A comparison of traditional and rough set approaches to missing attribute values in data mining

J. W. Grzymala-Busse
Department of Electrical Engineering and Computer Science, University of Kansas, USA
Institute of Computer Science, Polish Academy of Sciences, Poland

Abstract

Real-life data sets are often incomplete, i.e., some attribute values are missing. In this paper we compare traditional, frequently used methods of handling missing attribute values, which are based on preprocessing, with another class of methods dealing with missing attribute values in which rule induction is performed directly on incomplete data sets, i.e., handling missing attribute values and rule induction are conducted concurrently. In our experiments four traditional methods of handling missing attribute values were applied: Most Common Value, Concept Most Common Value, Closest Fit, and Concept Closest Fit. Both Closest Fit methods were enhanced by a rough set approach to missing attribute values. On the same typical data sets experiments were conducted using three different rough-set interpretations of missing attribute values: lost values, "do not care" conditions and attribute-concept values using the MLEM2 rule induction algorithm, based on rough set theory. The best method is the Concept Closest Fit enhanced by interpreting remaining missing attribute values as lost values.
Keywords: missing attribute values, incomplete data sets, concept approximations, LERS data mining system, MLEM2 algorithm.

1 Introduction

Real-life data sets are often incomplete, i.e., some attribute values are missing. Mining such incomplete data is challenging. In general, methods to handle missing

WIT Transactions on Information and Communication Technologies, Vol 42, © 2009 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA090161

attribute values may be categorized as *sequential* and *parallel*. In sequential methods, (also called *preprocessing methods*) firstly some technique is used to handle missing attribute values, then the main process of acquiring knowledge is conducted. In parallel methods missing attribute values are taken into account during knowledge acquisition, i.e., both processes are performed concurrently.

Sequential methods include techniques based on deleting cases with missing attribute values, replacing a missing attribute value by the most common value of that attribute, assigning all possible values to the missing attribute value, replacing a missing attribute value by the mean for numerical attributes, assigning to a missing attribute value the corresponding value taken from the closest fit case, or replacing a missing attribute value by a new value, computed from a new data set, considering the original attribute as a decision.

Parallel methods to handle missing attribute values include MLEM2 (Modified Learning from Examples Module, version 2) rule induction algorithm in which rules are induced form the original data set, with missing attribute values considered to be lost values, attribute-concept values, or "do not care" conditions [1, 2]. MLEM2 is an option of the LERS (Learning from Examples based on Rough Sets) data mining system. The C4.5 [3] and CART [4] approaches to missing attribute values are other examples of methods from this group.

There are three important interpretations of missing attribute values. The first is a *lost value* interpretation [5, 6], where we consider a missing attribute value as potentially important but unavailable because it was mistakenly erased or mistakenly not recorded. Here we cannot replace the missing attribute value by any specified attribute value.

In the second interpretation, called a *"do not care" condition* [5,7], it is assumed that a missing attribute value was either irrelevant during data collection or the responded refused to answer a corresponding question. For example, a patient may refuse to answer a question about his/her weight. Typically, a "do not care" condition can be replaced by any possible value from the attribute domain.

The third possibility is an *attribute-concept value* where a missing attribute value may be replaced by any possible value from the attribute domain restricted to the concept to which the case belongs. The *concept* (or *class*) is a set of all cases classified the same way. For example, if we want to use the attribute-concept interpretation of a missing attribute value for the attribute *Temperature* for a patient who is sick with flu, and other patients sick with flu have values of *Temperature* either *high* or *very_high*, then the typical values are *high* and *very_high*. The attribute-concept approach to missing attribute values was introduced in [5].

2 Traditional methods to handle missing attribute values

We will describe four traditional methods to handle missing attribute values, all four are sequential and are considered to be the most successful.

Table 1: Data sets used for experiments.

