flu*, *maybe*)], there exists only one probabilistic approximation (equal to the lower and upper approximation of this concept and equal to {8, 9}.

In this paper, for the first time, the results of experiments on probabilistic approximations are presented. This paper is a continuation of research presented in [16], where certain, boundary, and possible rule sets were compared. However, the main focus of [16] was to compare these three types of rules, especially to study usefulness of boundary rules, for $\alpha \geq 0.6$. Since the conclusion of [16] was that the boundary rules are the worst among the three type of rules, we no longer study this type of rules. In [15] we also compared performance of ordinary lower and upper

Table III
A DECISION TABLE

	Attributes			Decision
Case	Temperature	Headache	Cough	Flu
1	high	yes	no	SPECIAL
2	high	no	yes	no
3	normal	no	no	no
4	normal	no	no	no
5	high	yes	no	SPECIAL
6	high	yes	no	SPECIAL
7	high	no	yes	yes
8	high	no	no	SPECIAL
9	high	no	no	SPECIAL

approximations with probabilistic approximations, however, with two different systems classifying testing cases: for probabilistic approximations the partial matching factor [17] was set to one. In experiments presented in this paper the system classifying testing cases was the same for all approximations. Additionally, we explore all possible values of the parameter α.

IV. Rule Induction with LERS

The LERS (Learning from Examples based on Rough Sets) data mining system [3], [18] starts from computing lower and upper approximations for every concept and then it induces rules using the MLEM2 (Modified Learning from Examples Module version 2) rule induction algorithm. Rules induced from lower and upper approximations are called *certain* and *possible*, respectively [19].

MLEM2 explores the search space of attribute-value pairs. Its input data set is a lower or upper approximation of a concept. In general, MLEM2 computes a local covering and then converts it into a rule set [18].

In order to induce probabilistic rules we have to modify input data sets. For every probabilistic approximation of the concept X = [(d, w)], the corresponding region will be unchanged (every entry will be the same as in the original data set). For all remaining cases, the decision value will be set to a special value, not listed in any attribute domain in the original data set, e.g., let us use the value SPECIAL. Then we will induce a *possible* rule set [3] using the MLEM2 rule induction algorithm. From the induced rule set, only rules with (d, w) on the right hand side will survive, all remaining rules (for other values of d and for values SPECIAL) should be deleted. The final rule set is a union of all rule sets computed this way separately for all values of d.

For example, if we want to induce probabilistic rules with $\alpha = 0.5$ and X = [(*Flu*, *no*)] = $\{1, 2, 3, 4\}$ for the data set presented on Table I we should construct the decision table presented as Table III.

Table IV
DATA SETS USED FOR EXPERIMENTS

Data set	Number of			Consistency
	cases	attributes	concepts	
Glass	214	9	6	55.14
Hepatitis	155	19	2	65.81
Iris	150	4	3	56.0
Postoperative patient	90	8	3	84.44
Primary tumor	339	17	21	72.27
Wine recognition	178	13	3	61.80

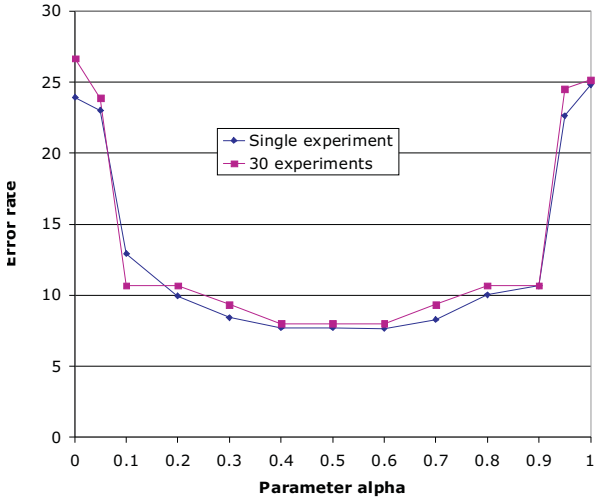

Figure 1. Results of experiments with *Iris* data set

From Table III, the MLEM2 rule induction algorithm induced the following possible rules with (Flu, no) on right hand side:

1, 2, 2
$(Temperature, normal) \rightarrow (Flu, no)$,
1, 1, 2
$(Cough, yes) \rightarrow (Flu, no)$.

Rules for remaining two concepts must be computed separately. Rules are presented in the LERS format, every rule is associated with three numbers: the total number of attribute-value pairs on the left-hand side of the rule, the total number of cases correctly classified by the rule during training, and the total number of training cases matching the left-hand side of the rule, i.e., the rule domain size.

