How Good Are Probabilistic Approximations for Rule Induction from Data with Missing Attribute Values?

Patrick G. Clark[1], Jerzy W. Grzymala-Busse[1,2], and Zdzislaw S. Hippe[3]

[1] Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS 66045, USA
[2] Institute of Computer Science, Polish Academy of Sciences, 01–237 Warsaw, Poland
pclark@ku.edu, jerzy@ku.edu
[3] Department of Expert Systems and Artificial Intelligence, University of Information Technology and Management, 35-225 Rzeszow, Poland
zhippe@wsiz.rzeszow.pl

Abstract. The main objective of our research was to test whether the probabilistic approximations should be used in rule induction from incomplete data. Probabilistic approximations, well known for many years, are used in variable precision rough set models and similar approaches to uncertainty.

For our experiments we used five standard data sets. Three data sets were incomplete to begin with and two data sets had missing attribute values that were randomly inserted. We used two interpretations of missing attribute values: lost values and "do not care" conditions. Among these ten combinations of a data set and a type of missing attribute values, in one combination the error rate (the result of ten-fold cross validation) was smaller than for ordinary approximations; for other two combinations, the error rate was larger than for ordinary approximations.

1 Introduction

One of the fundamental concepts of rough set theory is an idea of lower and upper approximations. A generalization of such approximations, a probabilistic approximation, introduced in [1], was applied in variable precision rough set models, Bayesian rough sets and decision-theoretic rough set models [2–10]. The probabilistic approximation is associated with some parameter α (interpreted as a probability). If α is very small, say 0.001 (this number depends on the size of the data set), the probabilistic approximation is reduced to the upper approximation; if α is equal to 1.0, the probabilistic approximation becomes the lower approximation. The problem is how useful are *proper* probabilistic approximations (with α larger than 0.001 but smaller than 1.0). We studied usefulness of proper probabilistic approximations for inconsistent data sets [11],

J.T. Yao et al. (Eds.): RSCTC 2012, LNAI 7413, pp. 46–55, 2012.

where we concluded that proper probabilistic approximations are not frequently better than ordinary lower and upper approximations.

In this paper we study usefulness of the proper probabilistic approximations applied for rule induction from incomplete data. We will use two interpretations of missing attribute values, as *lost values* (the original attribute values are not longer accessible, for details see [12, 13]) and as *"do not care" conditions* (the original values were irrelevant, see [14, 15]).

For data sets with missing attribute values there exist many definitions of approximations [16], we use one of the most successful options (from the view point of rule induction) called *concept* approximations [16]. Concept approximations were generalized to concept probabilistic approximations in [17].

Our experiments on five data sets with two types of missing attribute values (altogether ten combinations) show that the proper concept probabilistic approximations are not very useful for rule induction from incomplete data sets: for one combination the error rate (result of ten-fold cross validation) was smaller than for ordinary concept approximations, for two combinations such error rate was larger than for ordinary concept approximations, for remaining seven combinations the error rate was neither smaller nor larger.

2 Incomplete Data Sets

The data sets are presented in the form of a *decision table*. Rows of the decision table represent *cases*, while columns are labeled by *variables*. The set of all cases will be denoted by U. In Table 1, $U = \{1, 2, 3, 4, 5, 6, 7, 8\}$. Independent variables are called *attributes* and a dependent variable is called a *decision* and is denoted by d. The set of all attributes will be denoted by A. In Table 1, $A = \{$*Wind*, *Humidity*, *Temperature*$\}$. 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*, denoted by "?", and *"do not care" conditions*, denoted by "*". Table 1 present an incomplete data set affected by both lost values and "do not care" conditions.

One of the most important ideas of rough set theory [18, 19] is 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 follows:

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

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 *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 *Trip* is the set $\{1, 3, 5, 7\}$. 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

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

Table 1. A decision table

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

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 $\{x \in U \mid a(x) = v\}$ [20].