Data set	Number of cases	Number of attributes	Number of concepts	Type of attributes
Bankruptcy	66	5	2	numerical
Breast cancer	277	9	2	symbolic
Echocardiogram	74	7	2	numerical
Hepatitis	155	19	2	numerical
House	435	16	2	symbolic
Image segmentation	210	19	7	symbolic
Iris	150	4	3	symbolic
Lymphography	148	18	4	symbolic
Wine	178	12	3	symbolic

2.1 Most common value for symbolic attributes, average value for numerical attributes

In this method missing attribute values are replaced by the most common value of the attribute. Thus, a missing attribute value is replaced by the most probable known attribute value. Additionally, every missing attribute value for a numerical attribute is replaced by the average of known attribute values.

2.2 Concept most common value for symbolic attributes, concept average value for numerical attributes

In this method the most common value of the attribute restricted to the concept is used instead of the most common value for all cases. Such a concept is the same concept that contains the case with missing attribute value. A missing attribute value of a numerical attribute is replaced by the average of all known values of the attribute restricted to the concept.

2.3 Global closest fit

The global closes fit method [8] is based on replacing a missing attribute value by the known value in another case that resembles as much as possible the case with the missing attribute value. In searching for the closest fit case we compare two vectors of attribute values, one vector corresponds to the case with a missing attribute value, the other vector is a candidate for the closest fit. The search is conducted for all cases, hence the name global closest fit. For each case a distance is computed, the case for which the distance is the smallest is the closest fitting case

that is used to determine the missing attribute value. Let x and y be two cases. The value of x for the attribute a_i will be denoted by $a_i(x)$, where $i = 1, 2, \ldots, n$. The distance between cases x and y is computed as follows

$$distance\ (x, y) = \sum_{i=1}^{n} distance\ (a_i(x), a_i(y)),$$

where

$$distance\ (a_i(x), a_i(y)) = \begin{cases} 0 & \text{if both } a_i(x) \text{ and } a_i(y) \text{ are} \\ & \text{specified and } a_i(x) = a_i(y)\ , \\ \frac{|a_i(x) - a_i(y)|}{r} & \text{if } a_i(x) \text{ and } a_i(y) \text{ are numbers} \\ & \text{and } a_i(x) \neq a_i(y)\ , \\ 1 & \text{otherwise,} \end{cases}$$

r is the difference between the maximum and minimum of the known values of the numerical attribute with a missing value. If there is a tie for two cases with the same distance, a kind of heuristics is necessary, for example, select the first case. In general, using the global closest fit method may result in data sets in which some missing attribute values are not replaced by known values. Additional iterations of using this method may reduce the number of missing attribute values, but may not end up with all missing attribute values being replaced by known attribute values.

2.4 Concept closest fit

This method is similar to the global closest fit method. The original data set, containing missing attribute values, is first split into smaller data sets, each smaller data set corresponds to a concept from the original data set, then the Global Closest Fit is used for the smaller data sets.

3 Rough set approach to missing attribute values

In this section we will quote some basic ideas of the rough set theory. We will assume that lost values will be denoted by "?", "do not care" conditions will be denoted by "*", and attribute-concept values will be denoted by "–".

An important tool to analyze decision tables 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 $a(x) = v$. 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,

Table 2: Number of missing attribute values.

Name of the data set	Number of missing attribute values in input data set	data set outputted by Global Closest Fit	data set outputted by Local Closest Fit
Bank	113	5	5
Breast cancer	869	117	149
Echocardiogram	21	0	0
Hepatitis	1,035	74	110
House	376	2	2
Image segmentation	1,394	151	137
Iris	210	29	44
Lymphography	931	91	79
Wine	806	119	107

- 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,
- If for an attribute a there exists a case x such that the corresponding value is an attribute-concept value, i.e., $a(x) = -$, then the corresponding case x should be included in blocks $[(a, v)]$ for all specified values $v \in V(x, a)$ of attribute a, where

$$V(x, a) = \{a(y) \mid a(y) \textit{ is specified}, \ y \in U, \ d(y) = d(x)\}.$$

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 $K(x, a) = U$,
- If $a(x) = -$, then the corresponding set $K(x, a)$ is equal to the union of all blocks of attribute-value pairs (a, v), where $v \in V(x, a)$ if $V(x, a)$ is nonempty. If $V(x, a)$ is empty, $K(x, a) = U$.

