A Comparison of Three Approximation Strategies for Incomplete Data Sets

Jerzy W. Grzymala-Busse
Department of Electrical Engineering
and Computer Science
University of Kansas
Lawrence, KS 66045–7621, USA
and
Institute of Computer Science
Polish Academy of Sciences
01-237 Warsaw, Poland
jerzy@ku.edu

Witold J. Grzymala-Busse
Touchnet Information Systems, Inc.
Lenexa, KS 66219, USA
wgrzymala@touchnet.com

Zdzislaw S. Hippe
Department of Expert Systems
and Artificial Intelligence
University of Information
Technology and Management
35–225 Rzeszow, Poland
zhippe@wsiz.rzeszow.pl

Wojciech Rzasa
Department of Computer Science
University of Rzeszow
35–310 Rzeszow, Poland
wrzasa@univ.rzeszow.pl

Abstract

In this paper we consider incomplete data sets, i.e., data sets with missing attribute values. Two different types of missing attribute values are studied: lost and "do not care". Furthermore, three definitions of approximations are discussed: singleton, subset, and concept. Theoretically, singleton approximations should not be used in data mining since concepts approximated by singleton approximations are not definable. However, we conducted a number of experiments on 44 different incomplete data sets using all three approximation definitions and our results show that none of these approximations is superior to the other.

1 Introduction

Many approaches to mining incomplete data sets are used in practice, see, e.g., [6]. In our paper we will study how to approximate incomplete data using rough set theory.

We will consider two types of missing attribute values: lost values and "do not care" conditions. Lost values are interpreted as originally specified, but currently unavailable since these values were incidentally erased, forgotten to be recorded, etc. A rough set approach to incomplete data sets in which all attribute values were lost was presented for the first time in [10], where two algorithms for rule induction, modified to handle lost attribute values, were introduced.

The second possibility is a "do not care" condition. Such missing attribute values were irrelevant during collection of data since an expert decided that the attribute value was irrelevant for a classification or diagnosis of the case. For example, a data set describing flu patients may contain, among other attributes, an attribute *Height* with three possible values: *short*, *medium*, and *tall*. It seems that this attribute is irrelevant to flu diagnosis and many patients may leave it unspecified. If we suspect that this attribute does matter, the best interpretation for missing attribute values is replacing them by all possible existing attribute values. A rough set approach to incomplete data sets in which all attribute values were "do not care" conditions was presented for the first time in [2], where a method for rule induction was introduced in which each missing attribute value was replaced by all values from the domain of the attribute.

For incomplete decision tables there are three important and different possibilities to define lower and upper approximations, called singleton, subset, and concept approximations [5]. Singleton lower and upper approximations were

0-7695-3032-X/07 $25.00 © 2007 IEEE
DOI 10.1109/GrC.2007.119

Table 1. An example of the decision table

	Attributes			Decision
Case	Temperature	Hemoglobin	Blood_Pressure	Comfort
1	low	fair	low	low
2	?	fair	normal	low
3	normal	good	low	low
4	normal	good	*	medium
5	*	good	?	medium
6	low	?	normal	medium
7	normal	*	normal	medium
8	high	good	normal	medium

studied, e.g., in [12, 13, 17, 18]. Note that similar definitions of lower and upper approximations, though not for incomplete decision tables, were studied in [14, 16, 19]. Further definitions of approximations were discussed in [8, 9]. Additionally, note that some other rough-set approaches to missing attribute values were presented in [2] as well.

Frequently, singleton approximations, both lower and upper, are not definable. An example of such a data set is included in this paper. So, theoretically, such approximations should not be used for data mining. The main objective of this paper was to test the quality of all three approximations: singleton, subset and concept in terms of an error rate using many experiments with incomplete data and ten-fold cross validation. In our experiments, for singleton approximations rules were not induced from these approximations but rather from maximal definable sets for lower approximations and minimal definable sets for upper approximations. The conclusion is that among the three approximations there is no superior; so, in practice, we may use either; or even better, for a specific data set use the best approximation, selected by testing all three.

Similar research, reported in [7], was aimed at comparing different interpretations of missing attribute values while using the same definition of approximation (concept) in all experiments.

2 Blocks of attribute-value pairs

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 1.

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, ..., 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 = {*Temperature*, *Hemoglobin*, *Blood_Pressure*}. Any decision table defines a function ρ that maps the direct product of U and A into the set of all values. For example, in Table 1, $\rho(1, Temperature) = low$. A decision table with an incompletely specified function ρ will be called *incomplete*.