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 1 the blocks of attribute-value pairs are:

[(Wind, low)] = {2, 3, 4, 6},
[(Wind, high)] = {4, 5, 7, 8},
[(Humidity, high)] = {1, 3, 5, 7},
[(Humidity, low)] = {2, 3, 4, 8},
[(Temperature, high)] = {1, 2, 6},
[(Temperature, low)] = {3, 4, 6, 7, 8}.

For a case $x \in U$ and $B \subseteq A$, the *characteristic set* $K_B(x)$ is defined as the intersection of the sets $K(x, a)$, for all $a \in B$, where the set $K(x, a)$ is defined in the following way:

- If $a(x)$ is specified, then $K(x, a)$ is the block $[(a, a(x))]$ of attribute a and its value $a(x)$,

– If $a(x) =?$ or $a(x) = *$ then the set $K(x, a) = U$, where U is the set of all cases.

For Table 1 and $B = A$,

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

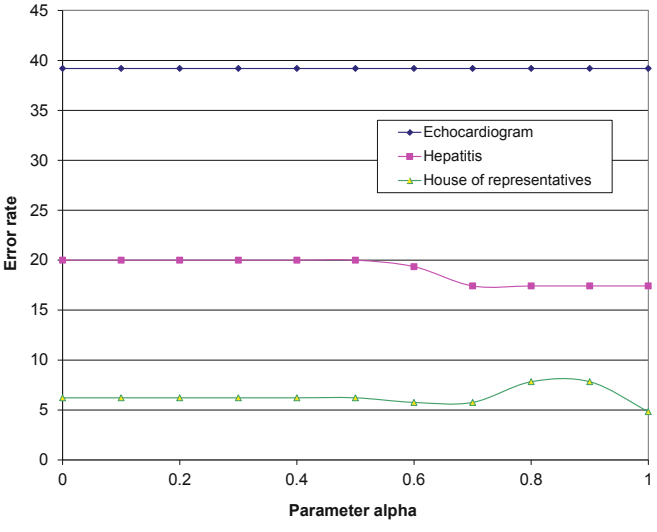

Fig. 1. Error rates for data sets *Echocardiogram*, *Hepatitis*, and *House of representatives* with lost values

Note that for incomplete data there is a few possible ways to define approximations [16], we use *concept* approximations [17]. A B-*concept lower approximation* of the concept X is defined as follows:

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

A B-*concept upper approximation* of the concept X is defined as follows:

$$\overline{B}X = \cup\{K_B(x) \mid x \in X, K_B(x) \cap X \neq \emptyset\} =$$
$$= \cup\{K_B(x) \mid x \in X\}.$$

For Table 1, A-concept lower and A-concept upper approximations of the two concepts: {1, 3, 5, 7} and {2, 4, 6, 8} are:

$\underline{A}\{1, 3, 5, 7\} = \{1, 5, 7\}$, $\overline{A}\{1, 3, 5, 7\} = \{1, 3, 4, 5, 6, 7\}$,
$\underline{A}\{2, 4, 6, 8\} = \{2, 4, 8\}$, $\overline{A}\{2, 4, 6, 8\} = \{2, 3, 4, 6, 8\}$.

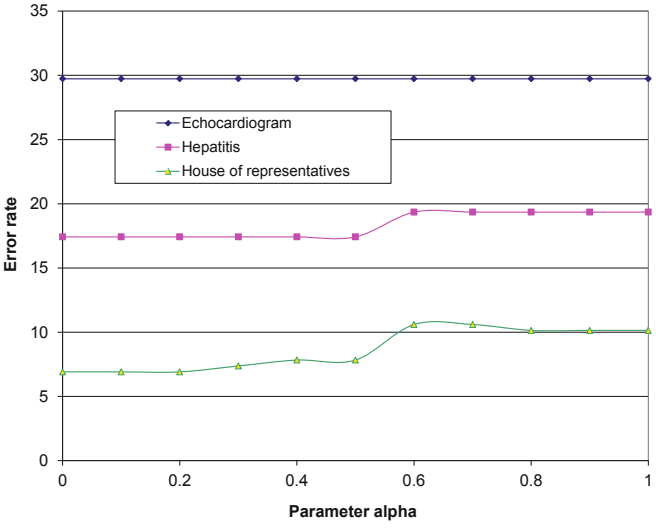

Fig. 2. Error rates for data sets *Echocardiogram*, *Hepatitis*, and *House of representatives* with "do not care" conditions