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

$$(x, y) \in R(B) \textit{ if and only if } y \in K_B(x).$$

Recently characteristic relations were investigated in a number of papers, see, e.g., [9–12].

Table 3: Error rates - I.

Method	Data sets					
	Bankruptcy		Breast cancer		Echocardiogram	
	certain rules	possible rules	certain rules	possible rules	certain rules	possible rules
1	15.15%	15.15%	27.80%	28.88%	37.84%	37.84%
2	7.58%	7.58%	23.47%	24.55%	31.08%	31.08%
3	13.64%	13.64%	29.24%	31.41%	29.73%	29.73%
4	4.55%	4.55%	23.10%	25.27%	37.84%	37.84%
5	13.64%	13.64%	28.88%	29.24%	40.54%	40.54%
6	45.45%	31.82%	29.60%	28.88%	29.73%	29.73%
7	24.24%	24.24%	32.49%	28.52%	29.73%	29.73%

Note that for incomplete data there is a few possible ways to define approximations, we used *concept* approximations [5]. 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 rule induction from incomplete data we used the MLEM2 data mining algorithm, for details see [1]. We used rough set methodology [13], i.e., for a given interpretation of missing attribute vales, *lower* and *upper approximations* were computed for all concepts and then rule sets were induced, *certain* rules from lower approximations and *possible* rules from upper approximations.

4 Experiments

In our experiments we used nine data sets, see Table 1. All of these data sets, except *bankruptcy*, are well-known data accessible at the University of California at Irvine Data Depository. The data set *bankruptcy* was collected by E. Altman and M. Heine at the Bew York University, School of Business, in 1968. Note that only two of data sets, *echocardiogram* and *house*, were used for experiments in their original form. In remaining seven data sets some attribute values, about 35%, were randomly removed, or more exactly, the original attribute values were replaced by symbols of missing attribute values. Two data sets: *image segmentation* and

Table 4: Error rates - II.

Method	Data sets					
	Hepatitis		House		Image	
	certain rules	possible rules	certain rules	possible rules	certain rules	possible rules
1	22.58%	20.65%	5.99%	5.30%	50.95%	52.38%
2	15.48%	12.90%	4.15%	5.30%	16.67%	9.52%
3	20.00%	21.94%	3.92%	4.15%	46.19%	41.90%
4	8.39%	8.39%	4.38%	4.61%	13.81%	12.38%
5	21.94%	21.94%	4.84%	6.45%	45.71%	43.81%
6	21.29%	19.35%	8.06%	6.91%	81.90%	48.10%
7	23.23%	21.29%	8.06%	6.91%	66.19%	57.62%

Table 5: Error rates - III.

Method	Data sets					
	Iris		Lymphography		Wine	
	certain rules	possible rules	certain rules	possible rules	certain rules	possible rules
1	21.33%	21.33%	30.41%	32.43%	28.09%	21.35%
2	4.00%	4.00%	6.08%	6.08%	3.37%	3.93%
3	10.00%	10.67%	24.37%	25.00%	13.48%	13.48%
4	2.67%	2.67%	9.46%	9.46%	5.62%	5.62%
5	16.00%	15.33%	24.32%	25.00%	18.54%	17.98%
6	48.00%	38.67%	40.54%	24.32%	41.57%	28.09%
7	32.00%	23.33%	45.95%	30.41%	35.96%	34.83%

wine were discretized using a discretization method based on agglomerative cluster analysis [14].

Seven methods of handling missing attribute values were used:

1. Most common value for symbolic attributes and average value for numerical attributes,
2. Most common value for symbolic attributes and average value for numerical attributes, both restricted to a concept,

3. Global closest fit, if this method outputs a data set with some missing attribute values, use Method 5,
4. Concept closest fit, if this method outputs a data set with some missing attribute values, use Method 5,
5. Missing attribute values interpreted as *lost values*,
6. Missing attribute values interpreted as *"do not care" conditions*,
7. Missing attribute values interpreted as *attribute-concept values*.

For all data sets except *echocardiogram*, use of Method 5 in Methods 3 and 4 was truly necessary since the corresponding data sets, outputted by Global and Local Closest Fit methods, still contained missing attribute values, see Table 2. The error rates, obtained by ten-fold cross validation, are presented in Tables 3–5.