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*. These data sets were enhanced by

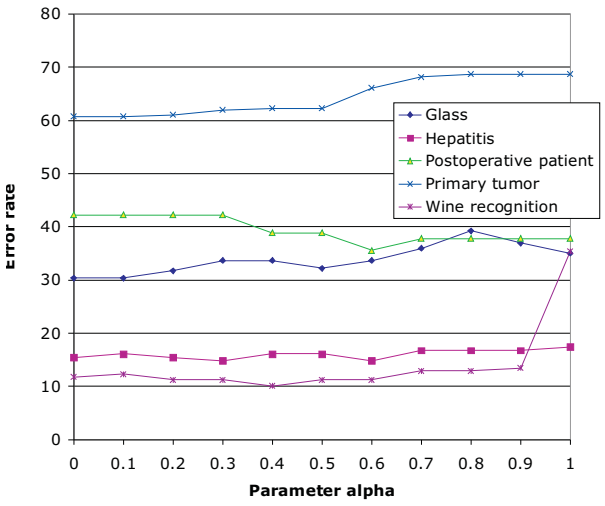

Figure 2. Results of single experiments with five remaining data sets

Table V
IRIS DATA SET

α	Error rate	Standard deviation
0.001	23.94	1.2022
0.05	23.01	0.9928
0.1	12.92	2.5101
0.2	9.95	0.3666
0.3	8.43	0.9114
0.4	7.70	0.4461
0.5	7.70	0.1637
0.6	7.66	0.2230
0.7	8.28	0.7779
0.8	10.02	0.3686
0.9	10.66	1.6050
0.95	22.63	1.4957
1	24.82	0.5007

reducing *consistency* (the percentage of cases not involved in any conflicts), see Table IV.

The main objective of our research was to test whether probabilistic approximations, different from lower and upper approximations, are truly better than lower and upper approximations. To accomplish this objective, we conducted experiments of a single ten-fold cross validation increasing the parameter α, with increments equal to 0.1, from 0 to 1.0. For a given data sets, in all of these eleven experiments we used identical ten pairs of larger (90%) and smaller (10%) data sets. If during such a sequence of eleven experiments, the error rate was smaller than the minimum of the error rates for lower and upper approximations or larger than maximum of the error rated for lower and upper approximations, we selected more precise values of the parameter α, to make sure that we are reaching an extremum (for example, for the glass data set, we concluded that the largest error rate is associated with the parameter α = 0.78). Additionally, for a value suspected to be an extremum, we conducted additional 30 experiments of ten-fold cross validation. We compared averages and the standard deviations, using the standard statistical test for the difference between two averages (two-tailed test with 5% of significance level).

We conducted extensive experiments for the *iris* data set to compare the error rate, a result of a single experiment of the ten-fold cross validation, using the same sampling for every experiment, with the error rate computed as an average of 30 experiments of ten-fold cross validation, where sampling was different for all experiments of ten-fold cross validation. As follows from our experiments, presented on Figure 1, a single experiment of ten-fold cross validation is a sufficient indicator of reaching an extremum associated with the average of 30 experiments. Therefore, in remaining experiments, for every value of the parameter α we used only a single experiment of ten-fold cross validation. However, a suspected extremum was tested using 30 experiments of ten-fold cross validation.

For the *iris* data set, as follows from Table V, we conclude that the difference is statistically significant and the upper approximation is better if we apply the standard statistical test for the difference between two averages (two tails and the significance level of 5%) for the upper approximation (the parameter α = 0.001) and the lower approximation (the parameter α = 1). Additionally, for any value of the parameter α equal to 0.1, 0.2,..., 0.9, and for the upper approximation, the same test indicates that the difference is statistically significant—the upper approximation is worse. In the rest of the paper, whenever we quote this statistical test, it will be always two two-tail test with the significance level of 5%.

For the *glass* data set, the best approximation is upper (the parameter α = 0.001). With increase of the parameter α, the error rate increases, up to α = 0.78, where the error rate is the largest. For the parameter α = 0.78, the average error rate of 30 experiments of ten-fold cross validation is 38.79%, while the error rate for 30 experiments of ten-fold cross validation for the parameter α = 0.001 is 37.85%, with the standard deviations equal to 1.9711 and 1.802, respectively. Thus, using the same standard statistical test for the difference between two averages we conclude that the difference between these two averages is significant, or, in different words, the probabilistic approximation for the parameter α = 0.78 is worse than the upper approximation.

For the *hepatitis* data set there are two promising values of the parameter α for which we may expect smaller error rates: 0.25 and 0.6. The average error rates and the standard deviations for 30 experiments of ten-fold cross validation are shown in Table VI.

It is not difficult to check that the differences between the averages for the upper approximation (the parameter α

Table VI
HEPATITIS DATA SET

α	Error rate	Standard deviation
0.001	17.18	1.5937
0.25	17.05	1.5029
0.6	16.84	1.5468
1.0	17.53	1.4184

= 0.001) and the probabilistic approximations for α = 0.25, 0.6, and the lower approximation (α = 1) are not significant. So, as in the glass data set, it is sufficient to consider only upper approximations and induce possible rule sets. We are not going to gain anything from considering probabilistic approximations different from the upper approximation.