3 Probabilistic Approximations

In this paper we explore all probabilistic approximations that can be defined for a given concept X. For completely specified data sets a *probabilistic approximation* is defined as follows

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

where $[x]$ is $[x]_A$ and α is a parameter, $0 < \alpha \leq 1$, see [17]. For discussion on how this definition is related to the value precision asymmetric rough sets see [11, 17].

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.

For incomplete data sets, a *B-concept probabilistic approximation* is defined by the following formula [17]

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

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 1 and the concept $X = [(Trip, no)] = \{1, 3, 5, 7\}$, for any characteristic set $K(x)$, $x \in U$, conditional probabilities $P(X|K_(x))$ are presented in Table 2.

Thus, for the concept $\{1, 3, 5, 7\}$ we may define only two distinct probabilistic approximations:

$$appr_{1.0}(\{1,3,5,7\}) = \{1,5,7\} \text{ and } appr_{0.333}(\{1,3,5,7\}) = \{1,3,4,5,6,7\}.$$

Table 2. Conditional probabilities

$K(x)$	{1}	{5, 7}	{7}	{3, 4, 6}	{3, 4, 8}	{2, 3, 4, 6}	{2}	{4, 8}
$P(\{1,3,5,7\} \mid K(x))$	1.0	1.0	1.0	0.333	0.333	0.25	0	0

Table 3. Data sets used for experiments

Data set	Number of cases	attributes	concepts	Percentage of missing attribute values
Echocardiogram	74	7	2	4.05
Hepatitis	155	19	2	5.67
House of Representatives	434	16	2	5.40
Image segmentation	210	19	7	70
Lymphography	148	18	4	70

4 Experiments

For our experiments we used five real-life data sets that are available on the University of California at Irvine *Machine Learning Repository*. Two of these data sets (*Image segmentation* and *Lymphography* were originally completely specified, i.e., they did not contain any missing attribute values. However, we replaced, randomly, 70% of existing attribute values by signs of missing attribute values, first by *lost* values and then we converted *lost values* to *"do not care" conditions*, see Table 3.

For rule induction we used the MLEM2 (Modified Learning from Examples Module version 2) rule induction algorithm, a component of the LERS (Learning from Examples based on Rough Sets) data mining system [20, 21].

The main objective of our research was to test whether proper probabilistic approximations are better than concept lower and upper approximations. We conducted experiments of a single ten-fold cross validation starting with 0.001 and then increasing the parameter α by 0.1 until reaching 1.0. For a given data set, in all of these eleven experiments we used identical ten pairs of larger (90%) and smaller (10%) data sets. Results of our experiments are shown in Figures 1–4. 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 and we conducted additional 30 experiments of ten-fold cross validation.

For example, for the *Echocardiogram* data set, affected by lost values, denoted by "?", the error rate was constant, so there is no need for additional 30 experiments, see Figure 1. Similarly, for the *Hepatitis* data set, also affected by *lost values*. But for the *House of representative* data set, affected by *lost values*, it is