For every data set and for both rule sets, containing *certain* and *possible* rules, we identified the best method for handling missing attribute values, i.e., the method producing the smallest error rate. Only three methods, out of seven, were winners in this competition. Eight times the winner was Method 4, six times the winner was Method 2, and four times the winner was Method 3.

5 Conclusions

Two traditional methods of handling missing attribute values (Methods 1 and 2), other two traditional methods enhanced by rough-set methodology (Methods 3 and 4, and three rough set methods (Methods 5, 6, and 7) were used in our experiments. Out of these seven methods, only three methods (2, 3, and 4) produced optimal results. Obviously, more experiments are needed, but for time being it is clear that the best methodology is based on the concept closest fit enhanced by interpreting remaining missing attribute values as *lost* (using rough-set methodology), the second best method is the most common value for symbolic attributes and average value for numerical attributes, both restricted to a concept, and the next best method is the global closest fit enhanced by interpreting remaining missing attribute values as *lost* (using rough-set methodology). Note that a choice between using certain and possible rules seems to be not important.

References

[1] Grzymala-Busse, J.W., MLEM2: A new algorithm for rule induction from imperfect data. *Proceedings of the 9th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pp. 243–250, 2002.

[2] Grzymala-Busse, J.W. & Grzymala-Busse, W.J., An experimental comparison of three rough set approaches to missing attribute values. *Transactions on Rough Sets*, **6**, pp. 31–50, 2007.

[3] Quinlan, J.R., *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers: San Mateo, CA, 1993.

[4] Breiman, L., Friedman, J.H., Olshen, R.A. & Stone, C.J., *Classification and Regression Trees*. Wadsworth & Brooks: Monterey, CA, 1984.
[5] Grzymala-Busse, J.W., Three approaches to missing attribute values—a rough set perspective. *Proceedings of the Workshop on Foundation of Data Mining, in conunction with the Fourth IEEE International Conference on Data Mining*, pp. 55–62, 2004.
[6] Stefanowski, J. & Tsoukias, A., On the extension of rough sets under incomplete information. *Proceedings of the RSFDGrC'1999, 7th International Workshop on New Directions in Rough Sets, Data Mining, and Granular-Soft Computing*, pp. 73–81, 1999.
[7] Kryszkiewicz, M., Rules in incomplete information systems. *Information Sciences*, **113(3-4)**, pp. 271–292, 1999.
[8] Grzymala-Busse, J.W., Grzymala-Busse, W.J. & Goodwin, L.K., A comparison of three closest fit approaches to missing attribute values in preterm birth data. *International Journal of Intelligent Systems*, **17(2)**, pp. 125–134, 2002.
[9] Li, T., Ruan, D., Geert, W., Song, J. & Xu, Y., A rough sets based characteristic relation approach for dynamic attribute generalization in data mining. *Knowledge-Based Systems*, **20(5)**, pp. 485 – 494, 2007.
[10] Song, J., Li, T. & Ruan, D., A new decision tree construction using the cloud transform and rough sets. *Proceedings of the Rough Sets and Knowledge Technology Conference*, pp. 524–531, 2008.
[11] Qi, Y., Wei, L., Sun, H., Song, Y. & Sun, Q., Characteristic relations in generalized incomplete information system. *International Workshop on Knowledge Discovery and Data Mining*, pp. 519–523, 2008.
[12] Yang, X.B., Yang, J.Y., Wu, C. & Yu, D.J., Further investigation of characteristic relation in incomplete information system. *Systems Engineering - Theory & Practice*, **27(6)**, pp. 155 – 160, 2007.
[13] Pawlak, Z., *Rough Sets. Theoretical Aspects of Reasoning about Data*. Kluwer Academic Publishers: Dordrecht, Boston, London, 1991.
[14] Chmielewski, M.R. & Grzymala-Busse, J.W., Global discretization of continuous attributes as preprocessing for machine learning. *International Journal of Approximate Reasoning*, **15(4)**, pp. 319–331, 1996.