Table 2. Basic data sets

Data set	Number of		
	cases	attributes	concepts
Global Warming	33	5	4
Hepatitis	155	19	2
Lymphography	148	18	4

For the rest of the paper we will assume that all decision values are specified, i.e., they are not missing. Also, we will assume that lost values will be denoted by "?" and "do not care" conditions by "*". Additionally, we will assume that for each case at least one attribute value is specified.

An important tool to analyze complete decision tables is a block of the attribute-value pair. Let a be an attribute, i.e., $a \in A$ and let v be a value of a for some case. For complete decision tables if $t = (a, v)$ is an attribute-value pair then a *block* of t, denoted $[t]$, is a set of all cases from U that for attribute a have value v. For incomplete decision tables the definition of a block of an attribute-value pair must be modified in the following way:

- If for an attribute a there exists a case x such that $\rho(x, a) =?$, 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., $\rho(x, a) = *$, then the case x should be included in blocks $[(a, v)]$ for all specified values v of attribute a.

For Table 1 the blocks of attribute-value pairs are:

[(Temperature, low)] = {1, 5, 6},
[(Temperature, normal)] = {3, 4, 5, 7},
[(Temperature, high)] = {5, 8},
[(Hemoglobin, fair)] = {1, 2, 7},
[(Hemoglobin, good)] = {3, 4, 5, 7, 8},
[(Blood_Pressure, low)] = {1, 3, 4},
[(Blood_Pressure, normal)] = {2, 4, 6, 7, 8},

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

Table 3. Error rates for certain rules and lost values

Percentage of missing attribute values	Data set		
	Global Warming	Hepatitis	Lymphography
30	69.70%	18.71%	25.68%
50	75.76%	21.29%	28.38%
70	78.79%	22.58%	33.11%
90	–	36.13%	57.43%

Table 4. Error rates for Global Warming, certain rules, and "do not care" conditions

Percentage of missing attribute values	Approximations	
	Singleton	Subset and Concept
30	93.94%	69.70%
50	100%	100%
70	100%	100%

- If $\rho(x,a)$ is specified, then $K(x,a)$ is the block $[(a,\rho(x,a)]$ of attribute a and its value $\rho(x,a)$,
- If $\rho(x,a) =?$ or $\rho(x,a) = *$ then the set $K(x,a) = U$.

For Table 1 and $B = A$,

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

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.

Table 5. Error rates for Global Warming, possible rules, and lost values

Percentage of missing attribute values	Approximations		
	Singleton	Subset	Concept
30	69.70%	63.64%	66.67%
50	63.64%	69.70%	60.61%
70	93.94%	75.76%	78.79%

3 Definability

For completely specified decision tables, any union of elementary sets of B is called a B-definable set [15]. 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. For example, the set $\{7\}$ is A-locally definable since $\{7\} = [(Temperature, normal)] \cap [(Hemoglobin, fair)]$. However, the set $\{7\}$ is not A-globally definable. On the other hand, the set $\{2\}$ is not even A-locally definable. Obviously, if a set is not B-locally definable then it cannot be expressed by rule sets using attributes from B. This is why it is important to distinguish between B-locally definable sets and those that are not B-locally definable.

4 Lower and upper approximations

For incomplete decision tables lower and upper approximations may be defined in a few different ways. Here we suggest three different definitions of lower and upper approximations for incomplete decision tables, following [5]. Let B be a subset of the set A of all attributes. Let X be any subset of the set U of all cases. The set X is called a *concept* and is usually defined as the set of all cases defined by a specific value of the decision.

Our first definition uses a similar idea as in the previous articles on incomplete decision tables [12, 13, 17, 18], i.e., lower and upper approximations are sets of singletons from the universe U satisfying some properties. We will call these approximations *singleton*. A singleton B-lower approximation of X is defined as follows:

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

Table 6. Error rates for Global Warming, possible rules, and "do not care" conditions

Percentage of missing	Approximations	
	Singleton and Concept	Subset
30	69.70%	72.73%
50	66.67%	66.67%
70	69.70%	66.67%

Table 7. Error rates for Hepatitis, certain rules, and "do not care" conditions

Percentage of missing attribute values	Approximations	
	Singleton	Subset and Concept
30	16.77%	21.29%
50	20.65%	20.65%
70	100%	100%
90	100%	100%

