Mining Incomplete Data—A Rough Set Approach

Jerzy W. Grzymala-Busse

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

Abstract. A rough set approach to mining incomplete data is presented in this paper. Our main tool is an attribute-value pair block. A characteristic set, a generalization of the elementary set well-known in rough set theory, may be computed using such blocks. For incomplete data sets three different types of global approximations: singleton, subset and concept are defined. Additionally, for incomplete data sets a local approximation is defined as well.

1 Introduction

Many real-life data sets are affected by missing attribute vales. Mining such incomplete data is very challenging. Recently we observe intensive activity of the rough set community in this area [1–38].

In a rough set approach to mining incomplete data we may take into account a source of incompleteness. If an attribute value was accidentally erased or is unreadable, we may use the most cautious approach to missing attribute values and mine data using only specified attribute values. This type of missing attribute values will be called *lost* and denoted by "?". Mining incomplete data affected by lost values was studied for the first time in [22]. In this paper two algorithms for rule induction from such data were presented. The same data sets were studied later, see, e.g., [36, 37].

Another type of missing attribute values may happen when a respondent refuses to answer a question that seems to be irrelevant. For example, a patient is tested for flu and one of the questions is a color of hair. This type of missing attribute values will be called a *"do not care" condition* and denoted by "*". The first study of "do not care" conditions, again using rough set theory, was presented in [6], where a method for rule induction in which missing attribute values were replaced by all values from the domain of the attribute was introduced. "Do not care" conditions were also studied later, see, e.g. [24, 25].

In a special case of the "do not care" condition, called an *attribute-concept* value, and denoted by "−", we know that the corresponding case belongs to a specific concept X, and, as a result, we replace the missing attribute value by

J.T. Yao et al. (Eds): RSKT 2011, LNCS 6954, pp. 1–7, 2011.

attribute values for all cases from the same concept X. A *concept* (class) is a set of all cases classified (or diagnosed) the same way. For example, if for a patient the value of an attribute *Temperature* is missing, this patient is sick with *Flu*, and all remaining patients sick with *Flu* have *Temperature* values *high* then using the interpretation of the missing attribute value as the attribute-concept value, we will replace the missing attribute value with *high*. This approach was introduced in [10].

An approach to mining incomplete data presented in this paper is based on the idea of an attribute-value block. A characteristic set, defined by means of such blocks, is a generalization of the elementary set, well-known in rough set theory [39–41]. A characteristic relation, defined from characteristic sets, is, in turn, a generalization of the indiscernibilty relation. As it was shown in [7], incomplete data are described by three different types of approximations: singleton, subset and concept. For rule induction from incomplete data it is the most natural to use the MLEM2 (Modified Learning form Examples Module, version 2) since this algorithm is also based on attribute-value pair blocks.

2 Rough Set Approaches to Missing Attribute Values

Our basic tool to analyze data sets is a *block of an attribute-value pair*. Let (a, v) be an attribute-value pair. For *complete* data sets, i.e., data sets 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 data sets 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.
- 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 case x should be included in blocks $[(a, v)]$ for all known values $v \in V(x, a)$ of attribute a. 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

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

The characteristic relation $R(B)$ is reflexive but—in general—does not need to be symmetric or transitive.

2.1 Global Approximations

Note that for incomplete data there is a few possible ways to define approximations [10, 42]. We will start from global approximations.

Let X be a concept, let B be a subset of the set A of all attributes, and let $R(B)$ be the characteristic relation of the incomplete decision table with characteristic sets $K_B(x)$, where $x \in U$. A *singleton* B-lower approximation of X is defined as follows:

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

A *singleton* B-upper approximation of X is

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

The second method of defining global lower and upper approximations for complete decision tables uses another idea: lower and upper approximations are unions of characteristic sets, subsets of U. There are two possibilities. 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

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

The second possibility 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:

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

Obviously, the subset B-lower approximation of X is the same set as the concept B-lower approximation of X. A *concept* B-upper approximation of the concept X is defined as follows:

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

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.

2.2 Local Approximations

An idea of local approximations was introduced in [20]. A set T of attribute-value pairs, where all attributes belong to the set B and are distinct, will be called a *B-complex.* A block of T, denoted by $[T]$, is the intersection of all blocks of attribute-value pairs (a, v) from T. A B-*local lower* approximation of the concept X is defined as follows

$$\cup\{[T] \mid T \textit{ is a B-complex of } X,\ [T] \subseteq X\}.$$

A B-*local upper* approximation of the concept X is defined as the minimal set containing X and defined in the following way

$$\cup\{[T] \mid \exists \textit{ a family } \mathcal{T} \textit{ of B-complexes of } X \textit{ with } \forall\ T \in \mathcal{T},\ [T] \cap X \neq \emptyset\}.$$

Note that a concept may have more than one local upper approximation [20].