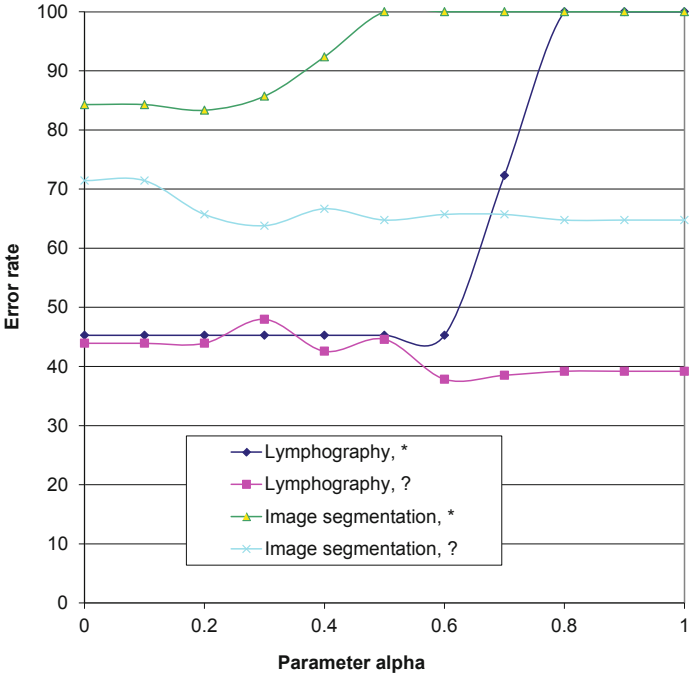

Fig. 3. Error rates for data sets *Image segmentation* and *Lymphography*

clear that we should look more closely at the parameter α around the values 0.65 and 0.85. Results are presented in Table 4. Using the standard statistical test for the difference between two averages (two tails and the significance level of 5%) we may conclude that there is no statistically significant difference between the probabilistic approximation associated with $\alpha = 0.65$ and the upper approximation ($\alpha = 0.001$). The same test indicates that the probabilistic approximation, associated with $\alpha = 0.85$ is worse than the upper approximation ($\alpha = 0.001$), as well as the lower approximation ($\alpha = 1.0$). Results of all remaining 30 experiments of ten-fold cross validation are presented in Tables 5–8.

In particular, for the *House of representatives* data set with "do not car" conditions as missing attribute values, for $\alpha = 0.65$, the corresponding probabilistic approximation is worse than both lower ($\alpha = 1.0$) and upper ($\alpha = 0.001$) approximations. On the other hand, for the *Image segmentation* data set with "do not care" conditions, for $\alpha = 0.2$, the error rate is significantly better than for both lower and upper approximations. In experiments reported in this paper this is the only situation of this type. For remaining data sets, no matter with lost values or "do not care" conditions, probabilistic approximations for α between 0.1 and 0.9 are neither worse than the worst for the two: lower and upper approximations nor better than the best of the two.

Table 4. Results of 30 experiments of ten-fold cross validation for *House of representatives*, lost values

α	Error rate	Standard deviation
0.001	6.59	0.6159
0.65	6.42	0.6396
0.85	7.31	0.7055
1.0	5.44	0.5885

Table 5. Results of 30 experiments of ten-fold cross validation for *House of representatives*, "do not care" conditions

α	Error rate	Standard deviation
0.001	5.97	0.5147
0.65	10.14	0.6819
1.0	9.72	0.7584

Table 6. Results of 30 experiments of ten-fold cross validation for *Image segmentation*, lost values

α	Error rate	Standard deviation
0.3	65.56	2.6567
1.0	63.44	2.5982

Table 7. Results of 30 experiments of ten-fold cross validation for *Image segmentation*, "do not care" conditions

α	Error rate	Standard deviation
0.001	85.20	1.1525
0.2	84.20	1.1191

Table 8. Results of 30 experiments of ten-fold cross validation for *Lymphography*, lost values

α	Error rate	Standard deviation
0.001	44.84	2.1767
0.3	44.64	2.4647
0.4	41.24	2.1031
1.0	37.61	2.2227

5 Conclusions

As follows from our experiments, the *proper* probabilistic approximations (ones with α between 0.1 and 0.9) were neither better nor worse than ordinary lower ($\alpha = 1.0$) and upper ($\alpha = 0.001$) approximations, except for three situations. In one of them (the *Image segmentation* data set with "do not care" conditions was better than ordinary approximations, in other two situations (both for the *House of representatives* data set, with lost values and "do not care" conditions) the proper probabilistic approximations were worse than ordinary approximations.