A singleton B-upper approximation of X is

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

In our example of the decision table presented in Table 1 let us say that $B = A$. Then the singleton A-lower and A-upper approximations of the two concepts: $\{1, 2, 3\}$ and $\{4, 5, 6, 7, 8\}$ are:

$$\underline{A}\{1, 2, 3\} = \{1\},$$

$$\underline{A}\{4, 5, 6, 7, 8\} = \{6, 7, 8\},$$

$$\overline{A}\{1, 2, 3\} = \{1, 2, 3, 4, 5\},$$

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

We may easily observe that the set $\{1, 2, 3, 4, 5\} = \overline{A}\{1, 2, 3\}$ is not A-locally definable since in all blocks of attribute-value pairs cases 2 and 7 are inseparable. Thus, as it was observed in, e.g., [5], singleton approximations should not be used, theoretically, for data mining and, in particular, for rule induction. The next possibility is to define lower and upper approximations for incomplete decision tables using characteristic sets instead of singletons.

Table 8. Error rates for Hepatitis, possible rules, and lost values

Percentage of missing attribute values	Approximations		
	Singleton	Subset	Concept
30	23.23%	20.00%	20.00%
50	21.94%	23.87%	21.94%
70	17.42%	20.00%	20.65%
90	38.06%	38.06%	38.71%

There are two ways to do this. Using the first way, a *subset* B-lower approximation of X is defined as follows:

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

A *subset* B-upper approximation of X is

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

For the same decision table, presented in Table 1, the subset A-lower and A-upper approximations are

$$\underline{A}\{1, 2, 3\} = \{1\},$$

$$\underline{A}\{4, 5, 6, 7, 8\} = \{4, 6, 7, 8\},$$

$$\overline{A}\{1, 2, 3\} = U,$$

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

The second way is to modify the subset definition of lower and upper approximation by replacing the universe U from the subset definition by a concept X. A *concept* B-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 *concept* B-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 the decision table presented in Table 1, the concept A-upper approximations are

$$\overline{A}\{1, 2, 3\} = \{1, 2, 3, 4, 7\},$$

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

Note that for complete decision tables, all three definitions of lower approximations, singleton, subset and concept, coalesce to the same definition. Also, for complete decision tables, all three definitions of upper approximations coalesce to the same definition. This is not true for incomplete decision tables, as our example shows.

Table 9. Error rates for Hepatitis, possible rules, and "do not care" conditions

Percentage of missing attribute values	Approximations	
	Singleton and Concept	Subset
30	17.42%	19.35%
50	20.65%	20.65%
70	20.65%	20.65%
90	23.23%	23.23%

Table 10. Error rates for Lymphography, certain rules, and "do not care" conditions

Percentage of missing attribute values	Approximations	
	Singleton	Subset and Concept
30	34.46%	35.81%
50	44.59%	41.89%
70	100%	100%
90	100%	100%

Table 11. Error rates for Lymphography, possible rules, and lost values

Percentage of missing attribute values	Approximations		
	Singleton	Subset	Concept
30	24.32%	24.32%	24.32%
50	31.08%	28.38%	27.70%
70	41.22%	42.57%	36.49%
90	52.70%	59.46%	56.76%

5 Experiments

The MLEM2 algorithm [4] of the LERS (Learning from Examples based on Rough Sets) data mining system [3] was used to induce rule sets from data. Rules induced from the lower approximation of the concept *certainly* describe the concept, hence such rules are called *certain* [1]. On the other hand, rules induced from the upper approximation of the concept describe the concept *possibly*, so these rules are called *possible* [1]. The MLEM2 algorithm handles missing attribute values using rough set theory, i.e., for a concept the user has a choice between singleton, subset and concept approximations.

For our experiments we used many data sets derived from three basic data sets: *Global Warming*, *Hepatitis*, and *Lymphography*, see Table 2. The *Global Warming* data set was described, e.g., in [11]. This data set presents the Earth global temperature between 1958 and 1990. Remaining two basic data sets, *Hepatitis* and *Lymphography*, are well-known data retrieved from the University of California at Irvine Data Depository. All three data sets were complete. In each basic data set we replaced some existing values by question marks (lost values) and asterisks ("do not care" conditions), starting from 30% of the total number of attribute values, incrementally, with the 20% increment, until the limit of 90% was reached. Thus a basic data set was replaced by eight new data sets, with 30%, 50%, 70% and 90% of lost values and with 30%, 50%, 70% and 90% of "do not care" conditions. Each of eight new data sets contained only one type of missing attribute values.