For rule induction from incomplete data, using rough set approach, the most natural is to use the MLEM2 data mining algorithm, for details see [43], since MLEM2 is based on attribute-value pair block as well.

3 Conclusions

An idea of the attribute-value block is extremely useful. We may use it for computing characteristic sets that are used for determining lower and upper approximations. Even more, the same idea is used in rule induction in the MLEM2 algorithm. Note that for completely specified data sets the characteristic relation is reduced to the indiscernibility relation and all three type of global approximations are reduced to ordinary approximations, well-known from rough set theory.

References

1. Cyran, K.A.: Modified indiscernibility relation in the theory of rough sets with real-valued attributes: Application to recognition of fraunhofer diffraction patterns. Transactions on Rough Sets 9, 14–34 (2008)
2. Dai, J., Xu, Q., Wang, W.: A comparative study on strategies of rule induction for incomplete data based on rough set approach. International Journal of Advancements in Computing Technology 3, 176–183 (2011)
3. Dardzinska, A., Ras, Z.W.: Chasing unknown values in incomplete information systems. In: Workshop Notes, Foundations and New Directions of Data Mining, in Conjunction with the 3-rd International Conference on Data Mining, pp. 24–30 (2003)
4. Dardzinska, A., Ras, Z.W.: On rule discovery from incomplete information systems. In: Workshop Notes, Foundations and New Directions of Data Mining, in conjunction with the 3-rd International Conference on Data Mining, pp. 24–30 (2003)

5. Greco, S., Matarazzo, B., Slowinski, R.: Dealing with missing data in rough set analysis of multi-attribute and multi-criteria decision problems. In: Zanakis, H., Doukidis, G., Zopounidised, Z. (eds.) Decision Making: Recent Developments and Worldwide Applications, pp. 295–316. Kluwer Academic Publishers, Dordrecht (2000)
6. 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)
7. 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 3-rd International Conference on Data Mining, pp. 56–63 (2003)
8. Grzymala-Busse, J.W.: Data with missing attribute values: Generalization of indiscernibility relation and rule induction. Transactions on Rough Sets 1, 78–95 (2004)
9. Grzymała-Busse, J.W.: Characteristic relations for incomplete data: A generalization of the indiscernibility relation. In: Tsumoto, S., Słowiński, R., Komorowski, J., Grzymała-Busse, J.W. (eds.) RSCTC 2004. Current Trends, vol. 3066, pp. 244–253. Springer, Heidelberg (2004)
10. Grzymala-Busse, J.W.: Three approaches to missing attribute values—a rough set perspective. In: Proceedings of the Workshop on Foundation of Data Mining, in Conjunction with the Fourth IEEE International Conference on Data Mining, pp. 55–62 (2004)
11. Grzymała-Busse, J.W.: Incomplete data and generalization of indiscernibility relation, definability, and approximations. In: Ślęzak, D., Wang, G., Szczuka, M.S., Düntsch, I., Yao, Y. (eds.) RSFDGrC 2005. LNCS (LNAI), vol. 3641, pp. 244–253. Springer, Heidelberg (2005)
12. Grzymala-Busse, J.W.: A comparison of traditional and rough set approaches to missing attribute values in data mining. In: Proceedings of the 10-th International Conference on Data Mining, Detection, Protection and Security, Royal Mare Village, Crete, pp. 155–163 (2009)
13. Grzymala-Busse, J.W.: Mining data with missing attribute values: A comparison of probabilistic and rough set approaches. In: Proceedings of the 4-th International Conference on Intelligent Systems and Knowledge Engineering, pp. 153–158 (2009)
14. Grzymala-Busse, J.W., Grzymala-Busse, W.J.: Handling missing attribute values. In: Maimon, O., Rokach, L. (eds.) Data Mining and Knowledge Discovery Handbook, pp. 37–57. Springer-Verlag, Heidelberg (2005)
15. 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, 31–50 (2007)
16. Grzymala-Busse, J.W., Grzymala-Busse, W.J.: Improving quality of rule sets by increasing incompleteness of data sets. In: Cordeiro, J., Shishkov, B., Ranchordas, A., Helfert, M. (eds.) ICSOFT 2008. Communications in Computer and Information Science, vol. 47, pp. 241–248. Springer, Heidelberg (2009)
17. 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), 125–134 (2002)
18. Grzymala-Busse, J.W., Grzymala-Busse, W.J., Hippe, Z.S., Rzasa, W.: An improved comparison of three rough set approaches to missing attribute values. In: Proceedings of the 16-th Int. Conference on Intelligent Information Systems, pp. 141–150 (2008)