References

1. Wong, S.K.M., Ziarko, W.: INFER—an adaptive decision support system based on the probabilistic approximate classification. In: Proceedings of the 6th International Workshop on Expert Systems and their Applications, pp. 713–726 (1986)
2. Grzymala-Busse, J.W., Ziarko, W.: Data mining based on rough sets. In: Wang, J. (ed.) Data Mining: Opportunities and Challenges, pp. 142–173. Idea Group Publ., Hershey (2003)
3. Pawlak, Z., Skowron, A.: Rough sets: Some extensions. Information Sciences 177, 28–40 (2007)
4. Pawlak, Z., Wong, S.K.M., Ziarko, W.: Rough sets: probabilistic versus deterministic approach. International Journal of Man-Machine Studies 29, 81–95 (1988)
5. Ślęzak, D., Ziarko, W.: The investigation of the bayesian rough set model. International Journal of Approximate Reasoning 40, 81–91 (2005)
6. Yao, Y.Y.: Probabilistic rough set approximations. International Journal of Approximate Reasoning 49, 255–271 (2008)
7. Yao, Y.Y., Wong, S.K.M.: A decision theoretic framework for approximate concepts. International Journal of Man-Machine Studies 37, 793–809 (1992)
8. Yao, Y.Y., Wong, S.K.M., Lingras, P.: A decision-theoretic rough set model. In: Proceedings of the 5th International Symposium on Methodologies for Intelligent Systems, pp. 388–395 (1990)
9. Ziarko, W.: Variable precision rough set model. Journal of Computer and System Sciences 46(1), 39–59 (1993)
10. Ziarko, W.: Probabilistic approach to rough sets. International Journal of Approximate Reasoning 49, 272–284 (2008)
11. Clark, P.G., Grzymala-Busse, J.W.: Experiments on probabilistic approximations. In: Proceedings of the 2011 IEEE International Conference on Granular Computing, pp. 144–149 (2011)
12. Grzymala-Busse, J.W., Wang, A.Y.: 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 1997) at the Third Joint Conference on Information Sciences (JCIS 1997), pp. 69–72 (1997)
13. Stefanowski, J., Tsoukias, A.: Incomplete information tables and rough classification. Computational Intelligence 17(3), 545–566 (2001)
14. Grzymala-Busse, J.W.: On the unknown attribute values in learning from examples. In: Proceedings of the ISMIS 1991, 6th International Symposium on Methodologies for Intelligent Systems, pp. 368–377 (1991)

15. Kryszkiewicz, M.: Rough set approach to incomplete information systems. In: Proceedings of the Second Annual Joint Conference on Information Sciences, pp. 194–197 (1995)
16. Grzymala-Busse, J.W.: Rough set strategies to data with missing attribute values. In: Workshop Notes, Foundations and New Directions of Data Mining, in conjunction with the 3rd International Conference on Data Mining, pp. 56–63 (2003)
17. Grzymała-Busse, J.W.: Generalized Parameterized Approximations. In: Yao, J., Ramanna, S., Wang, G., Suraj, Z. (eds.) RSKT 2011. LNCS, vol. 6954, pp. 136–145. Springer, Heidelberg (2011)
18. Pawlak, Z.: Rough sets. International Journal of Computer and Information Sciences 11, 341–356 (1982)
19. Pawlak, Z.: Rough Sets. Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Dordrecht (1991)
20. Grzymala-Busse, J.W.: LERS—a system for learning from examples based on rough sets. In: Slowinski, R. (ed.) Intelligent Decision Support. Handbook of Applications and Advances of the Rough Set Theory, pp. 3–18. Handbook of Applications and Advances of the Rough Set Theory. Kluwer Academic Publishers, Dordrecht (1992)
21. Grzymala-Busse, J.W.: MLEM2: A new algorithm for rule induction from imperfect data. In: Proceedings of the 9th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, pp. 243–250 (2002)