For the *postoperative patient* data set a possible minimum of the error rate (Figure 2) is associated with the parameter α = 0.6. The corresponding error rate, a result of 30 experiments of ten-fold cross validation, is 38.63%, with the standard deviation = 2.5384. However, the error rate for the parameter α = 1 is 37.19% with the standard deviation = 2.6541, the difference is significant, and the lower approximation is better than the probabilistic approximation for the parameter α = 0.6. Moreover, for the upper approximation, the parameter α = 0.001, the error rate, a result of 30 experiments of ten-fold cross validation, is 39.63% with the standard deviation = 2.4121, so we may conclude that the difference between the lower approximation and the upper approximation is significant. Thus, for the postoperative patient data set, the lower approximation is the best option overall. It is the only data set among the six data sets for which the lower approximation is the best. On the other hand, any probabilistic approximation different from the lower approximation provides not better results.

For the *primary tumor* data set the error rate grows monotonically, with the increase of the parameter α = 0.001 to the parameter α = 1, so it is clear that we are not going to gain anything from using a probabilistic approximations that is different from the upper approximation.

The *wine recognition* data set is another example of the data set for which there exists a probabilistic approximation different from lower and upper approximations that is better than both lower and upper approximations. Namely, the error rate for the parameter α = 0.4, a result of 30 experiments of ten-fold cross validation, is 8.5%, with the standard deviation = 0.9978, while the error rate for the parameter α = 0.001 is 9.94% with the standard deviation = 1.0857, so the difference between the averages is significant.

Our secondary objective was to test how many different probabilistic approximations there exist for a given concept of the real-life data set. Results are listed in Tables VII-XII.

Table VII
GLASS DATA SET

Concept	Number of distinct probabilistic approximations
Glass-Type, 1	7
Glass-Type, 2	7
Glass-Type, 3	7

Table VIII
HEPATITIS DATA SET

Concept	Number of distinct probabilistic approximations
Class, 1	5
Class, 2	5

VI. CONCLUSION

The main objective of our research was to test whether probabilistic approximations, different from lower and upper approximations, are truly better than lower and upper approximations. As follows from our experiments, probabilistic approximations other than lower and upper approximations are better in two out of six real-life data sets.

However, for the iris data set there exists an entire spectrum of probabilistic approximations, different from lower and upper approximations, all of them better than either lower or upper approximations, and the difference in performance is spectacular. Thus, for some data sets we may gain a lot in performance by using probabilistic approximations different from lower and upper approximations.

Our secondary objective was to test how many different probabilistic approximations there exist for a given concept of the real-life data set. It turned out that such a number

Table IX
IRIS DATA SET

Concept	Number of distinct probabilistic approximations
Class, Iris-setosa	2
Class, Iris-versicolor	6
Class, Iris-viginica	5

Table X
POSTOPERATIVE PATIENT DATA SET

Concept	Number of distinct probabilistic approximations
ADM-DEC, A	3
ADM-DEC, I	1
ADM-DEC, S	3

Table XI
PRIMARY TUMOR DATA SET

Concept	Number of distinct probabilistic approximations
Class, 1	5
Class, 2	3
Class, 3	3
Class, 4	2
Class, 5	5
Class, 6	2
Class, 7	5
Class, 8	2
Class, 9	–
Class, 10	1
Class, 11	5
Class, 12	6
Class, 13	3
Class, 14	4
Class, 15	1
Class, 16	1
Class, 17	3
Class, 18	6
Class, 19	2
Class, 20	1
Class, 21	1
Class, 22	2

Table XII
WINE RECOGNITION DATA SET

Concept	Number of distinct probabilistic approximations
Class, 1	2
Class, 2	6
Class, 3	5

is not large. For some cases this number was equal to the smallest possible, equal to one, meaning that only one probabilistic approximation exists.

REFERENCES

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

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

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

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

[9] 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.

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

[11] 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.

[12] 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.

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

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

[15] J. W. Grzymala-Busse, S. R. Marepally, and Y. Yao, "An empirical comparison of rule sets induced by lers and probabilistic rough classification," in *Proceedings of the 7-th International Conference on Rough Sets and Current Trends in Computing*, 2010, pp. 590–599.

[16] ——, "A comparison of positive, boundary, and possible rules using the MLEM2 rule induction algorithm," in *Proceedings of the 10-th International Conference on Hybrid Intelligent Systems*, 2010, pp. 7–12.

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

[18] ——, "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.

[19] ——, "Knowledge acquisition under uncertainty—A rough set approach," *Journal of Intelligent & Robotic Systems*, vol. 1, pp. 3–16, 1988.