19. Grzymała-Busse, J.W., Hu, M.: A comparison of several approaches to missing attribute values in data mining. In: Ziarko, W.P., Yao, Y. (eds.) RSCTC 2000. LNCS (LNAI), vol. 2005, pp. 378–385. Springer, Heidelberg (2001)
20. Grzymala-Busse, J.W., Rzasa, W.: Local and global approximations for incomplete data. In: Greco, S., Hata, Y., Hirano, S., Inuiguchi, M., Miyamoto, S., Nguyen, H.S., Słowiński, R. (eds.) RSCTC 2006. LNCS (LNAI), vol. 4259, pp. 244–253. Springer, Heidelberg (2006)
21. Grzymala-Busse, J.W., Rzasa, W.: Local and global approximations for incomplete data. Transactions on Rough Sets 8, 21–34 (2008)
22. 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)
23. Hong, T.P., Tseng, L.H., Chien, B.C.: Learning coverage rules from incomplete data based on rough sets. In: Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, pp. 3226–3231 (2004)
24. Kryszkiewicz, M.: Rough set approach to incomplete information systems. In: Proceedings of the Second Annual Joint Conference on Information Sciences, pp. 194–197 (1995)
25. Kryszkiewicz, M.: Rules in incomplete information systems. Information Sciences 113(3-4), 271–292 (1999)
26. Latkowski, R.: On decomposition for incomplete data. Fundamenta Informaticae 54, 1–16 (2003)
27. Latkowski, R., Mikołajczyk, M.: Data decomposition and decision rule joining for classification of data with missing values. In: Tsumoto, S., Słowiński, R., Komorowski, J., Grzymała-Busse, J.W. (eds.) RSCTC 2004. LNCS (LNAI), vol. 3066, pp. 254–263. Springer, Heidelberg (2004)
28. Li, H., Yao, Y., Zhou, X., Huang, B.: Two-phase rule induction from incomplete data. In: Wang, G., Li, T., Grzymala-Busse, J.W., Miao, D., Skowron, A., Yao, Y. (eds.) RSKT 2008. LNCS (LNAI), vol. 5009, pp. 47–54. Springer, Heidelberg (2008)
29. Li, D., Deogun, I., Spaulding, W., Shuart, B.: Dealing with missing data: Algorithms based on fuzzy set and rough set theories. Transactions on Rough Sets 4, 37–57 (2005)
30. Peng, H., Zhu, S.: Handling of incomplete data sets using ICA and SOM in data mining. Neural Computing and Applications 16, 167–172 (2007)
31. 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), 485–494 (2007)
32. Nakata, M., Sakai, H.: Rough sets handling missing values probabilistically interpreted. In: Ślęzak, D., Wang, G., Szczuka, M.S., Düntsch, I., Yao, Y. (eds.) RSFDGrC 2005. LNCS (LNAI), vol. 3641, pp. 325–334. Springer, Heidelberg (2005)
33. Qi, Y.S., Sun, H., Yang, X.B., Song, Y., Sun, Q.: Approach to approximate distribution reduct in incomplete ordered decision system. Journal of Information and Computing Science 3, 189–198 (2008)
34. Qi, Y.S., Wei, L., Sun, H.J., Song, Y.Q., Sun, Q.S.: Characteristic relations in generalized incomplete information systems. In: International Workshop on Knowledge Discovery and Data Mining, pp. 519–523 (2008)
35. Song, J., Li, T., Ruan, D.: A new decision tree construction using the cloud transform and rough sets. In: Wang, G., Li, T., Grzymala-Busse, J.W., Miao, D.,

Skowron, A., Yao, Y. (eds.) RSKT 2008. LNCS (LNAI), vol. 5009, pp. 524–531. Springer, Heidelberg (2008)
36. Stefanowski, J., Tsoukiàs, A.: On the extension of rough sets under incomplete information. In: Zhong, N., Skowron, A., Ohsuga, S. (eds.) RSFDGrC 1999. LNCS (LNAI), vol. 1711, pp. 73–82. Springer, Heidelberg (1999)
37. Stefanowski, J., Tsoukias, A.: Incomplete information tables and rough classification. Computational Intelligence 17(3), 545–566 (2001)
38. Wang, G.: Extension of rough set under incomplete information systems. In: Proceedings of the IEEE International Conference on Fuzzy Systems, pp. 1098–1103 (2002)
39. Pawlak, Z.: Rough sets. International Journal of Computer and Information Sciences 11, 341–356 (1982)
40. Pawlak, Z.: Rough Sets. In: Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Dordrecht (1991)
41. Pawlak, Z., Grzymala-Busse, J.W., Slowinski, R., Ziarko, W.: Rough sets. Communications of the ACM 38, 89–95 (1995)
42. Grzymala-Busse, J.W., Rzasa, W.: A local version of the MLEM2 algorithm for rule induction. Fundamenta Informaticae 100, 99–116 (2010)
43. 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)