Note that for the *Global Warming* data set we generated only six new data sets since with 90% of missing attribute values and only five attributes our assumption that every case should be described by at least one specified attribute value was violated. Additionally, these data sets were generated separately for lower and upper approximations. The total number of data sets used for experiments was 44. Note that though missing attribute values were assigned randomly, in order to compare the three types of approximations: singleton, subset and concept, for every series of corresponding three ten-fold cross validation experiments the same data set (one out of 44) was used and in all three ten-fold cross validation experiments a partition of the data set into ten pairs of subsets was also the same. Error rates computed as a result of ten-fold cross validation are presented in Tables 3–12. Note that for lower approximations and lost values all three kinds of approximations (singleton, subset and concept) are the same, so error rates are also the same. Results from Table 3 are presented only for completeness. Similarly, for "do not care" conditions the corresponding singleton and concept upper approximations and also the same.

6 Conclusions

Results of our experiments (Tables 3-12) show that none of the three definitions of approximation: single, subset,

Table 12. Error rates for Lymphography, possible rules, and "do not care" conditions

Percentage of missing attribute values	Approximations	
	Singleton and Concept	Subset
30	29.73%	35.14%
50	50.00%	45.27%
70	45.27%	45.27%
90	45.95%	46.62%

and concept is better than others. This observation is based on the Wilcoxon matched-pairs signed test at the 5% significance level using a two-tailed test. This test was used three times, to compare any possible pair of approximations among the three approximations: singleton, subset, and concept.

Our experiments support previous observations from [7], namely, that in such experiments on data with large number of missing attribute values, error rates show large variance. Additionally, for "do not care" conditions, lower approximations of concepts for data with large number of missing attribute values were empty, so the induced rule sets were also empty, and, as a result, the error rate was 100%.

References

[1] J. W. Grzymala-Busse. Knowledge acquisition under uncertainty—A rough set approach. *Journal of Intelligent & Robotic Systems*, 1:3–16, 1988.

[2] J. W. Grzymala-Busse. On the unknown attribute values in learning from examples. In *Proceedings of the ISMIS-91, 6th International Symposium on Methodologies for Intelligent Systems*, pages 368–377, 1991.

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

[4] 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, (IPMU 2002)*, pages 243–250, 2002.

[5] 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*, pages 56–63, 2003.

[6] J. W. Grzymala-Busse and W. J. Grzymala-Busse. Handling missing attribute values. In O. Maimon and L. Rokach, editors, *Data Mining and Knowledge Discovery Handbook*, pages 37–57. Springer-Verlag, Berlin, Heidelberg, 2005.

[7] J. W. Grzymala-Busse and W. J. Grzymala-Busse. An experimental comparison of three rough set approaches to missing attribute values. In J. F. Peters and A. Skowron, editors, *Transactions on Rough Sets*, pages 31–50. Springer-Verlag, Berlin, Heidelberg, 2007.

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

[9] J. W. Grzymala-Busse and W. Rzasa. Definability of approximations for a generalization of the indiscernibility relation. In *Proceedings of the 2007 IEEE Symposium on Foundations of Computational Intelligence (IEEE FOCI 2007)*, pages 65–72, 2007.

[10] 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)*, pages 69–72, 1997.

[11] J. D. Gunn and J. W. Grzymala-Busse. Global temperature stability by rule induction: An interdisciplinary bridge. *Human Ecology*, 22:59–81, 1994.

[12] M. Kryszkiewicz. Rough set approach to incomplete information systems. In *Proceedings of the Second Annual Joint Conference on Information Sciences*, pages 194–197, 1995.

[13] M. Kryszkiewicz. Rules in incomplete information systems. *Information Sciences*, 113:271–292, 1999.

[14] T. Y. Lin. Topological and fuzzy rough sets. In R. Slowinski, editor, *Intelligent Decision Support. Handbook of Applications and Advances of the Rough Sets Theory*, pages 287–304. Kluwer Academic Publishers, Dordrecht, Boston, London, 1992.

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

[16] R. Slowinski and D. Vanderpooten. A generalized definition of rough approximations based on similarity. *IEEE Transactions on Knowledge and Data Engineering*, 12:331–336, 2000.

[17] J. Stefanowski and A. Tsoukias. On the extension of rough sets under incomplete information. In *Proceedings of the RSFDGrC'1999, 7th International Workshop on New Directions in Rough Sets, Data Mining, and Granular-Soft Computing*, pages 73–81, 1999.

[18] J. Stefanowski and A. Tsoukias. Incomplete information tables and rough classification. *Computational Intelligence*, 17:545–566, 2001.

[19] Y. Y. Yao. Relational interpretations of neighborhood operators and rough set approximation operators. *Information Sciences*, 111:239–259, 1998.